Do Fraudulent Firms Produce Abnormal Disclosure?

Gerard Hoberg* and Craig Lewis*

November 24, 2014

ABSTRACT

We present two new hypotheses regarding the textual disclosures of fraudulent firms. First, these firms discuss performance in a manner that is similar to their industry peers. Second, their qualitative disclosures are distinct from their industry peers but instead are similar to other fraudulent firms. We use text-based analysis of 10-K MD&A disclosures to compare disclosures of firms involved in SEC enforcement actions to various counterfactuals including each firm's own disclosure both before and after the alleged violations. We find evidence that fraudulent firms do not make qualitative disclosures that resemble their industry peers but instead cluster with other fraudulent peer firms. Content analysis reveals that fraudulent firms under-disclose details relating to governance, financial liquidity and explaining revenues.

*University of Southern California and Vanderbilt University, respectively. We thank Ken Ahern, Christopher Ball, Kathleen Hanley, Tim Loughran, Vojislav Maksimovic, Bill McDonald, Gordon Phillips, and Harvey Westbrook for excellent comments and suggestions. We also thank seminar participants at Columbia University, Duke University, George Washington University, London Business School, London School of Economics, U. S. Securities and Exchange Commission, 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. Any remaining errors are ours alone.

Many studies suggest that managers committing fraud likely do so to achieve various objectives such as getting access to low cost capital (Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010)) or to conceal diminishing performance (Dechow, Ge, Larson and Sloan (2011)).[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 decision to commit fraud. This issue should be particularly salient to managers committing fraud because the SEC is tasked with identifying and pursuing enforcement actions against these 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.

The high degree of discretion associated with managements' qualitative discussion of its operating performance in the 10-K creates opposing forces regarding how a firm might position its results. On one hand, fear of detection might lead fraudulent managers to disclose in a way that attempts to minimize detection. For example, a manager might under-disclose explanations of their fraudulent revenue or expense calculations. Alternatively, they may over-disclose complex transactions to increase the cost of detection. Such managers might also use excessively difficult to read text, or they may simply mimic the disclosures of their industry peers to appear "ordinary". These considerations suggest that qualitative disclosures may be designed to strategically mislead financial statement readers.

While it may be the case that some firms strategically design qualitative disclosures, an alternative and nearly empirically indistinguishable explanation is that the same economic conditions that induce managers to commit fraud generate qualitative discussions that are consistent with our main hypotheses. For example, a firm may experience a negative shock, such as a labor strike or a product liability claim, that requires discussion in the MD&A, which differentiates it from its industry peers. Alternatively, a negative industry-wide shock could result in similar qualitative discussions among industry peers as they describe

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

similar economic conditions. In both of these examples, common qualitative disclosures by fraudulent firms may be artifacts of either idiosyncratic or industry-wide conditions, and do not necessarily reflect proactive attempts to influence perceptions. Although we acknowledge and explore this possibility, it should be noted that these firms have already chosen to pro-actively and fraudulently disclose their quantitative performance. Hence, on the margin, the cost associated with attempting to disguise this fraud using manipulated verbal disclosure might be relatively low.

Regardless of the underlying motive, the possibility that MD&A might contain a signal that predicts whether firms have engaged in accounting fraud is an important research question. While it is further important to then differentiate between potentially competing explanations, the ability to improve the prediction of accounting fraud is practically relevant to future researchers, investors and regulators alike, and it motivates future theoretical and empirical research to understand why.

We use an empirical framework that incorporates the possibility that common disclosure is either strategic or incidental by first examining whether verbal disclosure is abnormal relative to three different benchmarks. The first examines whether each firm's raw disclosure is similar to that of industry peers of similar size and age. The second considers abnormal disclosure that is common to fraudulent firms after controlling for the disclosure of the industry peers. This entails purging each firm's disclosure of the common industry component related to similar size and age, and then examining whether firms involved in alleged fraud have systematically different disclosures. The third focuses on the time series behavior of the fraudulent firm itself, and examines if firms have disclosures that differ from themselves in the years prior to and after their alleged fraud. We find strong and uniform support for our central hypothesis that firms committing fraud have a strong common component in their disclosures, and this component is much less prevalent among firms not committing fraud. It also cannot be explained by the disclosure of industry peers or firm fixed effects, and hence this disclosure only appears abnormally during the specific

years firms are committing fraud.

Having established the presence of abnormal disclosure, we turn our attention to the question of why it is present. We first test two specific non-strategic hypotheses: (1) disclosure might appear aberrational relative to peers because certain disclosures are correlated with the likelihood of an SEC review, and (2) certain disclosures are simply correlated with poor economic conditions, which in turn are correlated with fraud. If (1) is true, then the links we find between disclosure and alleged fraud may be partially attributable to internal SEC procedures. To consider (1), we conduct tests examining the link between disclosure and the issuance of Comment Letters, which are produced by the SEC's Division of Corporation Finance, which is tasked with the intake and evaluation of verbal disclosure. Under this alternative, vocabulary linked to comment letters should coincide strongly with the vocabulary linked to fraud. Our findings reject this hypothesis, as we find that the thematic drivers of comment letters have little overlap with the thematic drivers of fraud. These findings are not surprising given that Dyck, Morse and Zingales (2010) find that SEC reviews likely account for very little fraud detection (employees and the media being more relevant).

Regarding the economic conditions hypothesis (2), we consider three tests. First, we examine disclosures relative to size-age-industry matched peers in each year after controlling for both firm and year fixed effects. This difference-based approach absorbs variation associated with economic conditions (these peers, being in the same industry, likely face similar economic conditions). Second, and perhaps even more directly, we compute the "disclosure implied" economic conditions of the firm. To do this, we control for the implied Tobins' Q and profitability of each firm in each year based on the actual Tobins Q and profitability of other firms that have MD&A sections of their 10-K that are most similar to the given firm. If disclosures are systematically representative of poor economic conditions when firms engage in fraud, these measures should subsume our existing fraud variables in our key regressions. Third, and as we discuss next, we identify the verbal themes that

explain why our disclosure variables are informative regarding firms that are involved in fraud. We interpret these themes, which are based on the LDA factors of Ball, Hoberg and Maksimovic (2013), and discuss their potential links to each hypothesis. Although we cannot rule out a role for economic conditions, these three tests do not support the conclusion that economic conditions can fully explain our results.

Overall, we find that fraudulent managers make qualitative disclosures that are different from their industry peers, but are common among fraudulent-firm peers. We draw this conclusion based on conservative specifications using both cross sectional and time series differences with firm fixed effects, and we stress test them using interpretable thematic analysis of the specific vocabulary that drives our key results. Our study relies on text analytic methods including the cosine similarity method and Latent Dirichlet Allocation (LDA). Our use of cosine similarities is through the lens of whether fraudulent firms disclose in a way that is similar to or different from industry peers of similar size and age, and also whether their disclosure is similar to or different from other firms committing fraud. The cosine similarity method is a standard approach used in computational linguistics (See Sebastiani (2002) for example). It is easy to interpret given its range in the interval [-1,1] and its standardization, which controls for document length.

We also consider LDA content analysis to identify the key themes that firms involved in AAERs use relative to peers not involved in AAERs. LDA is a topic modeling technique developed by Blei, Ng and Jordan (2003). It is a generative model solved using likelihood analysis that discovers clusters of text (referred to as "topics") that frequently appear in various documents. LDA is intuitively akin to a sophisticated text-based analog of factor analysis (commonly used for numerical data). We discuss LDA in greater detail in Section 5. Specifically, we consider the MD&A LDA factors from Ball, Hoberg and Maksimovic (2013), and examine which particular themes are associated with firms committing fraud.

Our content analysis suggests that fraudulent firms tend to under-discuss factors that might explain potentially fraudulent accounting, including for example, attribution text

explaining revenues, financial market liquidity, and discussion of the management team itself. These firms also excessively discuss acquisitions, product lines, and growth strategies. These verbal clusters can be consistent with specific motives managers might have to commit fraud, and hence they might be viewed as evidence for strategic disclosure. However, these tests do not rule out a role for economic conditions, as economic conditions that induce fraud might relate to these variables through obscure channels.

In a final test, we thus consider the specific strategic hypothesis that managers commit fraud to artificially improve their odds of issuing equity. Although other studies find evidence consistent with this motive[2], no existing studies report supportive evidence in verbal disclosures. Using an exogenous shock to equity market liquidity, which increases the motive to commit fraud for this reason, we find that treated firms produce disclosure that becomes more similar to fraudulent firms. In turn, we also find that the use of this common fraudulent disclosure is associated with higher rates of equity issuance. These results provide some suggestive evidence that at least some of our findings might be due to managers using verbal disclosure to further achieve the same goals that drive them to commit fraud in the first place.

The remainder of this article is organized as follows. Section I reviews the existing literature and presents our hypotheses. Section II describes our data and methodology. Section III presents our data and summary statistics and Section IV presents our central disclosure regressions. Section V presents content analysis and summarizes the vocabulary of fraudulent firms, and Section VI concludes.

[2] See Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010).

1 Literature and Hypotheses

1.1 Existing Literature

Many studies examine the links between accounting, stock returns and AAERs. Feroz, Park, and Pastena (1991) and Karpoff, Lee and Martin (2008b) examine the issues that motivate fraud and their consequences. Although we cannot summarize all literature in the area due to space constraints, we refer readers to Dechow, Ge and Schrand (2010) for a thorough review.

Earlier work links standard accounting variables with fraudulent activity. Beneish (1997) considers a Jones model, and examines whether firms that manipulate earnings can be separated from those that merely have more aggressive accruals. Beneish (1999) considers a host of accounting ratios and constructs an index. Dechow, Sloan and Sweeney (1996) find that a strong motive for earnings management is the desire to attract low cost financing. Beneish (1999) finds that managers are more likely sell their own shares when earnings are overstated.

More recent studies extend these earlier works and provide more depth. Dechow, Ge, Larson and Sloan (2011) find that mis-stating firms hide diminishing performance, have higher relative prices, and have abnormal reductions in the number of employees. Wang (2013) addresses the partial-observability of fraud, and finds that R&D increases the likelihood of fraud while also reducing the likelihood of detection. Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010) show theoretically and empirically that the incentive to commit fraud is more intense during industry booms.

Dyck, Morse and Zingales (2010) take a different approach and examine who is most likely to "blow the whistle". The authors find that investors, the SEC, and auditors play only a small role, whereas employees and the media play a larger role. Kedia and Philippon (2009) find links to corporate hiring, executive option exercise, and firm productivity. Kedia and Rajgopal (2011) show that firm locations relative to SEC offices, and areas with past

enforcement activity, are less likely to be involved in restatements.

The body of existing work regarding the incentives to commit fraud, and evidence that can be used to detect fraud, is extensive. Yet, although many studies use text in SEC disclosures to test various hypotheses, none have made a specific link to accounting fraud[3].

1.2 Hypotheses

We briefly describe four hypotheses that might explain why firms committing fraud might produce verbal disclosures that differ from similar firms that are not committing fraud. Two hypotheses suggest that managers might act strategically, and two suggest that common disclosure might be incidental or related to economic conditions.

H1 [Strategic Disclosure to Conceal Fraud]: *Managers committing fraud engage in strategic disclosure to conceal the fraud and to reduce the likelihood of detection.*

There are at least three channels through which H1 might be implemented. First, managers might under-disclose specific discussions, such as detailed explanations of revenue or expenses. Second, managers might strategically "herd" with industry peers to appear "ordinary" from the regulator's perspective. Third, managers might produce disclosures that are difficult to read (e.g., a high fog index). We examine all three channels.

H2 [Strategic Disclosure to Achieve Fraud-Driven Motives]: *Managers of firms committing fraud will over-disclose or under-disclose various topics to achieve the same specific benefits that led the firm to commit fraud.*

An example is that managers might under-disclose problems with financial market liquidity to further increase the likelihood of getting access to low cost capital. Similarly, managers committing fraud to report strong revenue growth and attract more customers might excessively grandstand their firm's growth.

[3] Antweiler and Frank (2004) and Tetlock (2007) are among the earliest studies in Finance. Also see Cole and Jones (2005), Feldman, Govindaraj, Livnat, and Segal (2010), Li (2008), Li (2010), Brown and Tucker (2011), Hanley and Hoberg (2010), Hoberg and Phillips (2010), Kothari, Li and Short (2009) , Loughran and McDonald (2011), and Bryan (1997). These studies examine textual tone, information content, readability, links to economic quantities, revision intensity, and the cost of capital.

H3 [Likelihood of Regulatory Review]: *Managers do not act strategically. Instead, some LDA topics correlate with AAER actions because firms with certain disclosures are more likely to be reviewed by the SEC.*

Because reviews by the SEC's Division of Corporate Finance, and the comment letter process, is the primary intake and review process used by the SEC regarding 10-K disclosures, H3 further predicts that the verbal content used by fraudulent managers should strongly overlap with the verbal content that most associates with comment letters.

H4 [Economic Conditions]: *Firms involved in fraud might have common components in their disclosure that relate to poor economic conditions. For example, firms committing fraud might be experiencing losses.*

Hypothesis H4 is perhaps the most difficult to separate empirically from the other hypotheses. It predicts that controls for the disclosure of industry peers facing common industry conditions should explain much of the common disclosure produced by fraudulent firms. These tests rely on a number of measures to absorb this variation, and further include controls for the economic state of the firm that is implied by its MD&A disclosure (we include a control for the profitability and Tobins q of firms with similar MD&A disclosures).

2 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 (extracted using software provided by metaHeuristica LLC) of annual firm 10-Ks.

We first extract COMPUSTAT observations from 1997 to 2008 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 87,887 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 71,637 observations. After requiring that observations are in the CRSP database, we have 60,853 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 2008 as this is the final year of our AAER database.

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.[4] 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.[5] These requirements leave us with a final sample of 49,039 firm-year observations having adequate data.

2.1 Accounting and Auditing Enforcement Releases

We obtain data on Accounting and Auditing Enforcement Releases (AAERs) from the Securities and Exchange Commission website[6]. Our hand collected sample includes AAERs indicating fraudulent behavior from 1997 to 2008. In addition to firm identifying data, which is needed to link AAER firms to our Compustat universe, we also collect the filing date of each AAER, and the beginning and ending dates each AAER alleges fraudulent activity. We define our AAER dummy to be one for firm fiscal years ending in calendar 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 each AAER, we also identify a year that is definitively prior to the alleged fraudulent

[4] We use the WRDS SEC Analytics package to link 10-Ks to Compustat.

[5] 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.

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

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, that 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.

2.2 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".[7]

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).[8] The resulting list of words is stable over time, as 99.1% of randomly drawn words using our 1997-based screen would 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 $ISA_{i,t}$.

[7] 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.

[8] 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 5,000 words. Because we also do not see a material degree of improvement in going from 5,000 to 10,000 words, we thus conclude that our universe is sufficiently refined to provide a relevant signal for testing our key hypotheses.

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

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.

2.3 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):

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

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 2008 for training as we wish to preserve these years for assessing the out of sample 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:

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

3 Data and Summary Statistics

Table 1 displays summary statistics for our panel of 49,039 firm-year observations from 1997 to 2008 having machine readable MD&As. 1.5% of firm year observations are AAER-years. As it is based on cosine similarities between positive and negative word vectors, the Fraud Similarity Score has a distribution in the interval [-1,+1] and a mean that is close to zero. Intuitively, because AAER years are rare, the average firm does not have a vocabulary that correlates highly with fraudulent firms. The industry similarity score is based on cosine similarities of non-negative vectors, and is bounded in the interval [0,1]. Its mean of 0.667 indicates that the average firm shares a substantial amount of disclosure with its ISA peers. However, the average firm also has much unique content.

[Insert Table 1 Here]

Table 2 displays Pearson correlation coefficients. The positive 8.2% 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 2.6%. Remarkably, the fraud similarity score is more correlated with the AAER dummy than any of the other displayed variables including firm size (7.0% correlation).

[Insert Table 2 Here]

Fraud similarity is 9.0% correlated with industry similarity (significant at the 1% level). Given that both variables are functions of firm disclosures, this is somewhat modest. 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, 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.

[Insert Table 3 Here]

Table 3 displays time series summary statistics regarding AAER-year observations in our sample from 1997 to 2008. The table shows a peak in 2000 to 2002 following the internet bubble's collapse, and also a steady stream of AAER years throughout our sample with the exception of the last three 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. We also note that, in all, 2.9% of our sample firms (249 of 8510) were involved in an AAER at some point in time in our sample. The relatively low rate of AAERs during the financial crisis of 2007 and 2008 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 during the crisis.

3.1 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.

[Insert Table 4 Here]

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 3.7% compared to just 0.5% for decile 1. The positive link between industry similarity and AAER incidence is weaker with high to low decile range of 2.7% to 1.0%.

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 2.3). Hence, our assessment of the link between AAERs and the fraud scores in 2002 to 2008 is an out of sample test on all levels. We continue to observe strong positive associations with AAER incidence rates for fraud similarity, and the inter-decile range is 0.7% to 2.2%. 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, 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 vocabulary used by AAER firms, that is distinct from industry-size-age peers, has remained stable over time.

3.2 Fraud Similarity Distributions

In this section, we examine the distribution of fraud similarity. Figure 1 shows the empirical density function of this variable over its domain [-1,1]. The distribution is centered near zero and is nearly bell shaped. However, it is somewhat asymmetric and right skewed, indicating that observations are potentially drawn from a mixed distribution where potentially fraudulent firms have a higher mean than non-fraudulent firms. 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 2.55% of the total mass. This is materially larger than the observed 1.5% AAER rate indicated in Table 1.

[Insert Figure 1 Here]

We consider whether the rate of undetected fraud can be estimated. To do so, we make two assumptions that are unique to this exercise. These assumptions are not relevant to the tests in other parts of the paper. First, we assume that non-fraudulent firms have symmetrically distributed fraud similarities. Second, we assume that firms engaged in fraud that is not yet detected have a similar distribution compared to those that are detected. These assumptions allow us to estimate the extent of undetected fraud based on how many firms would have to be removed from the sample to eliminate the observed asymmetry. We note that whether or not these assumptions hold likely depends critically on the nature of how fraud is detected, and whether the mechanism strongly relates to verbal text in the disclosure, even after controlling for ISA peers. Although it is unlikely that these assumptions hold precisely, the results in Dyck, Morse and Zingales (2010) suggest that they might only be weakly violated. In particular, the authors find that the primary consumers of 10-Ks (investors, the SEC, and auditors) play only a small role in detecting frauld. Employees and the media play a larger role.

[Insert Figure 2 Here]

We next assess the extent to which the removal of known AAER firm-years reduces asymmetry. Figure 2 plots the density function of fraud similarity separately for firms not involved in AAERs (upper figure) and involved in AAERs (lower figure). The figure shows that the density function retains a substantial degree of asymmetry even when known AAER firm-years are excluded, as the right-skewed mass only decreases from 2.55% to 2.10%. We thus compute the upper bound regarding the rate of undetected fraud as the fraction of the sample that would have to be removed to eliminate all observed asymmetric

mass. This calculation suggests that just 17.6% ($\frac{2.55-2.10}{2.55}$) of fraudulent firms have been detected and hence fraud is 5.6x as pervasive as observed. We compute a lower bound by assuming that the 2.1% of remaining asymmetry in Figure 2 is due to 2.1% of undetected firms being engaged in fraud. This would imply that fraud is 2.4x as pervasive as observed. Because the observed rate of known AAER firm years is 1.5%, these estimates indicate that the actual rate of committed AAERs likely lies in the range (3.6%, 8.5%) of all firm-years. This range is substantially higher than the 1.5% detection rate in our sample.

The lower plot in Figure 2 further illustrates why our approach might have good power for estimating undetected 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 asymmetry than any of the other figures, indicating that fraud similarity is effective in separating AAER firms from non-AAER firms. The degree of asymmetric mass is 41.0%, which is far larger than the 2.1% in the upper figure.

Figure 3 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.

[Insert Figure 3 Here]

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 0.025. During the period of alleged fraud, this score more than doubles to over 0.05, and remains near this level during the years of alleged fraud. After the period of alleged fraud ends, fraud similarity then drops sharply to 0.025 and then 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 automatically account for firm fixed effects.

4 Disclosure and Fraud Regressions

In this section, we use regression analysis to test our strategic 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. In Panel 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 is based on within-firm variation only (they include controls for firm and year fixed effects). Standard errors are adjusted for clustering by firm. We also include several controls 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).[9]

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 industry fixed effects instead of firm fixed effects.

[Insert Table 5 Here]

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 6.58, and is significant well beyond the 1% level. The results for industry similarity are not significant with a t-statistic of 0.8. 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 Figure 3, 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 commitment of the fraud itself.

Panels B and C of Table 5 show that fraud profile similarity is robust at the 1% level for both large and 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 produce verbal disclosures that have a strong common component that cannot be explained by industry, size and age (ISA peers).

Although these results are stark, they do not strongly rule out any of our hypotheses. For example, they are consistent with potential strategic disclosure (H1 and H2) during fraud years. They also are consistent with institutional review mechanisms (H3). However, it should be noted that our findings do not support the auxiliary prediction of H3 that fraud scores will remain high after the years fraud is committed given that fraud detection

[9]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.

typically occurs much later.

Finally, the results remain consistent with economic conditions (H4). This explanation, however, is less persuasive on the margin because the regression specifications explicitly control for a number of proxies for economic conditions, i.e., firm and year fixed effects plus the implied economic conditions of firms with similar profitability and Tobins q.

Panel D, shows that our results remain robust during the out of sample period from 2002 to 2008. This test is particularly stringent, as the sample is smaller, and the impact of firm fixed effects on remaining degrees of freedom is more extreme. Nevertheless, the fraud similarity variable remains significant at the 5% level with a t-statistic of 2.31.

In Panel E, we further challenge our specification by including four additional control variables: restatements, litigation, uncertainty and mergers.[10] 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 replacing firm fixed effects with less stringent SIC-2 industry fixed effects. Not surprisingly, the results are stronger. This indicates that although our results are primarily driven by within-firm variation, variation across industries also goes in the same direction, further supporting our key hypotheses.

Table 6 uses the same framework as Table 5, except that we consider the future AAER

[10]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.

dummy (a dummy that is one if the firm will be involved in an AAER in the next fiscal year) as an explanatory variable instead of the actual AAER dummy. As a result, we are implicitly testing if fraud similarity is elevated in the year prior to the fraud period. This allows us to test hypotheses predicting that disclosure will strictly relate to the act of committing fraud, and not to passive long term firm characteristics. We thus expect that the results should be substantially weaker than those in Table 5.

[Insert Table 6 Here]

Table 6 shows uniformly weak and statistically insignificant links between fraud profile similarity and the future AAER dummy. These results are thus much weaker than those in Table 5. This reinforces the graphical depiction of the average fraud score in Figure 3, which shows that fraud scores are close to zero prior to the fraud period. We conclude that our evidence in Table 5 is strongly linked to the years that firms are allegedly engaged in fraud and our results cannot be explained by passive long-term firm characteristics.

[Insert Table 7 Here]

Table 7 is similar to Table 6, except we replace the future AAER dummy with the past AAER dummy. Hence, the dummy identifies firms that have committed fraud in the past, but are no longer committing fraud. The results of Table 7 are similar to those of Table 6 in that fraud profile similarity is not positively related to AAERs. In some specifications, we in fact find a negative link. 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, these results further show that our results are not related to passive firm characteristics, and are unique to firms allegedly committing fraud.

5 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, this finding does not provide strong separation of our four hypotheses. In this section, we consider content analysis using the 75 verbal factors based on Latent Dirichlet Allocation (LDA) from Ball, Hoberg and Maksimovic (2013). In particular, we report the key vocabulary themes from LDA that distinguish firms involved in AAERs from non-AAER firms. By again focusing on a stringent difference-based framework that includes firm fixed effects, the reported results in this section are also based on within-firm variation. We thus identify the specific verbal topics that appear while firms are allegedly involved in fraud, as compared to the same firms in the years they are not involved in fraud.

We also report which verbal themes appear in counter-factual periods: the year prior to AAER-years and in the year after AAER-years. We then interpret each of the themes through the lens of our key hypotheses. These tests not only provide evidence that can support or reject hypotheses, but in addition, they can reveal information concerning the specific mechanisms through which managers involved in fraud alter their disclosure. We attribute this testing framework to Ball, Hoberg and Maksimovic (2013), who examine business change.

We then conduct similar analysis for SEC comment letters. This test is particularly relevant to examining hypothesis H3. This test is motivated by the role of reviews by the Division of Corporate Finance of the SEC in reviewing and commenting on verbal disclosure such as MD&A. This division is tasked with providing comment letters directly to the issuer when there are concerns, and hence this test of H3 is quite direct regarding whether our results relate to detection mechanisms in place at the SEC relating to the intake of disclosure. Under H3, we predict that the thematic drivers of AAERs and comment letters

should be very similar. If they are sharply different, it is more difficult for H3 to explain our results.

5.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) 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 and is replicable. It also does not require researcher prejudgment in that the researcher is not required to make assumptions about specific topics to be found in the document, or the associated word distributions.

The approach is commonly referred to as a "bag-of-words" technique because the relative frequency of words in a document is vitally important but not their specific ordering. A particular topic can be characterized as a distribution over a common vocabulary of words where the relative probability weight assigned to each word indicates its relative importance to that topic. For example, the word "oil" may receive relative high probability weights in topics that are associated with Manufacturing and Natural Resources. By contrast, "electric" may have nontrivial probability weights in both topics but have a relatively higher weight in the Manufacturing topic.

Each document is modeled as a random mixture over these topics. Intuitively, the weight of each topic that is assigned to a particular document reflects its relative importance to the document. For example, the MD&A sections of British Petroleum and General Motors would both be expected to use the word "oil" but the documents might be expected to place greater emphasis respectively on the Manufacturing and Natural Resources topics respectively - both of which place relative high weight on this word.

LDA was developed by Blei, Ng and Jordan (2003) to provide an analytic framework that allows one to estimate the topic densities from a corpus of documents. We provide only a brief summary here, and refer readers to these articles for more detail. For our purposes, LDA is a generalization of factor analysis (used in numerical data) to textual data. LDA

uses Gibbs Sampling and likelihood analysis and discovers clusters of text ("topics") that frequently appear in a corpus. We use the LDA topics from Ball, Hoberg and Maksimovic (2013), which were generated using the metaHeuristica software program.

LDA generates two detailed data structures. The first data structure 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. As do Ball, Hoberg and Maksimovic (2013), we fit the LDA model vocabularies using only the first year of our sample (1997) to ensure there is no look ahead bias in the regressions that use LDA text.

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.[11] This data structure is a detailed summary of MD&A content. It has reduced dimensionality because it summarizes each document using a vector of length 75, whereas raw MD&As have a dimensionality exceeding 50,000, which is the number of unique words in the corpus of MD&As.

We use these two LDA-generated data structures to refine our understanding of disclosure during episodes of fraud. In particular, we use the highest frequency commongrams to provide labels for the 75 topics. A "commongram" is a set of two or more contiguous words that appears with high frequency in the corpus.

We then use the panel data 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

[11] Generating the database of topic loadings is achieved by projecting the distribution of text for any given MD&A on the 75 vectors representing the distribution of text for each of the topics. See Ball, Hoberg and Maksimovic (2013) for details regarding this regression-based approach. Because LDA generates topics reflecting nearly orthogonal clusters of vocabulary, this projection is not susceptible to multi-collinearity and we implement the procedure using this simple approach. This allows us to build a database of topic loadings for our entire sample 1997 to 2011 using the topic vocabularies from 1997 as discussed above.

different can then be interpreted and discussed regarding their potential consistency with our central hypotheses.

5.2 LDA Content Analysis in AAER-Years

Table 8 displays the results of 75 regressions that treat each of the LDA topic loadings as a dependent variable. We control for year and firm fixed effects, and we focus on the AAER dummy as the independent variable of interest in the first column. We run analogous tests for the pre-AAER dummy and the post-AAER dummy to create the second and third columns. This allows us to examine which topics AAER firms uniquely disclose or under-disclose during periods of fraud. Strong results are those that are statistically significant in the first column, and that are not significant with the same sign in the second and third columns. All standard errors are clustered by firm. Because we control for firm fixed effects, all reported links between fraud-years and LDA factors are conservative and based on within-firm identification.

[Insert Table 8 Here]

The table shows that twelve of the 75 topics are significantly linked to AAER years. These twelve themes are abnormally disclosed by firms involved in fraud relative to the same firms in non-AAER years. A negative coefficient indicates under-disclosure, and a positive coefficient indicates abnormally high levels of disclosure. We note that each topic is well-described by its machine-generated commongram labels as displayed in the first column. These commongram topic labels are generated automatically by the metaHeuristica software (as provided by Ball, Hoberg and Maksimovic (2013)), and are thus not subjected to researcher prejudice. We also note that finding twelve significant topics, some significant at the 1% level, is well beyond what one would expect by chance for 75 topics.

We next interpret the results. The first topic, which includes "offsets" and "primarily due", is a performance attribution topic. The table shows that fraudulent firms disclose less of this attribution text, and thus provide fewer details that would help a reviewer to

evaluate the drivers of the firm's performance. This result is especially consistent with H1, suggesting that firms might provide less attribution text to increase the cost of detecting the fraud. This result might also be consistent with Hypothesis H3 (institutional reviewing methods) if reviewers place more scrutiny on firms that provide less attribution. Although the bar is rather high for Hypothesis H4 in this setting, it is also possible that firms in economic distress also report less attribution text.

The second significant topic focuses on the management team and likely reflects the firm's governance. This topic is under-disclosed by fraudulent firms. This result is consistent with H1, H2, or H3. Relating to H1, fraudulent managers might be less willing to cite their own qualifications, which might be weak relative to individuals that typically hold managerial positions. As a result, managers might feel that under-discussing these attributes would reduce the likelihood of red flags. Relevant to H2, managers might strategically omit information about the firm's weak governance in order to give investors the impression that the firm is well-governed. Also relevant to H1 and H2, managers committing fraud might omit information about themselves to avoid associating their own names with the fraudulent accounting. Regarding H3, the SEC might place more scrutiny on firms that provide little information about the managerial team. We believe that this finding is harder to square with H4, as touting the managerial team's experience would seem equally relevant in all periods if fraud is not occurring and managerial integrity is less at stake.

We also note that neither of these first two topics are significant for pre-AAER or post-AAER years in the latter two columns. Along with the firm fixed effect controls, this finding further reinforces the fact that these deviations from standard disclosure are unique to firms committing fraud.

The third row shows that firms involved in fraud are less likely to disclose legal proceedings or bankruptcy issues, yet they discuss legal issues more often in the year after the fraud is announced. This result is likely due to the fact that firms disclose the fraud itself after it is made public. Consistent with this view, we note that the most significant

individual words that are used more after a fraud is announced include a large number of terms citing the SEC investigation itself (See Table A4 in the Online Appendix to this paper for this evidence).

The fourth row indicates that fraudulent firms tend to under-disclose information relating to financial market liquidity and the sufficiency of their funds to meet ongoing liquidity needs. This result is most consistent with H2 and H4. Regarding H2, it suggests that managers who commit fraud to improve their odds at issuing equity also under-disclose liquidity problems to maximize the likelihood of being successful. Regarding H4, poor liquidity might be an example of poor economic conditions that can motivate firms to commit fraud more broadly. Because the strategic link between fraud and the ability to secure capital at favorable rates is a common theme in the existing literature (See Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010)), we consider direct tests of H2 in this setting using an exogenous liquidity shock later in the next section.

The 5th row relates to acquisitions and is consistent with H1, as managers might focus on complex transactions to increase the cost of fraud detection. Yet this finding might also be consistent with H3 if mergers attract more scrutiny, and with H4 if mergers coincide with periods of poor performance. The elevated disclosure relating to product markets and growth strategies also are potentially consistent with strategic disclosure to achieve specific benefits (H2), as grandstanding products or future growth can give investors false impressions of the firm's prospects, allowing managers to elevate stock prices and potentially issue equity at artificially low cost. This finding is less consistent with H4, as higher growth is not consistent with poor economic conditions.

Overall the results of Table 8 provide much clarity regarding the specific disclosures and mechanisms that drive our empirical results. In many cases, results are potentially consistent with more than one hypothesis. Hence, our results can be seen as strong guidance for future research, both empirical and theoretical, to assess why disclosures relating espe-

cially to attribution of performance, managerial characteristics, financial market liquidity, mergers, and growth strategies might play a role.

We view the results, in balance, as most suggestive of at least some role for strategic motives (H1 and H2), and also a link to economic conditions (H4). Support for H3 seems less likely, especially given the next set of tests on comment letters. Also regarding H4, the most likely economic conditions that might be linked to fraud would include challenges relating to financial constraints and the possibility of slowing growth. In turn, because fraud is by definition strategic, the boundary between H4 and the strategic hypotheses H1 and H2 is not perfectly black and white, and a combination of these factors might be at play. These broad findings support consideration of more specialized tests, which we consider next.

In Online Appendix Table A5, we separately report results for revenue AAERs and expense AAERs. We briefly summarize the results here. Although power is lower for these events, we find that firms committing revenue fraud under-disclose attribution text, litigation and financial market liquidity, and over-disclose information about total revenues relative to non-fraud peers. These results are consistent with stock market investors focusing on revenue performance for high growth firms. Such firms have incentives to grandstand their revenue performance, and to potentially conceal liquidity issues. These results also might be consistent with economic conditions in the form of financial constraints and slowing growth. Regarding expense fraud, firms under-disclose information relating to efforts to reduce costs, economies of scale, and cost cutting in general. They also over-discuss issues relating to research and development, and new products.

5.3 Comment Letters

In this section, we examine the key topics related to the issuance of comment letters by the SEC. In particular, we compare these results to our AAER results. We use the same methodology as in Table 8. This test allows us to more directly challenge hypothesis H3,

which predicts that our above results for AAERs are driven by institutional features of the regulatory review process. Because the comment letter review process is the primary way that verbal disclosure is evaluated by the SEC, this test is quite direct. H3 predicts that the topic themes that dominate our AAER tests should be the same as those that dominate our comment letter tests.

We consider the universe of comment letters from Audit Analytics for which a comment letter was written that referred to the content in the MD&A Section of the 10-K. This data is available from 2005 to 2008. Despite this shorter sample, we have adequate power to examine verbal themes because comment letters are far more common (18.3% of our sample firms) than AAERs (1.5% of our sample). Nevertheless, we acknowledge that the smaller sample size here is a limitation.

[Insert Table 9 Here]

Table 9 displays the results of these tests. The table shows that among the five topics that are significantly related to comment letters, only two overlap with the twelve topics that are significant for AAERs. This level of overlap is rather modest, and thus provides only modest support for H3. The two topics that do overlap suggest that firms that excessively disclose information about new products, and firms that under-disclose information about governance and the managerial team, are both more likely to receive a comment letter and also be involved in an AAER.

Overall, the fact that comment letters and AAERs appear to be distinct from a verbal content perspective, indicates that H3, at best, explains only a small fraction of our results. The most plausible link to H3 is the governance and managerial team topic. However, this specific topic might also relate to other hypotheses, as discussed earlier. In all, the rather modest results for H3, along with the fact that Dyck, Morse and Zingales (2010) find that SEC reviews likely account for very little fraud detection (employees and the media being more relevant), suggest that H3 likely explains little of our results.

5.4 Fog Index

In this section, we test the hypothesis, relating to H1, that 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 H1, 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 $0.4[\frac{\#words}{\#sentences} + \frac{\#complexwords}{\#words}]$, where complex words are those with three or more syllables. We also consider the Automated Readability Index and the Flesch Kinkaid Index for robustness.

[Insert Table 10 Here]

The results are reported in Table 10. The results in Panel A, which are based on the AAER year, reject the hypothesis that managers use complex text when they are involved in AAERs. In contrast, for two of the three indices, the AAER dummy is negative and significant. Interestingly, the coefficient becomes positive and significant only in Panel C, which is based on the post-AAER year. The likely explanation is that once the AAER becomes public, firms disclose the legal implications of the AAER itself, and the use of legal jargon likely increases the difficulty of reading the document and hence the various fog indices.

Overall, these tests reject the specific hypothesis relating to H1 that firms use complex language to obfuscate the interpretation of their disclosures.

5.5 Individual Words

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 restatements, which indicates a history of poor accounting beyond the AAER itself. We also observe that AAER firms disclose more information about acquisitions and international vocabulary including region and country names such as Africa and Brazil. It is possible that more difficult to trace international transactions might facilitate fraudulent accounting. Firms involved in AAERs also disclose more vocabulary indicative of uncertainty and speculation: "believe", "feasibility", "fluctuating", and "instability".

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 report single word results for pre-AAER and post-AAER firms in the Online Appendix tables A3 and A4. These tables also confirm that AAER years are unique. Table A4 confirms that firms involved in AAERs disclose information about the AAER itself after the AAER investigation is made public. We also present a list of the top 25 most representative AAERs in Table A6, which lists the AAERs that have the highest fraud similarity scores.

6 Equity Market Liquidity

In this section, we examine the link between fraudulent firm disclosure, equity market liquidity and equity issuance. These tests examine the specific hypothesis that managers might commit fraud to get access to an artificially lower cost of capital (see Dechow, Sloan and Sweeney (1996), Povel, Singh and Winton (2007) and Wang, Winton and Yu (2010)). We consider whether, following exogenous negative shocks specifically to equity market liquidity, managers are more likely to commit fraud and to produce disclosure with a higher fraud similarity score, potentially to inflate their odds of issuing equity.

We consider the Edmans, Goldstein, and Jiang (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 moderately long lasting, perhaps one to 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.

[Insert Table 11 Here]

The results are presented in Table 11, Panel A (industry and year fixed effects) and Panel B (firm and year fixed effects). Both panels support our prediction that negative shocks to equity market liquidity, through the lens of the mutual fund selling instrument, 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. These results are highly significant and robust to both industry and firm fixed effects.

In panel C, we examine regressions in which the dependent variable is equity issuance, and the key independent variable is the fraud profile similarity. 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 potentially 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.

7 Conclusions

We consider four hypotheses predicting common disclosures in the MD&A Section of the 10-K among firms committing fraud. The first two relate to strategic disclosure (to conceal fraud, or to achieve fraud-driven objectives). The third relates to whether the process of fraud detection used by institutions such as the Securities and Exchange Commission, which can also generate observed common disclosures. The fourth relates to whether common economic conditions faced by firms committing fraud. In many regards, these hypotheses are empirically indistinguishable.

We first examine if firms committing fraud produce disclosure that (1) is highly similar to industry peers or (2) that is different from industry peers but similar to other firms committing fraud. Our results strongly favor the second conclusion, and we find that similarity to a cluster of vocabulary unique to firms committing fraud strongly predicts observed fraud both in sample and out of sample. The results are economically large, as being in the lowest decile regarding a firm's use of this vocabulary predicts fraud at a rate of 0.5%, and being in the highest decile predicts fraud at a rate of 3.5% overall. In contrast, we do not find support for the conclusion that fraudulent firms cluster strongly with industry peers. These results are particularly striking along one dimension. We find results for firms involved in AAERs even when compared to the same firms before and after the AAER. These results suggest that disclosures are revised materially as a firm evolves from a pre-AAER firm, to a firm involved in AAER actions, and to a firm that has been revealed as allegedly committing fraud. Content analysis reveals much granularity regarding the discussions firms disclose over this cycle.

Having established the presence of a strong common verbal signature among fraudulent firms, we turn our attention to content analysis to refine our ability to test the four

aforementioned hypotheses. These tests reveal a link between fraudulent firms and the under-reporting of managerial characteristics, financial liquidity, and text that explains performance in detail. In contrast, fraudulent firms over-report mergers and acquisitions, new product introductions, and growth strategies. These results reveal specific mechanisms that drive common verbal content among fraudulent firms, and greatly reduce the set of specific hypotheses that can explain our results. They also provide motivation for future researchers, theoretical and empirical, to further assess these channels and their roots.

We note three additional findings. First, we find little overlap in the verbal content of firms involved in AAERs and those receiving comment letters from the SEC indicating problems with a firm's MD&A. Second, we find no positive link between various fog indices and the nature of verbal disclosure by fraudulent firms. Third, we find that negative exogenous shocks to equity market liquidity are associated with increased incidence rates of fraud, and also increased use of the vocabulary that is common among fraudulent firms.

Overall, our findings provide least support for the conclusion that our results are driven by details in how fraud is detected by institutions like the SEC (H3). We also find no support for the conclusion that managers strategically use either excessively complex text or that they herd with industry peers to increase the cost of detection (H1). Our results are most consistent with strategic disclosure to achieve fraud driven motives such as attaining an artificially low cost of capital (H2), and also with a potential role for economic conditions faced by fraudulent firms generating common disclosures (H4). The specific economic conditions that might matter most are financial constraints and slowing growth.

References

Antweiler, Werner, and Murray Frank, 2004, Is all that talk just noise? The information content of internet stock message boards, *Journal of Finance* 52, 1259–1294.

Ball, Christopher, and Gerard Hoberg, and Vojislav Maksimovic, 2013, Disclosure Informativeness and the Tradeoff Hypothesis: A Text-Based Analysis, University of Maryland Working Paper.

Beck, Thorsten, Asli Demirguc-Kunt, Asli and Vojislav Maksimovic, Vojislav,2005, Financial and Legal Constraints to Growth: Does Firm Size Matter? *Journal of Finance* 60, 137–77.

Beneish, Messod, 1997, Detecting GAAP Violation: Implications for Assessing Earnings Management among Firms with Extreme Financial Performance, *Journal of Accounting and Public Policy* 16, 271–309.

Beneish, Messod, 1999, The Detection of Earnings Manipulation, *Financial Analysts Journal* 55, 24–36.

Beneish, Messod, 1999, Incentives and Penalties Related to Earnings Overstatements that Violate GAAP, *The Accounting Review* 74, 425–457.

Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003, Latent Dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1002.

Brown, Stephen V., and Jennifer Wu Tucker, 2011, Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications, *Journal of Accounting Research* 49, 309–346.

Bryan, S. H., 1997, Incremental Information Content of Required Disclosures Contained in Management Discussion and Analysis, *The Accounting Review* 72, 285–301.

Cimiano, Phillip, 2010, *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*, Springer, New York.

Cole, C. J., and C. L. Jones, 2005, Management Discussion and Analysis: A Review and Implications for Future Research, *Journal of Accounting Literature* 24, 135–74.

Darrough, Masako N., 1993, Disclosure policy and competition: Cournot vs. Bertrand, *Accounting Review* 534–561.

Dechow, Patricia, and Weili Ge, and Chad Larson, and Richard Sloan, 2011, Predicting Material Accounting Misstatements, *Contemporary Accounting Research* 28, 17–82.

Dechow, Patricia, and Weili Ge, and Catherine Schrand, 2010, Understanding earnings quality: A review of the proxies, their determinants and their consequences, *Journal of Accounting and Economics* 2, 344–401.

Dechow, Patricia, and Richard Sloan, and Amy Sweeney, 1996, Causes and Consequences of Earnings Manipulation: An Analysis of Firms Subject to Enforcement Actions by the SEC, *Contemporary Accounting Research* 13, 1–36.

Devenow, Andrea, and Ivo Welch, 1996, Rational Herding in Financial Economics, *European Economic Review* 40, 603–615.

Dye, Ronald A., and Sri S. Sridhar, 1995, Industry-wide disclosure dynamics, *Journal of accounting research* 157–174.

Dyck, Alexander, and Adair Morse, and Luigi Zingales, 2010, Who Blows the Whistle on Corporate Fraud?, *Journal of Finance* 65, 2213–53.

Edmans, A., I. Goldstein, and W. Jiang, 2012, The Real Effects of Financial Markets: The Impact of Prices on Takeovers, *The Journal of Finance* 67, 933–971.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, Management's Tone Change, Post Earnings Announcement Drift and Accruals, *Review of Accounting Studies* 15, 915–53.

Feroz, Ehsan, and Kyungjoo Park, and Vector Pastena, 1991, The Financial and Market Effects of the SEC's Accounting and Auditing Enforcement Releases, *Journal of Accounting Research* 29, 107–142.

Hanley, Kathleen, and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review of Financial Studies* 23, 2821–2864.

Hanley, Kathleen, and Gerard Hoberg, 2012, Litigation risk and the underpricing of initial public offerings, *Journal of Financial Economics* 103, 235–254.

Hoberg, Gerard, and Vojislav Maksimovic, 2012, Redefining Financial Constraints: a Text-Based Analysis, *University of Maryland Working Paper*.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies in mergers and acquisitions: A text based analysis, *Review of Financial Studies* 23, 3773–3811.

Hoberg, Gerard, and Gordon Phillips, 2012, New dynamic product based industry classifications and endogenous product differentiation, *University of Maryland Working Paper*.

Hughes, Patricia J., and Anjan V. Thakor, 1992, Litigation risk, intermediation, and the underpricing of initial public offerings, *Review of Financial Studies* 5, 709–742.

Karpoff, Jonathan, and Scott Lee and Gerald Martin, 2008, The Consequences to Managers for Financial Misrepresentation, *Journal of Financial Economics* 88, 193–215.

Karpoff, Jonathan, and Scott Lee and Gerald Martin, 2008, The Cost to Firms of Cooking the Books, *Journal of Quantitative and Financial Analysis* 43, 581–612.

Kedia, Simi, and Thomas Philippon, 2009, The Economics of Fraudulent Accounting, *Review of Financial Studies* 22, 2169–2199.

Kedia, Simi, and Shiva Rajgopal, 2011, Do the SEC's Enforcement Preferences Affect Corporate Misconduct?, *Journal of Accounting and Economics* 51, 259–278.

Kothari, S. P., Xu Li, and James E. Short, 2009, The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis, *The Accounting Review* 84, 1639–70.

Li, Feng, 2010, Information Content of the Forward-Looking Statements in Corporate Filings: A Naive Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

Li, Feng, 2008, Annual Report Readability, Current Earnings, and Earnings Persistence, *Journal of Accounting and Economics* 45, 221–47.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-ks, *Journal of Finance* 66, 35–65.

Povel, Paul, and Rajdeep Singh, and Andrew Winton, 2007, Booms, Busts, and Fraud, *Review of Financial Studies* 20, 1219–1254.

Sebastiani, Fabrizio, 2002, Machine learning in automated text categorization, *acmcs*.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Wang, Tracy, 2013, Corporate Securities Fraud: Insights from a New Empirical Framework, *Journal of Law and Economics* Forthcoming.

Wang, Tracy, Andrew Winton, Xiaoyun Yu, 2010, Corporate Fraud and Business Conditions: Evidence from IPOs, *Journal of Finance* 65, 2255–2292.

Table 1: Summary Statistics

Summary statistics are reported for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the 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. These peers are identified by sorting firms in each two digit SIC code first into above and below median firm sizes, and then into above and below median firm ages for each group. Median size and age are computed separately for each year. 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. R&D/sales and CAPX/sales are Compustat values of R&D and capital expenditures scaled by sales. All ratios are winsorized at the 1% and 99% level, and any values of operating income/sales less than minus one are set to minus one.

Variable	Mean	Std. Dev.	Minimum	Median	Maximum
Panel A: Data on Payout Status and Cash Holdings					
AAER Dummy	0.015	0.120	0.000	0.000	1.000
Industry Similarity Score	0.667	0.080	0.410	0.671	0.839
Fraud Similarity Score	0.002	0.077	-0.191	-0.002	0.251
Log Sales	4.917	2.127	0.001	4.866	12.326
Operating Income/Sales	-0.006	0.353	-1.000	0.081	0.703
R&D/Sales	0.190	0.770	0.000	0.000	11.230
CAPX/Sales	0.123	0.345	0.000	0.037	9.276

Table 2: Pearson Correlation Coefficients

Pearson Correlation Coefficients are reported for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. 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
			Correlation Coefficients				
(1)	Fraud Similarity Score	0.082					
(2)	Industry Similarity Score	0.026	0.090				
(3)	Log Sales	0.070	-0.005	0.061			
(4)	Operating Income/Sales	0.022	-0.026	-0.040	0.522		
(5)	R&D/Sales	-0.012	0.044	0.081	-0.302	-0.518	
(6)	CAPX/Sales	-0.010	-0.002	0.042	-0.145	-0.156	0.195

Table 3: AAER Timeseries Statistics

The table reports time series statistics for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the the current year.

Row	Year	Number AAER Firm Years	Number of Firms in Sample	Fraction AAER Firm Years
1	1997	28	4670	0.006
2	1998	48	4663	0.010
3	1999	80	4727	0.017
4	2000	110	4647	0.024
5	2001	125	4406	0.028
6	2002	104	4173	0.025
7	2003	80	4009	0.020
8	2004	68	3915	0.017
9	2005	46	3522	0.013
10	2006	17	3396	0.005
11	2007	10	3420	0.003
12	2008	4	3491	0.001

Table 4: AAERs versus Fraud Similarities and Industry Similarity Deciles

The table displays decile statistics for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. 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). 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 (1997-2008)		
1	-0.124	0.005	0.514	0.010
2	-0.076	0.007	0.585	0.012
3	-0.050	0.008	0.617	0.012
4	-0.030	0.011	0.641	0.012
5	-0.011	0.010	0.662	0.012
6	0.007	0.011	0.682	0.015
7	0.027	0.014	0.702	0.017
8	0.050	0.020	0.724	0.013
9	0.081	0.023	0.750	0.017
10	0.147	0.037	0.792	0.027
		Panel B: Out of Sample (2002-2008)		
0	-0.112	0.007	0.519	0.009
1	-0.069	0.008	0.587	0.008
2	-0.045	0.009	0.616	0.008
3	-0.026	0.014	0.639	0.009
4	-0.010	0.012	0.657	0.010
5	0.007	0.010	0.676	0.013
6	0.024	0.008	0.696	0.016
7	0.044	0.018	0.718	0.011
8	0.070	0.017	0.745	0.018
9	0.129	0.022	0.789	0.024

Table 5: Disclosure Outcome Regressions (AAER-year)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. 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 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 25,926 annual firm observations from 2002 to 2008. 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 replaces the firm fixed effects with SIC-2 industry fixed effects.

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.	0.029	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(6.58)	(–0.01)	(1.40)	(2.68)	(5.26)	(11.63)	(–3.66)	
(2)	Industry Similarity	0.003	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(0.77)	(–4.87)	(3.86)	(2.32)	(9.41)	(–2.65)	(2.88)	
				Panel B: Above Median Firm Size Only					
(3)	Fraud Profile Sim.	0.026	0.020	0.046	0.012	0.008	0.004	–0.001	24,523
		(4.93)	(3.57)	(2.49)	(3.58)	(4.76)	(8.37)	(–2.14)	
(4)	Industry Similarity	–0.001	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(–0.13)	(–0.41)	(–0.32)	(–0.52)	(2.26)	(–2.88)	(1.35)	
				Panel C: Below Median Firm Size Only					
(5)	Fraud Profile Sim.	0.031	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(3.44)	(–1.46)	(1.06)	(1.84)	(3.95)	(7.92)	(–3.02)	
(6)	Industry Similarity	0.008	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(1.09)	(–6.13)	(3.99)	(2.85)	(8.77)	(–1.58)	(2.46)	
				Panel D: Same as Panel A, but Out of Sample Years Only					
(7)	Fraud Profile Sim.	0.014	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(2.31)	(–1.66)	(1.31)	(1.95)	(2.10)	(4.03)	(–3.22)	
(8)	Industry Similarity	0.008	–0.014	0.003	0.008	0.009	–0.002	0.000	25,926
		(1.41)	(–3.56)	(2.47)	(2.72)	(5.80)	(–2.49)	(1.40)	
				Panel E: Same as Panel A, but Add Additional Controls					
(9)	Fraud Profile Sim.	0.029	0.002	0.001	0.004	0.004	0.003	–0.001	49,039
		(6.76)	(0.65)	(1.33)	(2.75)	(4.36)	(11.73)	(–3.49)	
(10)	Industry Similarity	0.003	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(0.78)	(–4.21)	(3.68)	(2.37)	(8.63)	(–2.71)	(3.06)	
				Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects					
(11)	Fraud Profile Sim.	0.049	0.003	0.004	-0.003	0.001	0.006	-0.000	49,745
		(9.00)	(1.21)	(5.08)	(-2.25)	(2.21)	(14.33)	(-1.23)	0.031
(12)	Industry Similarity	0.005	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(1.04)	(-6.27)	(11.23)	(2.64)	(19.45)	(2.45)	(2.65)	0.198

Table 6: Disclosure Outcome Regressions (Pre-AAER Disclosures)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. 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 Future AAER dummy is our primary variable of interest and is one if the firm was involved in fraudulent activity in the the year after the current year of the observation. 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 25,926 annual firm observations from 2002 to 2008. 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 five 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 replaces the firm fixed effects with SIC-2 industry fixed effects.

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.	–0.001	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(–0.18)	(–0.07)	(1.43)	(2.77)	(5.53)	(11.68)	(–3.73)	
(2)	Industry Similarity	–0.004	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(–0.71)	(–4.87)	(3.87)	(2.33)	(9.45)	(–2.65)	(2.88)	
				Panel B: Above Median Firm Size Only					
(3)	Fraud Profile Sim.	–0.001	0.020	0.047	0.013	0.009	0.004	–0.001	24,523
		(–0.12)	(3.53)	(2.51)	(3.61)	(4.96)	(8.66)	(–2.35)	
(4)	Industry Similarity	–0.010	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(–1.41)	(–0.42)	(–0.33)	(–0.52)	(2.25)	(–2.89)	(1.38)	
				Panel C: Below Median Firm Size Only					
(5)	Fraud Profile Sim.	–0.005	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(–0.39)	(–1.52)	(1.09)	(1.90)	(4.10)	(7.90)	(–3.02)	
(6)	Industry Similarity	0.001	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(0.12)	(–6.15)	(4.00)	(2.87)	(8.83)	(–1.59)	(2.46)	
				Panel D: Entire Sample (Out of Sample Years Only)					
(7)	Fraud Profile Sim.	0.001	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(0.05)	(–1.69)	(1.32)	(2.00)	(2.19)	(4.06)	(–3.23)	
(8)	Industry Similarity	–0.016	–0.014	0.003	0.008	0.010	–0.002	0.000	25,926
		(–0.78)	(–3.57)	(2.47)	(2.74)	(5.82)	(–2.47)	(1.40)	
				Panel E: Same as Panel A, but Add Additional Controls					
(9)	Fraud Profile Sim.	0.001	0.001	0.001	0.004	0.004	0.003	–0.001	49,039
		(0.22)	(0.59)	(1.35)	(2.84)	(4.61)	(11.79)	(–3.56)	
(10)	Industry Similarity	–0.002	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(–0.34)	(–4.21)	(3.69)	(2.38)	(8.67)	(–2.70)	(3.06)	
				Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects					
(11)	Fraud Profile Sim.	0.010	0.002	0.004	-0.003	0.001	0.006	-0.000	49,745
		(1.24)	(1.15)	(5.07)	(-2.14)	(2.98)	(14.38)	(-1.36)	0.025
(12)	Industry Similarity	-0.001	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(-0.15)	(-6.28)	(11.23)	(2.64)	(19.54)	(2.48)	(2.64)	0.198

Table 7: Disclosure Outcome Regressions (Post-AAER Disclosures)

In Panels A to C, the table reports our baseline OLS regressions for our sample of 49,039 firm-year observations based on annual firm observations from 1997 to 2008. 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 Past AAER dummy is our primary variable of interest and is one if the firm was involved in fraudulent activity in the the year prior to the current year of the observation. 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 25,926 annual firm observations from 2002 to 2008. 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 replaces the firm fixed effects with SIC-2 industry fixed effects.

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.	–0.018	0.000	0.002	0.004	0.005	0.003	–0.001	49,039
		(–3.10)	(–0.09)	(1.43)	(2.76)	(5.55)	(11.68)	(–3.71)	
(2)	Industry Similarity	0.001	–0.012	0.003	0.004	0.009	–0.001	0.000	49,039
		(0.25)	(–4.87)	(3.87)	(2.33)	(9.45)	(–2.64)	(2.87)	
				Panel B: Above Median Firm Size Only					
(3)	Fraud Profile Sim.	–0.009	0.020	0.047	0.012	0.009	0.004	–0.001	24,523
		(–1.36)	(3.52)	(2.51)	(3.61)	(4.96)	(8.65)	(–2.34)	
(4)	Industry Similarity	0.008	–0.003	–0.005	–0.002	0.004	–0.001	0.000	24,523
		(1.17)	(–0.41)	(–0.31)	(–0.52)	(2.26)	(–2.87)	(1.35)	
				Panel C: Below Median Firm Size Only					
(5)	Fraud Profile Sim.	–0.026	–0.004	0.001	0.003	0.005	0.003	–0.001	24,516
		(–2.58)	(–1.54)	(1.09)	(1.90)	(4.13)	(7.89)	(–3.00)	
(6)	Industry Similarity	–0.008	–0.017	0.004	0.005	0.011	0.000	0.000	24,516
		(–0.80)	(–6.15)	(4.00)	(2.87)	(8.84)	(–1.60)	(2.46)	
				Panel D: Entire Sample (Out of Sample Years Only)					
(7)	Fraud Profile Sim.	–0.011	–0.006	0.002	0.004	0.003	0.003	–0.001	25,926
		(–1.91)	(–1.70)	(1.33)	(1.98)	(2.21)	(4.04)	(–3.22)	
(8)	Industry Similarity	–0.005	–0.014	0.003	0.008	0.010	–0.002	0.000	25,926
		(–0.78)	(–3.58)	(2.47)	(2.73)	(5.85)	(–2.48)	(1.40)	
				Panel E: Same as Panel A, but Add Additional Controls					
(9)	Fraud Profile Sim.	–0.018	0.001	0.001	0.004	0.004	0.003	–0.001	49,039
		(–3.26)	(0.58)	(1.36)	(2.83)	(4.62)	(11.78)	(–3.54)	
(10)	Industry Similarity	0.000	–0.010	0.003	0.004	0.008	–0.001	0.001	49,039
		(0.04)	(–4.21)	(3.69)	(2.38)	(8.67)	(–2.70)	(3.06)	
				Panel F: Same as Panel A, but Replace Firm Effects with Industry Effects					
(11)	Fraud Profile Sim.	0.011	0.003	0.004	-0.003	0.001	0.006	-0.000	49,745
		(2.02)	(1.16)	(5.08)	(-2.14)	(2.97)	(14.38)	(-1.37)	0.025
(12)	Industry Similarity	0.003	-0.013	0.009	0.005	0.008	0.001	0.000	49,745
		(0.62)	(-6.27)	(11.23)	(2.65)	(19.51)	(2.48)	(2.64)	0.198

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 (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, Hoberg and Maksimovic (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	partially offset, primarily due, offset decrease, due primarily, decreased decrease	-0.173 (-3.55)		
2	board directors, executive officers, officers directors, vice president, directors officers	-0.196 (-3.20)		
3	legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	-0.185 (-2.73)	-0.172 (-2.32)	0.222 (2.89)
4	sufficient meet, additional financing, sources liquidity, raise additional, additional funds	-0.115 (-2.63)		
5	acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	0.157 (2.61)		-0.134 (-2.40)
6	marketing expenses, professional fees, salaries benefits, expenses related, related expenses	-0.102 (-2.49)		0.107 (2.25)
7	product line, product lines, product sales, distribution channels, product introductions	0.086 (2.23)		
8	cash flow, cash flows, cash cash equivalents, cash provided, cash investing activities	-0.119 (-2.14)		
9	payments made, principal payments, payment dividends, pay dividends, dividends paid	-0.090 (-2.06)		0.149 (2.63)
10	gain sale, held sale, sale leaseback, gains sale, realized gains	-0.114 (-2.03)		
11	continued growth, business strategy, growth strategy, business opportunities, core business	0.126 (2.03)		
12	clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	-0.020 (-2.02)		
13	license fees, consulting services, consulting fees, service fees, services provided		-0.145 (-3.10)	
14	past years, recent years, significant portion, substantial portion, years company		-0.160 (-2.95)	
15	senior notes, principal amount, notes payable, subordinated notes, senior subordinated notes		0.219 (2.76)	
16	laws regulations, government regulation, federal state, government agencies, change control		-0.158 (-2.52)	
17	generally accepted, conducted audits accordance generally accepted auditing, based, principles significant estimates made management		-0.138 (-2.39)	
18	life insurance, premiums written, insurance premiums, premiums earned, insurance coverage		-0.119 (-2.35)	
19	foreign currency, foreign exchange, north america, currency exchange, domestic international		0.130 (2.02)	-0.103 (-2.08)
20	restructuring charge, restructuring charges, write downs, special charges, fourth quarter			0.259 (3.68)
21	interest rates, certificates deposit, asset liability, assets liabilities, balance sheet			-0.115 (-2.03)
22	entered agreement, agreement dated, terms agreement, pursuant terms, agreement entered			0.111 (2.02)

Table 9: LDA Topics Associated with Comment Letters (Compared to AAERs)

The table lists the Topic Model Factors found to be statistically significant regarding their link to firms receiving comment letters as compared to firms not receiving comment letters (first column after Topic Descriptions). Our sample is restricted to the years 2005 to 2008 for the comment letter tests due to comment letter data availability. We also report significant topics for firms involved in AAERs (last column). The table displays coefficients and t-statistics for regressions where firm-year topic loadings are regressed on the comment letter dummy and the AAER dummy, respectively. We only report results for the 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (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 Comment Letter AAER years	Different in AAER years
1	effective january, september september, effective july, effective september, effective october	-0.029 (-2.39)	
2	short term, long term, long term debt, short term borrowings, short term investments	0.037 (2.31)	
3	accounts receivable, accounts payable, doubtful accounts, accounts payable accrued, accounts receivable inventory	0.027 (2.17)	
4	product line, product lines, product sales, distribution channels, product introductions	0.020 (2.13)	0.086 (2.23)
5	board directors, executive officers, officers directors, vice president, directors officers	-0.042 (-2.06)	-0.196 (-3.20)
6	partially offset, primarily due, offset decrease, due primarily, decreased decrease		-0.173 (-3.55)
7	legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights		-0.185 (-2.73)
8	sufficient meet, additional financing, sources liquidity, raise additional, additional funds		-0.115 (-2.63)
9	acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted		0.157 (2.61)
10	marketing expenses, professional fees, salaries benefits, expenses related, related expenses		-0.102 (-2.49)
11	cash flow, cash flows, cash cash equivalents, cash provided, cash investing activities		-0.119 (-2.14)
12	payments made, principal payments, payment dividends, pay dividends, dividends paid		-0.090 (-2.06)
13	gain sale, held sale, sale leaseback, gains sale, realized gains		-0.114 (-2.03)
14	continued growth, business strategy, growth strategy, business opportunities, core business		0.126 (2.03)
15	clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals		-0.020 (-2.02)

Table 10: Fog Index Regressions

The table reports OLS regressions for our sample of observations based on annual firm observations from 1997 to 2008. One observation is one firm in one year. The dependent variable is a fog index or readability index as noted in the first column. All three readability indices are constructed such that a higher value indicates greater difficulty in reading. Panels A to C differ in how the AAER Dummy is lagged. The AAER dummy in Panel A is one if an AAER action indicates that the firm was involved in fraudulent activity in the the current year. In Panel B, the AAER dummy is one in the year prior to a year in which a given firm was involved in an AAER, and In Panel C, the AAER dummy is one in the year after a given firm was involved in an AAER. See Table 1 for the description of our key variables. All regressions are estimated with year and firm fixed effects, and standard errors are clustered by firm. t-statistics are in parentheses.

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: AAER-year Results					
(1)	Automated Read. Index	–0.128	–0.276	0.022	0.054	0.083	–0.013	–0.004	49,039
		(–1.74)	(–6.01)	(1.42)	(1.94)	(4.46)	(–3.48)	(–1.20)	
(2)	Gunning Index	–0.183	–0.157	0.004	0.050	0.014	–0.011	–0.001	49,039
		(–2.88)	(–3.97)	(0.29)	(2.02)	(0.91)	(–3.50)	(–0.26)	
(3)	Flesch Kinkaid Index	–0.288	–0.628	0.093	0.302	0.293	–0.010	0.009	49,039
		(–0.97)	(–3.91)	(1.77)	(3.05)	(4.44)	(–0.80)	(0.72)	
				Panel B: pre-AAER-year Results					
(4)	Automated Read. Index	–0.120	–0.275	0.022	0.054	0.082	–0.013	–0.004	49,039
		(–1.08)	(–5.99)	(1.41)	(1.92)	(4.39)	(–3.51)	(–1.18)	
(5)	Gunning Index	–0.047	–0.156	0.003	0.050	0.013	–0.012	–0.001	49,039
		(–0.47)	(–3.95)	(0.28)	(1.99)	(0.80)	(–3.54)	(–0.24)	
(6)	Flesch Kinkaid Index	0.275	–0.627	0.093	0.301	0.291	–0.010	0.009	49,039
		(0.69)	(–3.91)	(1.76)	(3.04)	(4.40)	(–0.82)	(0.73)	
				Panel C: Post-AAER-year Results					
(7)	Automated Read. Index	0.241	–0.275	0.022	0.054	0.082	–0.013	–0.004	49,039
		(2.68)	(–5.98)	(1.41)	(1.93)	(4.39)	(–3.49)	(–1.20)	
(8)	Gunning Index	0.211	–0.156	0.003	0.050	0.013	–0.012	–0.001	49,039
		(2.72)	(–3.93)	(0.27)	(2.00)	(0.79)	(–3.52)	(–0.25)	
(9)	Flesch Kinkaid Index	0.689	–0.624	0.093	0.301	0.290	–0.010	0.009	49,039
		(2.26)	(–3.89)	(1.76)	(3.05)	(4.39)	(–0.81)	(0.72)	

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 2008. 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. The regressions include industry and year fixed effects in Panels A and C, and firm and year fixed effects in Panels B and D. 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 C 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	0.007	-0.000	0.003	-0.005	0.000	0.007	-0.000	30,683
		(7.45)	(-0.10)	(2.09)	(-2.18)	(0.08)	(2.70)	(-0.66)	0.032
(2)	AAER Dummy	0.006	-0.004	-0.000	0.004	0.006	0.001	-0.000	30,683
		(6.04)	(-0.80)	(-0.15)	(1.60)	(3.32)	(5.58)	(-0.48)	0.022
				Panel B: Firm and Year Fixed Effects					
(3)	Fraud Score	0.003	0.000	0.001	0.005	0.006	0.003	-0.001	30,683
		(5.16)	(0.11)	(0.83)	(2.18)	(4.09)	(10.38)	(-2.96)	0.012
(4)	AAER Dummy	0.005	-0.010	0.001	0.000	0.007	0.001	-0.000	30,683
		(3.97)	(-1.47)	(0.48)	(0.16)	(1.97)	(1.74)	(-0.08)	0.002

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	0.085	-0.040	0.005	0.033	-0.037	0.017	-0.004	30,683
		(4.59)	(-3.45)	(1.07)	(4.32)	(-10.18)	(8.67)	(-3.90)	0.064
(6)	SDC Public SEO Issuance	0.051	0.017	0.002	0.018	-0.006	0.005	-0.000	30,683
		(4.67)	(2.52)	(0.74)	(3.59)	(-3.06)	(7.66)	(-0.41)	0.012

Figure 1: Empirical distribution of firm Fraud Similarities. The distribution is based on our entire sample including both firms that were involved in AAERs and firms that were not. 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.

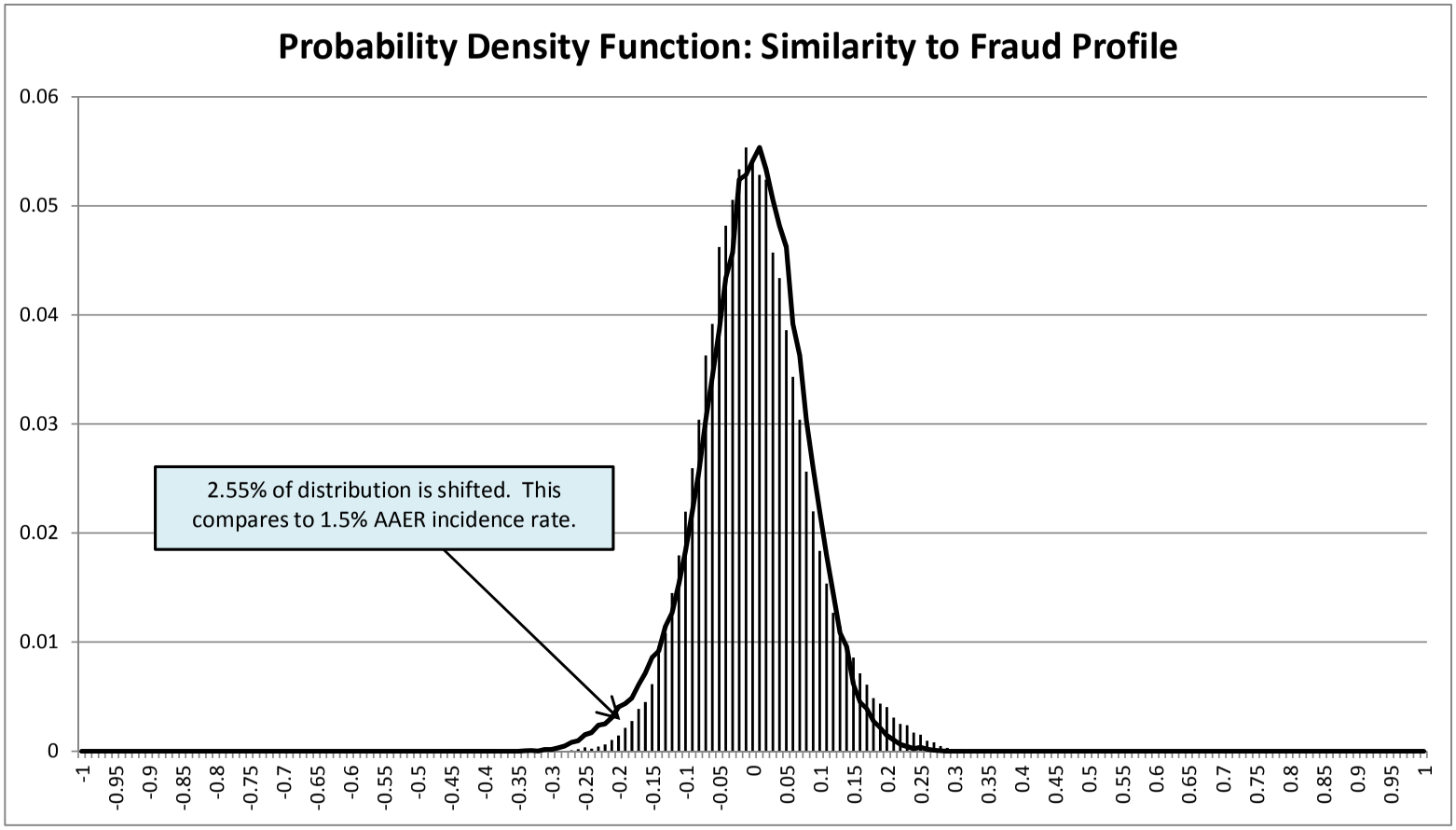
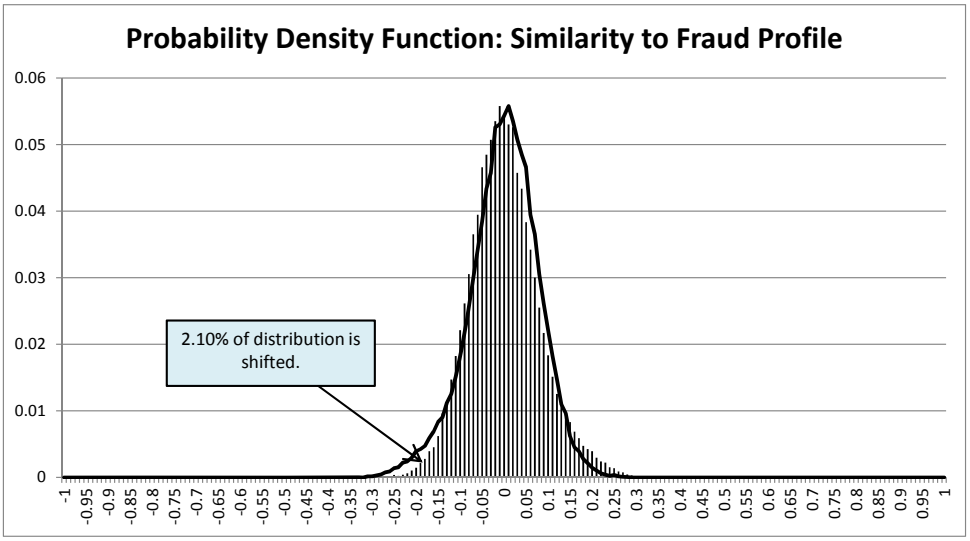


Figure 2: Empirical distribution of firm Fraud Similarities for two subsamples. The upper figure's distribution is based on all firms in our sample excluding firm years involved in AAERs. The lower figure reports the fraud similarity distribution only for firms-years involved in AAERs. In both figures, 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.



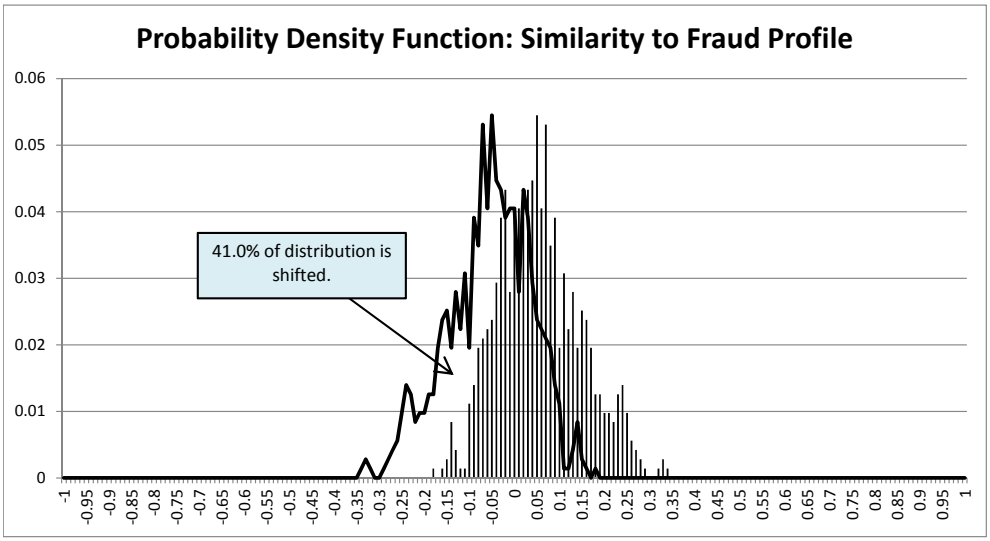


Figure 3: Average Fraud Similarities over time for firms involved in AAERs. The figure displays the average fraud similarity score during the period of time that the AAER alleges fraud occurred, and also during the period of time preceding and after the period of the alleged fraud. Regardless of duration 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.