Do Fraudulent Firms Engage in Disclosure Herding?

Gerard Hoberg* and Craig Lewis*

December 1, 2013

*The University of Maryland, and the Securities and Exchange Commission and Vanderbilt University, respectively. We thank Christopher Ball, Kathleen Hanley, Vojislav Maksimovic, and Harvey Westbrook for excellent comments and suggestions. We also thank seminar participants at the Securities and Exchange Commission. Any remaining errors are ours alone. The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statement by any of its employees. The views expressed herein are those of the authors and do not necessarily reflect the views of the Commission or of the authors' colleagues on the staff of the Commission.

Do Fraudulent Firms Engage in Disclosure Herding?

ABSTRACT

We present two new hypotheses regarding the strategic qualitative disclosure choices of firms involved in potentially fraudulent activity. First, these firms have incentives to herd with industry peers in order to escape detection. Second, these firms have incentives to locally anti-herd with the same peers on specific aspects of disclosure consistent with achieving fraud-driven objectives. We use text-based analysis of firm disclosures and compare disclosures across firms involved in SEC enforcement actions to benchmarks based on industry, size and age, and also to each firm's own disclosure before and after SEC alleged violations. We find especially strong support for the conclusion that firms involved in alleged fraud anti-herd with industry peers on localized disclosure dimensions. We find moderately strong support for the conclusion that they herd with industry peers on broader disclosure dimensions. Content analysis then reveals key vocabulary terms used by firms involved in enforcement actions, and suggests that these firms use complexity to potentially conceal fraudulent activity, and these firms often discuss issues relating to uncertainty, litigation, and speculative statements.

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 how a firm's 10-K MD&A disclosure to the SEC is influenced by this decision. This issue should be particularly salient to managers committing fraud, as this disclosure is submitted to the Securities and Exchange Commission, which is tasked with identifying and pursuing enforcement actions against firms that commit fraud.

This tension, when coupled with the desire to achieve an objective, creates opposing forces regarding how a firm might structure its qualitative disclosure, which offers managers a high degree of discretion regarding the specific information they choose to disclose. On one hand, fear of detection might lead managers committing fraud to herd with their industry peers, as managers might believe that disclosure that appears more standard might reduce the risk of red flags and possible regulator attention.

However, herding with industry peers on all dimensions might reduce the firm's ability to maximize the very objectives that led it to commit fraud in the first place. For example, a firm committing revenue fraud in order to entice investors to provide low cost equity might benefit from under-disclosing information relating to its weak liquidity position in the issuance market. Furthermore, a firm committing fraud may need to deviate from its industry peers on specific issues to mask the fraud itself. For example, such a firm might mask what is ultimately fabricated revenue by discussing overly complex foreign transactions. Herding with industry peers on this dimension

[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.

would make the fabrication too self-evident. Anti-herding on complexity dimensions can serve to raise the cost of review from the regulator's perspective. Central to this hypothesis is that the incentives to anti-herd are localized to specific topics directly related to fraud motives.

We refer to these two apparently competing hypotheses as the industry herding hypothesis and the localized anti-herding hypothesis. We assign these labels because the incentives to herd relate directly to industry peers, as SEC firm reviews are traditionally done in twelve industry groups.[2] Analogously, the anti-herding hypothesis is more localized, and the incentives to anti-herd relate to specific disclosure items such as the discussion of revenues, expenses, or financial market liquidity. Because the incentives to herd are broad and the incentives to anti-herd are localized to specific disclosures, it also is possible that firms engaged in fraudulent behavior engage in both practices at the same time.

We construct an empirical framework that examines both hypotheses, and also allows for the possibility that these hypotheses can be separated in the data. Our analysis of the industry herding hypothesis relies on measures of the extent to which each firm's raw disclosure is similar in word usage distributions to the raw disclosure of other firms of similar size and age in its industry. Analogously, our analysis of the localized anti-herding hypothesis first entails purging each firm's disclosure of content that is related to its industry, size, and age. We then examine whether firms involved in alleged fraud have incremental disclosures that have common anti-herding components. In particular, we ask if firms have these common anti-herding components when they are involved in SEC accounting and auditing enforcement actions, but not when they are not involved in these enforcement actions.

[2]See http://www.sec.gov/divisions/corpfin/cffilingreview.htm for details.

We examine these hypotheses using two differences. We compare firms involved in accounting and auditing enforcement actions (AAERs) to firms not involved in AAERs in cross section, and we also compare the disclosure of firms involved in AAERs to the disclosure of the same firms before and after the dates of alleged fraud indicated in the AAERs. In all, we find especially strong support for the localized anti-herding hypothesis, and moderately strong support for the industry herding hypothesis.

Our study relies heavily on text analytic methods and the cosine similarity method. The approach first uses cosine similarities directly to examine if each firm has raw disclosure that is similar to its industry peers. The second step is to assess each firm's abnormal disclosure, which is based on a firm's raw vocabulary adjusted for the average vocabulary of peers matched on the basis of industry, size, and age. In this second step, we measure the cosine similarity between each firm's abnormal disclosure and the average abnormal disclosure of firms involved in AAERs in the past. The cosine similarity method is a standard approach used in computational linguistics (See Sebastiani (2002) for example), and is frequently used due to its simple interpretation based on its range in the interval [-1,1] and its standardization which provides a natural control for document length. We also consider content analysis on the 10-K MD&A disclosures to identify the key words that firms involved in AAERs use relative to peer firms not involved in AAERs.

The results of our content analysis suggest that firms involved in alleged fraud use more vocabulary associated with complexity including acquisitions, international vocabulary, and litigation. They also disclose more vocabulary relating to uncertainty, and potentially speculative words including "believe", "feasibility", "fluctuating", and "instability". These firms also appear to under-disclose some items relative to

their peers including discussions of financial market liquidity and discussions that would attribute or explain changes in accounting. We also find that disclosures are remarkably different before and after the AAER years. After the AAER years, in particular, disclosures contain abnormally high levels of vocabulary discussing the AAER investigation itself. This reflects that fact that the AAER becomes public information when it is announced by the SEC, and managers disclose the investigation itself in detail. These results confirm that firms involved in AAERs likely engage in localized anti-herding, and typically focus on unique issues that are less prevalent in the population of non-fraudulent firms.

In all, the goal of our study is to examine whether firms are engaged in strategic disclosure at the time they are engaged in alleged fraudulent activity. The goal is not to address the detection of fraud, although understanding the roots of strategic disclosure can eventually inform that issue. Our study makes three primary contributions. First, it is among the first to use text-based analysis of MD&A to examine the link between alleged fraud and strategic disclosure strategies. Second, our study is the first to present new fraud-based industry herding and localized anti-herding hypotheses, and we find strong evidence consistent with both. Finally, we also present content analysis illustrating the key vocabularies used by fraudulent firms, and how they change before, during and after AAER years.

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

1 Literature and Hypotheses

1.1 Existing Literature

Many studies examine the links between accounting, stock returns and AAERs. Feroz, Park, and Pastena (1991) is an early paper that examines the types of issues that motivate SEC enforcement actions and they examine the consequences of enforcement actions. Dechow, Ge and Schrand (2010) provide a detailed review of the AAER and related fraud and manipulations literature more broadly. Although we do not have the scope to summarize all articles in this area, we note a number of studies that are more directly relevant.

Earlier work provides many relevant findings regarding the link between standard accounting variables and other numerical variables and potentially fraudulent activity. Beneish (1997) considers a group of AAER firms and examines, through a Jones model, whether firms that manipulate can be separated from those that merely have more aggressive accruals. Beneish (1999) considers a host of accounting ratios and constructs an index, for which higher values indicate that earnings management is more likely. Dechow, Sloan and Sweeney (1996) find that a strong motive for earnings management is the desire to attract external financing at low cost. Beneish (1999) finds that managers are more likely sell their own shares during times when earnings are overstated. These studies do not consider verbal disclosure or the herding hypotheses we consider.

More recent studies generally extend these earlier works and provide significantly more depth and recency regarding the conclusions. Dechow, Ge, Larson and Sloan (2011) for example find that mis-stating firms appear to be hiding diminishing performance, have higher relative prices, and have abnormal reductions in the num-

ber of employees. Wang (2013) employs new methodology to address the partial-observability problem associated with fraud data, and finds, for example, that R&D increases the likely commission of fraud yet also reduces 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 also relates to industry business conditions and is more intense during industry booms.

Dyck, Morse and Zingales (2010) take a different approach and examine the question of who is most likely to "blow the whistle" on corporate fraud. The authors find that traditional monitors such as investors, the SEC, and auditors play only a small role, whereas non-traditional players including employees and the media play a larger role. Kedia and Philippon (2009) examine the broader economic incentives faced by fraudulent firms, and find that fraud relates to corporate hiring, executive option exercise, and ultimately firm productivity. Kedia and Rajgopal (2011) show that firm locations relative to SEC offices and areas with greater past enforcement activity are less likely to be involved in restatements.

The body of existing work regarding the incentives to commit fraud, and the existence of evidence that can be used by regulators to detect fraud, is extensive. Our study is unique along two primary dimensions. First, our primary research objective is not to improve fraud detection, but rather is to examine whether fraudulent firms engage in disclosure herding and anti-herding. This matter remains unaddressed in the existing work. Second, none of these studies consider information in the verbal disclosures made by managers, whereas our focus is on the verbal disclosure in the MD&A section of the 10-K. Hence, little is known from the existing literature regarding whether managers alter their qualitative disclosures in years their firms are allegedly committing fraud, and why this may or may not be the case.

Our study is not the first to consider textual analysis of MD&A statements and 10-Ks to examine issues in accounting and finance. Cole and Jones (2005), Feldman, Govindaraj, Livnat, and Segal (2010), Li (2008), Li (2010), Brown and Tucker (2011), and Ball, Hoberg and Maksimovic (2013) examine the tone and information content in MD&A and the extent to which it is readable, informative about relevant economic quantities, and updated over time. Hanley and Hoberg (2010) show that informative content exists in IPO prospectus MD&As, and this information relates to ex-post IPO pricing. Bryan (1997) links the tone of disclosures to subsequent three-year outcomes. Kothari, Li and Short (2009) investigate the link between tone in MD&A and a firm's cost of capital. Loughran and McDonald (2011) examine the link between tone and many outcomes including class action lawsuits. Hoberg and Maksimovic (2012) use the content of Capitalization and Liquidity Subsection of firms' MD&A to infer the degree to which firms are constrained.[3] Although many of these studies also examine text in MD&A disclosures, none examines the link between alleged fraud and the possibility of industry herding and anti-herding.

1.2 Hypotheses

In this section we propose the two central hypotheses regarding disclosure of firms potentially committing fraud that we test in this paper. The objective of our study is to test whether firms allegedly committing fraud engage in herding or anti-herding behavior regarding the content in their qualitative disclosures.

The premise behind our first hypothesis is that managers engaged in potentially

[3]Our analysis is also related to a growing literature using text-based analysis to test Finance theory. Earlier work using SEC Edgar disclosures includes Hanley and Hoberg (2010), Hanley and Hoberg (2012), Hoberg and Phillips (2010), and Hoberg and Phillips and Prabhala (2010a). Antweiler and Frank (2004), Tetlock (2007) and Tetlock, Saar-Tsechanksy, and Macskassy (2008) further develop this line of research, and for example, consider the tone in message boards and the media and its link to asset prices.

fraudulent activity, all else equal, prefer to engage in qualitative disclosures that can help them to avoid detection. We first make the basic assumption that managers believe that disclosures are less likely to attract attention or raise red flags if they look "similar to their to peers". Under this assumption, we hypothesize that managers will in fact strategically choose their disclosures to fit this profile as follows:

H1 [Industry Herding Hypothesis]: Managers of firms potentially committing fraud are more likely to provide voluntary disclosure to regulators that is highly similar to industry peers in order to escape detection.

Although many studies define herding in general terms as behavior that is correlated across individuals, Devenow and Welch (1996) define rational herding more specifically as cases where such behavior is rational and where such behavior also leads to erroneous information and choices that are suboptimal relative to what is best in aggregate. Our use of the herding concept is close to the Devenow and Welch (1996) interpretation. In particular, herding is individually rational as it can reduce the likelihood of detection, and furthermore, H1 further predicts that agents who consume the disclosure including investors and fund managers will indeed use erroneous information, and may over-allocate capital to firms potentially committing fraud. Therefore, the equilibrium allocation of financial capital in the economy is likely to be suboptimal. For example, in a no-herding equilibrium, fraud detection should be improved and hence firms would commit less fraud, and the result would be improved aggregate welfare through improved resource allocation.

The main idea behind our second hypothesis is that firms potentially committing fraud also engage in anti-herding along some localized disclosure dimensions, and thus appear distinct from their industry peers on these specific disclosure items. There are two theoretical reasons why they may find it optimal to do so. The first,

once again, relates to reducing the likelihood of detection. For example, as many AAERs relate to revenue and expense manipulations, it is possible that fraudulent firms anti-herd with non-fraudulent firms in their discussions of these items, as a degree of complexity and uncertainty is needed in the disclosure of these items to potentially mask the wrongdoing.

Second, firms committing fraudulent acts may have common objectives. For example, following Dechow, Sloan and Sweeney (1996) they may have incentives to obtain financing at lower cost, or following Dechow, Ge, Larson and Sloan (2011) they may wish to conceal diminishing performance. Regardless of whether firms committing fraud need to distinguish themselves to escape detection or to achieve common fraud-driven objectives, the prediction is that they will appear to anti-herd from their peers along a detectable set of localized disclosures as follows:

H2 [Localized Anti-Herding Hypothesis]: Managers of firms potentially committing fraud are likely to anti-herd with industry peers on a set of common disclosure items that are consistent with concealing fraud and achieving common fraud-driven objectives.

Despite the fact that they initially appear at odds, we note that it is possible that both H1 and H2 can hold. H1 is a broad hypothesis, and we test H1 based on the overall similarity between each firm's disclosure and its industry peers. Our primary test of H1 examines whether this industry similarity is higher for firms allegedly committing fraud.

In contrast, H2 is localized, and the focus is on whether any subset of the overall disclosure is unique to firms allegedly committing fraud yet also distinct from industry peers. Our primary test of H2 thus relies on abnormal disclosure, where we first industry-size-age adjust each firm's vocabulary. We then assess whether

firms allegedly committing fraud use unique abnormal disclosures that have common components, which would lead to the conclusion that they engage in localized anti-herding from their industry peers.

In conclusion, our tests of H1 will be based on raw disclosures, whereas our tests of H2 will be based on industry, size and age adjusted disclosures. Our later results will illustrate that we find especially strong support for H2, and that we also find moderately strong support for H1.

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 (financials have unique MD&A 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 (which ensures our firm-years are non-trivial and publicly traded), 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 Management's Dis-

cussion and Analysis section with a valid link from the 10-K's cik to the Compustat database.[4] We use software provided by metaHeuristica to web crawl and parse 10-Ks and extract the Management's Discussion and Analysis section. MetaHeuristica software uses natural language processing to parse and organize textual data, and its pipeline employs "Chained Context Discovery".[5] 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 10-K MD&A is when MD&A is "incorporated by reference," and is not in the body of the 10-K itself.[6] These 10-K requirements leave us with a final sample of 49,039 firm-year observations having adequate numerical and textual data for our examination.

2.1 Accounting and Auditing Enforcement Releases

We obtain data on Accounting and Auditing Enforcement Releases (AAERs) from the Securities and Exchange Commission website.[7] 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 data on the filing date of each AAER, and the beginning and ending dates each AAER alleges that fraudulent activity has occurred. We define our AAER dummy to be one for firm fiscal years ending in calendar years that overlap in any way with these begin and end dates. This is our primary variable of interest, and our focus is on how firm disclosure varies in the years that fraudulent behavior is alleged.

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

[5]See Cimiano (2010) for details.

[6]The typical scenario under which a MD&A section is incorporated by reference is when the annual report is submitted along with or referenced by the 10-K (and the annual report contains MD&A), and thus MD&A is not available in the 10-K itself. Thus a 10-K parsing engine would not detect its presence.

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

For each AAER, we also identify a year that is definitively prior to the alleged fraudulent activity, and a year that is definitively subsequent to the public release of the AAER by the SEC. We refer to these as the pre-AAER year and the post-AAER year. Our assessing disclosure in three critical periods (prior to alleged fraud, during the years of alleged fraud, and after the years of alleged fraud) serves two purposes. First, this serves as a placebo test and under our hypothesis that fraudulent firms are proactive in choosing their disclosure policies, 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. Second, this allows us to fully understand the disclosure life cycle of fraudulent firms, and especially to examine how firms refine their disclosures in the years after the alleged fraud becomes public and the firm has faced direct scrutiny. Indeed we find materially different disclosure patterns in all three periods.

Due to the potentially approximate nature of alleged fraud periods, we take a conservative approach when identifying the pre-AAER year and the post-AAER year. We thus 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 fraudulent year. 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 or revise its disclosure subsequent to the alleged fraud becoming public information.

2.2 Disclosure Herding

In this section, we focus on the extent to which a firm disclosure-herds with industry peers. A firm engaged in fraudulent behavior may pursue herding to appear typical and escape detection. Our approach of identifying common disclosures across groups of firms is closely related to the methods employed by Hanley and Hoberg (2010), who document that content that is unique relative to industry peers is informative regarding IPO pricing and reduced ex-post uncertainty.

To compute our disclosure industry herding measure, we first group all firms into bins based on industry size and age. We base industry on two-digit SIC codes. For each industry grouping, we then create a small firm and a large firm bin based on the median size among firms in each industry bin. We then divide the bins once again based on age, and each industry-size bin is divided into a young firm and an old firm bin based on median firm listing vintage. We thus have four bins for each SIC-2 industry, and each of the four bins in each industry 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 can report that our findings are entirely robust to only using industry bins rather than these industry-size-age bins. We use these more refined bins due to the fact that we expect material systematic differences in disclosure across firms of different size and age, in addition to differences across industries themselves. We refer to a firm's peers in its same industry-size-age bin as its "ISA peers".

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 these vectors is based on the universe of words the researcher wishes to use to identify textual similarities and differences. 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 for example 99.1% of randomly drawn words using this 1997 screen would be included using an analogous screen based on 2008. Weighting words equally rather than by frequency, 74.6% of words from our 1997 screen would also be included using the analogous 2008 screen. Each firm-year observation has an MD&A disclosure that is thus represented by its word distribution vector $W_{i,t}$. This distribution vector sums to one, and each element indicates the relative frequency of the given word in the given firm year's MD&A. Our use of the first year of our sample for determining 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 herding with ISA peers, we next compute the average word usage vector for a given firm's ISA peers excluding itself. We define this vector as $ISA_{i,t}$, and this vector also has elements that sum to one. We note that it is important that the average of the ISA peers excludes the firm itself, as skipping this step would create a mechanistic degree of similarity that would increase as the number of ISA peers declines. Our measure of industry disclosure herding (H_{it}) is then the cosine similarity between $W_{i,t}$ and $ISA_{i,t}$:

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

[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.

We use the cosine similarity due to its use as a standard technique in computational linguistics for measuring the similarity of two documents (See Sebastiani (2002) for example). It is also easy to interpret, as two documents with no overlap in word usage have a similarity of zero, whereas two documents which use words with exactly the same frequency have a cosine similarity of 1. Finally, by virtue of its normalization to unit length as in the above equation, this method also has the property that it correlates only modestly with document length.

2.3 Disclosure Anti-Herding and Fraud Scores

In this section, we construct measures for the extent to which firms engaged in fraudulent behavior anti-herd with industry peers in common and predictable ways. This form of anti-herding, for example, might relate to pervasive under-discussion of details regarding revenue calculations, or overly-extensive discussion of expected growth in the future. Such disclosures would be lend support to a strategy in which fraud is committed in order to improve the firm's appeal and valuation to investors. A necessary assumption for identification of an anti-herding profile is that firms committing fraud have some commonality in their objectives and thus their incentives to provide disclosures that are unique.

As incentives to commit fraud should span many industries and firms of varying size and age, we first construct industry adjusted disclosure vectors for each firm as follows:

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

The vector $AW_{i,t}$ identifies a firm's abnormal disclosure, relative to what we might expect given its industry, size and age. This vector sums to zero, which must be the

case because 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, which we use for identifying the vocabulary associated with alleged fraudulent firms. 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 score variable. We note, however, that our results are slightly stronger if we do use all years in our sample for the computation of the $AAER_{vocab}$ vector. We view our construction, which permits out of sample assessment, to be a key conservative step in our framework that ensures that our results are not driven by look ahead bias.

We then define the fraud score for any firm in any year F_{it} as the cosine similarity between its industry-adjusted vocabulary $AW_{i,t}$ and the abnormal vocabulary of known-AAER firms in 1997 to 2001 $AAER_{vocab}$.

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

We use our industry herding measure $H_{i,t}$ and our fraud score measure $F_{i,t}$ to test our key hypotheses regarding industry herding and fraud-driven anti herding in subsequent sections. Importantly, we assess performance both in sample (using our entire sample 1997 to 2008) as well as out of sample (using only our out of sample

years 2002 to 2008).

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. In our sample, 1.5% of firm year observations are AAER-years, and are years during which an SEC enforcement action alleges that the firm was involved in fraudulent activity. As it is based on cosine similarities between positive and negative word vectors, the Fraud Disclosure Score has a distribution in the bounded interval [0,1] and a mean that is close to zero. Intuitively, as AAER years are rare, the average firm does not have an industry-size-age adjusted vocabulary that correlates highly with that of fraudulent firms.

[Insert Table 1 Here]

The industry herding score is based on cosine similarities of non-negative vectors, and hence is bounded in the interval [0,1]. Its mean of 0.667 indicates that the average firm shares a rather substantial amount of disclosure with its industry-size-age peers. However, as this mean also is far from one, the average firm also has a substantial amount of unique content.

Table 2 displays Pearson correlation coefficients between our key herding and anti-herding variables, and other key variables including the AAER dummy. The positive correlation between the AAER dummy and both the fraud score and the industry herding score (both significant at the 1% level) foreshadow our later multivariate results. These findings suggest that firms involved in potentially fraudulent activity have disclosures that are consistent with industry herding in general, and that are also consistent with anti-herding on specific disclosure dimensions that are common

among AAER firms. The correlation with the fraud score is stronger at 8.2% than is the correlation with the industry herding variable at 2.6%. Also, remarkably, the fraud score is more correlated with the AAER dummy than all of the displayed variables, even firm size, which is 7.0% correlated with the AAER dummy.

[Insert Table 2 Here]

The fraud score also is 9.0% correlated with the industry herding variable (significant at the 1% level). Given that both variables are functions of firm disclosures, this is somewhat modest in economic magnitude. The modest nature of this figure likely relates to the construction of the variables. One is a function of the disclosure level and the other is a function of abnormal disclosure relative to industry, size and age peers. We also note that the fraud score correlates little with firm size, which likely relates to its construction based on size-adjusted peers (in addition to industry and age adjustments). These aspects of the construction of our variables help to ensure a clear interpretation both in a univariate sense and in a multivariate sense. Finally, the table also suggests that multicollinearity is unlikely to be a concern given the overall modest nature of correlations.

[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.

3.1 Initial Evidence of Herding and Anti-Herding

In this section, we explore the distributional features of our herding and anti-herding measures, and their links to observed AAER Enforcement actions. In Table 4, we sort firms into deciles based on their fraud score and industry herding score disclosure measures. We then report the fraction of firms in each decile that are involved in AAER actions. For this initial test, we sort firms individually into these deciles and report decile statistics separately for each variable.

[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 score or the industry herding decile in which a firm resides. The results are economically large and decile sorting is close to monotonic. Regarding the fraud score, the incidence rate of AAERs in decile 10 is 3.7% compared to just 0.5% for decile 1. Regarding industry herding, there also is a positive link between industry herding and AAER incidence, but it is slightly weaker than that noted for the fraud score. For example, the AAER incidence rate in the highest AAER decile is 2.7% compared to just 1.0% for the lowest AAER decile.

Panel B of table 4 displays analogous results for the out of sample period from 2002 to 2008. Regarding both the fraud score and the industry herding score, we continue to observe strong positive associations with AAER incidence rates. For the fraud score, the inter-decile range is 0.7% to 2.2%. For the industry herding score, this range is 0.9% to 2.4%. Our later tests will show that these basic results are robust to formal multivariate regressions with controls including industry fixed effects.

In table 5, we sort firm years simultaneously by their fraud score and their industry herding score into quartiles, so we can assess the independent links between these variables and AAER incidence rates. We use quartiles to avoid overly small bins, as we have 16 groups based on fraud score and industry herding quartiles. The results in Panel A are based on the full sample. The table shows that both industry herding and fraud scores independently contribute to sorting firms by AAER score. The AAER incidence rate for the highest quartile in both variables is 3.7%, compared to just 0.5% for the lowest quartile for both variables. The fact that the incidence rate in the high quartiles is multiples larger than in the lowest quartile indicates that these findings are economically large, and foreshadows the strong statistical results we report in formal regressions in later sections.

The table also displays the overall sorts separately for both variables across the quartiles in the last row and the last column in each panel. These results are based on conditional sorts, as the averages in these groups for each variable are based on groupings that hold the other variable fixed. The results show that the fraud score more strongly sorts AAER incidence rates compared to the industry herding variable. However, both variables show some independent ability to generate unique variation in AAER incidence rates. These results echo our earlier finding that the fraud score and the industry herding variable are just 9% pairwise correlated and thus contain much distinct information.

[Insert Table 5 Here]

Panel B of table 5 shows that these results are highly robust in the out of sample period. These results indicate that the vocabulary used by AAER firms, that is distinct from industry-size-age peers, has remained stable over time and supports the conclusion that AAER firms engage in anti-herding behavior to accomplish specific

goals associated with achieving fraud-driven goals or to conceal fraudulent policies.

3.2 Fraud Score Distributions

In this section, we examine the distribution of the fraud score. Figure 1 shows the empirical density function of this variable over its full potential range [-1,1]. The figure shows that the distribution is centered near zero, and is nearly bell shaped. However, it also is 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 probability mass. This is materially larger than the observed 1.5% AAER rate indicated in Table 1.

[Insert Figure 1 Here]

We next consider whether upper and lower bounds on the rate of undetected fraud in our sample can be estimated. In order to do so, we have to make two central assumptions. Note that these assumptions only pertain to the estimation of undetected fraud in this section, and are not relevant to the tests in other parts of the paper, where the objective is not to estimate undetected fraud. First, we will assume that non-fraudulent firms have symmetrically distributed fraud scores, and hence the total skewness in the figure is generated by firms engaged in fraud. Second, we assume that firms engaged in fraud that are not detected have similar disclosure patterns as compared to those that are detected. These assumptions allow us to infer the degree of undetected fraud based on how many firms would have to be removed from the sample in order to eliminate the observed skewness.

In order to proceed in computing upper and lower bounds, we first assess the extent to which the removal of known-AAER firm-years reduces the level of observed skewness. Figure 2 thus plots the density function of the fraud score separately for firms not involved in AAERs (upper figure) and for firms that are involved in AAERs (lower figure).

[Insert Figure 2 Here]

Figure 2 shows that the density function retains a substantial degree of right skewness even when known AAER firm-years are excluded. In particular, the degree of skewed mass decreases from 2.55% to 2.10%. The remaining 2.10% is substantial. We compute the upper bound regarding the rate of undetected fraud as the fraction of the sample that would have to be removed in order to remove all observed skewness, again under the assumption that undetected fraudulent firms have similar distributions of the fraud score as do detected fraudulent firms. The result of this calculation is 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 on this figure by assuming that the 2.1% of remaining skewness in Figure 2 is due to 2.1% of undetected firms being engaged in fraud. The result of this calculations that 42.7% ($\frac{1.5}{2.10+1.5}$) of fraudulent firms have been detected and hence fraud is 2.4x as pervasive as observed.

Overall, although the assumptions underlying the estimate of undetected fraud may not hold, these figures illustrate that a substantial fraction of committed fraud may continue to go undetected. The observed rate of known AAER firm years in our sample is 1.5%. The estimated rate of committed AAERs may lie in the range (3.6%, 8.5%).

The lower plot in Figure 2 helps to illustrate why the estimate of undetected fraud

may be larger than what one might expect upon reviewing Figure 1 in isolation. The lower plot displays the density function of the fraud score for firms that are known to be involved in AAERs. The figure shows a far higher degree of skewness than any of the other figures, indicating that the fraud score is effective in separating firms associated with AAERs. The degree of skewed mass is 41.0%, which is far larger than the 2.55% in Figure 1. Although this figure is large, the fact that it is not 100% helps to explain why the upper bound estimate of 5.6x is as high as it is. In particular, the fraud score, while highly effective, is not 100% effective in separating AAER firms. Therefore, removing AAER firms from the distribution does not reduce the degree of skewed mass on a one to one basis.

[Insert Figure 3 Here]

Figure 3 displays fraud scores and industry herding scores over time: before, during and after a firm is involved in an AAER. We also explore the extent to which the score 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 score calculation for all three of these years. If the firm's alleged fraud lasted only one or two years, it will only be included in the first and second fraud year calculations, respectively. To ensure robustness, we also consider this calculation only for firms that experienced a fraud period of at least three years.

The results are displayed in Figure 3. The figure shows a trapezoidal pattern for

the fraud score. During the three years preceding the alleged fraud, the average fraud score 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, the fraud score 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.

The lower figure in Panel 3 suggests that the time series results for the industry herding variable are not as decisive. We do not observe a strong difference between the fraud period and the pre or post fraud periods. This echoes results in the next section, where we show that the overall results for the fraud score (anti-herding) are stronger than those for the industry herding variable. However, because we do find significant cross sectional results for the industry herding variable in these later tests, the figure thus suggests that our overall results for industry herding are based on cross sectional differences and not on time series differences, or they are due to lower frequency changes in the level of industry herding that would not be as easily observed in the figure. This is in contrast to the more decisive results for the fraud score variable, where we find sharp differences across AAER and non-AAER firms both in time series and in cross section.

4 Formal Analysis of Disclosure Herding and Anti-Herding

In this section, we use formal regression analysis to test our herding and anti-herding hypotheses. We consider these tests and assess disclosures in the year of an AAER,

the year prior, and the year after. Assessing disclosure in these three periods serves two purposes. First, it serves as a placebo test and under our hypothesis that fraudulent firms are proactive in choosing their disclosure policies, 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. Second, this approach allows us to fully understand the disclosure life cycle of fraudulent firms, and especially to examine how firms refine their disclosures in the years after the alleged fraud becomes public and the firm has faced direct scrutiny. We also consider these three periods in the next section where we document the actual content vocabulary used by AAER firms relative to non-AAER firms.

Table 6 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 the fraud score or the industry herding score. In Panel A to C, we report results for the entire sample, and in Panels D to F, we report results for the out of sample period. For each sample period, we report overall sample regression results, and results for large and small firms, where firm size is identified using median assets in each year.

[Insert Table 6 Here]

Panel A of Table 6 shows that firms engaged in alleged fraud have significantly higher fraud scores and industry herding scores. The results are particularly strong for the fraud score, as the t-statistic of 9.10 is highly significant at a level well beyond the 1% level. The results for industry herding are significant at the 10% level with a t-statistic of 1.8. We also note that all regressions include industry fixed effects based on two digit SIC industries and standard errors are also adjusted for clustering by firm.

Panels B and C of Table 6 show that the fraud score results are highly robust at the 1% level for large firms and for small firms, respectively. We also find that industry herding is only significant for smaller firms in Panel C (at the 1% level) and not for larger firms in Panel B. Hence, anti herding as identified by the fraud score is the dominant result in our overall sample, although we also find support for industry herding among smaller firms.

Panels D to F of Table 6 show that results for the fraud score and the industry herding score remain highly robust during the out of sample period from 2002 to 2008. The fraud score is significant in all three panels, but the industry herding result is significant only in the overall sample and in the large firm sample in Panels D and E, respectively. It retains its positive coefficient but falls below 10% level significance for small firms in Panel F. Overall, we conclude that our support for the anti herding hypothesis among AAER firms is strongly supported in all samples, and that the industry herding hypothesis is significant but overall is less reliable. Because the sample size for the small and large firm tests is only half as large, the differences we observe for small and large firms in the overall sample and the out of sample period are likely due to reduced power in these tests, and it is thus not clear whether industry herding varies strongly across firms of different size.

Table 7 uses the same framework as Table 6, except that the we consider the future AAER dummy as an explanatory variable instead of the actual AAER dummy. As a result, we are thus testing if the fraud score and the industry herding score are elevated in the year prior to which the SEC alleges fraudulent activity. If the anti herding and industry herding hypotheses strictly relate to the act of committing fraud, and not passive long term firm characteristics, we expect that the results for the future AAER dummy should be insignificant or at least substantially weaker

than those in Table 6.

[Insert Table 7 Here]

Table 7 shows only weak evidence of statistical links between our disclosure variables and the future AAER dummy (a dummy that is one if the firm will be involved in an AAER in the next fiscal year). In Panel A, which is based on the full sample, we find that the future AAER dummy is statistically significant, but only at the 10% level, when predicting the fraud score. Regarding large and small firms in Panels B and C, we find a 5% significant result for large firms and no significant link for small firms. In all cases, these results are far weaker than in Table 6, where results were reliably significant at the 1% level (with t-statistics ranging from more than five to more than nine. We do not find significant results for the industry herding score. The out of sample results in Panels D to F are more decisive, and we do not find any evidence that the future AAER dummy is associated with our fraud score and industry herding disclosure variables. In all, we conclude that our evidence in Table 6 is strongly linked to the years that firms are allegedly engaged in fraud. That is, our disclosure results are likely not related to passive long-term firm characteristics and are relatively unique to firms allegedly committing fraud in the year the AAER reports the firm is committing fraud.

[Insert Table 8 Here]

Table 8 is related to Table 7, except we replace the future AAER dummy with the past AAER dummy. Hence, the dummy variable identifies firms that have committed fraud in the past, but are no longer committing fraud. Remarkably, the results of Table 8 are very similar to those of Table 7. In particular, for the full sample, we find some modest links between the past AAER dummy and the fraud score disclosure variable. However, we find no evidence of a link in the out of sample period, and

we find very little evidence of a link for the industry herding variable. Overall, these results provide further evidence that our disclosure results are likely not related to passive long-term firm characteristics and are relatively unique to firms allegedly committing fraud in the year the AAER reports the firm is committing fraud.

5 Content Analysis

In this section we report the key vocabulary that distinguishes firms involved in AAERs from non-AAER firms. We also consider whether firms disclose unique language in the year prior to AAER years and in the year after AAER years. The intent of these tests is to examine the links we found in earlier sections, and to focus on the years in which firms allegedly commit fraud and how these years differ.

Panel A of Table 9 reports the top 50 words that are used more aggressively by firms involved in AAERs as compared to firms that are not involved in AAERs. 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 fifty words with the lowest p-values regarding the confidence in rejecting the hypothesis that the word is used equally among AAER and non-AAER firms are then chosen for presentation. For robustness, Panel B of Table 9 displays the results of analogous difference tests based on the LDA topics in Ball, Hoberg and Maksimovic (2013) (See Section 5.5 and Table 9 of their study for details).[9]

[Insert Table 9 Here]

The table shows that, in the years they are involved in AAERs, firms disclose

[9]The LDA topics in Ball, Hoberg and Maksimovic (2013) (BHM) are derived using computational linguistic algorithms that identify common textual factors explaining systematic variation in MD&A disclosures in general. This is akin to factor analysis, but LDA focuses on verbal data instead of numerical data. We label (and report in our tables) each topic using the top five commongrams each loads highly on, which is obtained from the metaHueristica software program (as do BHM).

more information about many topics. Some words load on litigation, and suggest that firms involved in AAERs are involved in other types of litigation suggesting they attract multiple forms of legal attention. They are also more likely to be involved in restatements, which indicates that they typically have 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 likely that complex transactions such as acquisitions and more difficult to trace international transactions may facilitate fraudulent accounting that is more difficult to review.

Firms involved in AAERs also disclose more vocabulary indicative of uncertainty and speculation: "believe", "feasibility", "fluctuating", and "instability". Some of the disclosure is consistent with proactive herding, as words including "similar" and "prevailing" are also more common in the abnormal disclosures of AAER firms. Finally, the word "dilutive" also appears indicating that managers are concerned about maintaining their ownership rights and the value of equity.

The topic analysis in Panel B confirms the conclusions from Panel A, with a couple additions. Rows three and four provide further support for the conclusion that AAER firms focus on transactions, and row seven confirms the focus on issues in foreign countries, confirming the added complexity in their disclosures. Also interesting is that the results in Panel B suggest that firms involved in AAERs also under-disclose some components of information relative to their industry peers. In particular, the first row suggests that they likely under-disclose attribution text explaining various changes in the firm's accounting, and the second row suggests that they likely under-disclose information relating to challenges in their financial liquidity (presumably to artificially obtain superior terms when they raise capital).

[Insert Table 10 Here]

Table 10 repeats this exercise and identifies abnormal vocabulary used by firms in the year prior to their AAER year. Hence, these vocabularies foreshadow potential issues faced by firms that may explain their incentives to commit fraud, or that may provide clues for less extreme behavior that may occur prior to more severe forms of fraud. The vocabulary used by these firms is substantially different from that displayed in Table 9. Notable categories include words relating to high growth including "growth" and "opportunities", and words relating to complex instruments or transactions including "derivatives", "instruments", "acquisition" and international vocabulary including "currencies", "pacific", and "Canada".

The results for topics in Panel B confirms a somewhat elevated discussion of acquisitions and foreign currencies as in Table 9, but we also note that the topic discussion results are notably weaker in the year prior to the AAER years relative to the actual AAER years. We also observe less discussion of dividends and attribution text relative to peers in the year prior to the AAER. In all, these results suggest that firms involved in more complex transactions and activities are more likely to become involved in AAERs in future years. A potential reason for this result is the fact that complexity may lead managers to believe that their performance is too complex to review, and that fraudulent activity may be harder to detect.

Table 11 repeats this exercise and identifies abnormal vocabulary used by firms in the year after their AAER year. Hence, these vocabularies should reflect the fact that the associated managers of these firms have gone through the experience of SEC scrutiny and the potential loss of reputation from the public announcement of their firm's alleged fraudulent activity. Our hypothesis is that the key indicators of fraud found in the past two tables should no longer be present, and moreover, the firm will

disclose information about the SEC review itself.

[Insert Table 11 Here]

The discussions in Table 11 strongly support these predictions, especially the prediction that managers disclose a material amount of information about the SEC review itself. Several of the top words in the table's list relate directly to the SEC: "SEC", "Lawsuits", "Audit", "File", "Proceedings", "Review", "Investigations", "Auditors", "Violations", and many more. These results confirm that managers react to the AAER proceedings and modify their disclosure, and these results also confirm that our framework is informative regarding the construction and timing of the variables we employ. The topic analysis in Panel B further support this conclusion, especially rows one and three, which suggest that these firms report more in the way of restructuring writedowns and they have elevated discussions of litigation.

[Insert Table 12 Here]

Table 12 displays examples of the firm AAER-years that are most consistent with our key anti-herding and industry herding hypotheses. In particular, we report the Top 25 AAERs in which the associated firm's fraud score is among the highest in our sample. These firms use an industry adjusted vocabulary that is highly similar to the average vocabulary used by AAER firms in general. For each firm, we also report its fraud score and its industry herding score. Because the unit of observation is a firm-year, it is possible that a firm can appear in our list more than once if an AAER exists listing the firm as committing fraud during more than one year. We also report the top 25 AAER firms that have the highest industry herding scores.

6 Conclusions

We hypothesize that firms involved in potentially fraudulent activity face tensions when providing qualitative disclosures to the Securities and Exchange Commission, the agency tasked with enforcing anti-fraud laws. Our focus is on the Management's Discussion and Analysis section of the 10-K, which is where managers have a high level of discretion to describe the key issues facing their firms and to describe their performance in detail. A primary motive is to escape detection, and managers who assume that the SEC is less likely to scrutinize disclosures that resemble industry peers, or that such disclosure is less likely to raise red flags, have incentives to herd with industry peers. On the other hand, the same objectives that lead managers to commit fraud may also provide incentives to anti-herd in their disclosure from industry peers. However, these latter incentives are likely more localized, and anti-herding would be predicted only on disclosure dimensions that might help managers to achieve these objectives.

We find especially strong support for the localized anti-herding hypothesis, and moderately strong support for the industry herding hypothesis. Indeed the same firms can choose disclosure strategically that both resembles industry peers overall, yet deviates from industry peers on a small number of important dimensions. These results provide strong support for the conclusion that strategic disclosure incentives are strong for firms involved in SEC accounting and auditing enforcement actions (AAERs). We can separately measure both types of disclosure because industry herding is directly tested by considering similarities between a firm's raw disclosure and that of its industry, size and age matched peers. Localized anti-herding is then measured using the clustering properties of abnormal disclosures (raw disclosure adjusted for that of industry-size-and age peers).

Our results are particularly striking along two dimensions. First, our identification examines firms involved in AAERs both relative to firms not involved in AAERs, and also relative to the same firms involved in AAERs but for the years before and after the AAER-influenced years as indicated in the AAER filing announcements. These results suggest that disclosures are revised frequently as a firm evolves from a pre-AAER firm, to a firm involved in AAER actions, to a firm that has been revealed as allegedly committing fraud. Content analysis reveals much granularity regarding the discussions firms disclose over this cycle. In particular, firms involved in fraud focus more than average on complexity including acquisitions and international words to potentially conceal fraudulent activity. These firms also discuss issues relating to uncertainty, litigation, and speculative statements more than their peers do. After firms are revealed by the SEC to allegedly have committed fraud, their disclosure is revised materially and focuses more than average on describing the SEC investigation itself in great detail.

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.

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, 13574.

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, 2213-2253.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, Managements Tone Change, Post Earnings Announcement Drift and Accruals, *Review of Accounting Studies* 15, 91553.

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*.

Hoberg, Gerard, Gordon Phillips, and Nagpurnanand Prabhala, 2013, Product Market Threats, Payouts, and Financial Flexibility, *Forthcoming: Journal of Finance*.

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.

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 SECs 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, 163970.

Li, Feng, 2010, Information Content of the Forward-Looking Statements in Corporate FilingsA Naive Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 10491102.

Li, Feng, 2008, Annual Report Readability, Current Earnings, and Earnings Persistence, *Journal of Accounting and Economics* 45, 221-247.

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, Maytal Saar-Tsechanksy, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437–1467.

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 herding 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 disclosure 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 disclosure 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.

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 Herding Score	0.667	0.080	0.410	0.671	0.839
Fraud Disclosure 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 Score	Industry Herding Score	Log Sales	Operating Income/ Sales	R&D Sales
				Correlation Coefficients			
(1)	Fraud Score	0.082					
(2)	Industry Herding 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 Scores and Industry Herding 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 scores (first two columns) and based on their industry herding 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 Score	Fraction AAER Firm Years	Industry Herding 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: AAERs versus Fraud Scores and Industry Herding (2-D Sorts)

The table displays the fraction of firm-year observations associated with AAERs for our sample of 49,039 observations based on annual firm observations from 1997 to 2008 based on 2-D quartile sorts. Within each year, firms are independently sorted into quartiles based on their fraud scores (rows) and based on their industry herding scores (columns). The fraction of firms involved in AAERs is then reported for each group. See Table 1 for the description of our key variables.

Fraud Score Quartile	Industry Herding Quartile 1	Industry Herding Quartile 2	Industry Herding Quartile 3	Industry Herding Quartile 4	Overall
		Panel A: Full Sample (1997 to 2008)			
0	0.005	0.005	0.009	0.009	0.007
1	0.008	0.010	0.011	0.012	0.010
2	0.012	0.010	0.013	0.018	0.013
3	0.023	0.023	0.027	0.037	0.028
Overall	0.012	0.012	0.015	0.019	0.015
		Panel B: Out of Sample (2002 to 2008)			
0	0.008	0.003	0.012	0.011	0.008
1	0.010	0.009	0.013	0.016	0.012
2	0.007	0.010	0.010	0.017	0.011
3	0.011	0.014	0.019	0.030	0.019
Overall	0.009	0.009	0.014	0.018	0.013

Table 6: Disclosure Outcome Regressions (AAER-year)

In Panels A to C, the table reports OLS regressions for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. In Panels D to F, the table reports OLS regressions for our out of sample period including 26,061 annual firm observations from 2002 to 2008. These out of sample tests are out of sample because the base vocabulary used to compute the fraud score is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. The dependent variable is based on the firm's disclosure and varies by row as indicated. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the the current year. See Table 1 for the description of our key variables. All regressions are estimated with year and industry 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	Obs./ R^2
			Panel A: Entire Sample				
(1)	Fraud Profile Sim.	0.052	-0.003	0.005	-0.003	0.000	49,039
		(9.10)	(-1.17)	(6.51)	(-1.92)	(1.23)	0.010
(2)	Industry Herding Sim.	0.008	-0.013	0.008	0.005	0.008	49,039
		(1.82)	(-6.35)	(11.06)	(2.68)	(20.03)	0.200
			Panel B: Above Median Firm Size Only				
(3)	Fraud Profile Sim.	0.048	0.036	0.229	-0.007	0.003	24,523
		(7.65)	(6.69)	(7.71)	(-2.21)	(4.56)	0.065
(4)	Industry Herding Sim.	0.005	-0.005	0.084	0.007	0.000	24,523
		(1.20)	(-1.02)	(6.13)	(1.42)	(0.35)	0.202
			Panel C: Below Median Firm Size Only				
(5)	Fraud Profile Sim.	0.048	-0.008	0.003	-0.002	0.003	24,516
		(5.34)	(-3.06)	(4.04)	(-1.23)	(3.36)	0.009
(6)	Industry Herding Sim.	0.016	-0.027	0.009	0.004	0.018	24,516
		(2.44)	(-11.55)	(11.40)	(2.02)	(23.39)	0.242
			Panel D: Entire Sample (Out of Sample Years Only)				
(1)	Fraud Profile Sim.	0.025	0.001	0.003	0.001	0.000	25,926
		(3.92)	(0.27)	(3.62)	(0.47)	(0.40)	0.003
(2)	Industry Herding Sim.	0.014	-0.011	0.007	0.005	0.007	25,926
		(3.06)	(-4.48)	(8.67)	(1.90)	(15.47)	0.228
			Panel E: Above Median Firm Size Only (Out of Sample Years Only)				
(3)	Fraud Profile Sim.	0.016	0.046	0.202	-0.013	0.004	12,965
		(2.14)	(6.27)	(8.78)	(-2.64)	(5.33)	0.054
(4)	Industry Herding Sim.	0.013	-0.010	0.066	0.013	0.000	12,965
		(2.58)	(-1.42)	(5.00)	(1.72)	(0.27)	0.236
			Panel F: Below Median Firm Size Only (Out of Sample Years Only)				
(5)	Fraud Profile Sim.	0.037	-0.002	0.001	0.001	0.001	12,961
		(3.69)	(-0.76)	(1.46)	(0.58)	(0.78)	0.006
(6)	Industry Herding Sim.	0.012	-0.023	0.008	0.004	0.015	12,961
		(1.45)	(-8.13)	(9.03)	(1.60)	(16.75)	0.258

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

In Panels A to C, the table reports OLS regressions for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. In Panels D to F, the table reports OLS regressions for our out of sample period including 26,061 annual firm observations from 2002 to 2008. These out of sample tests are out of sample because the base vocabulary used to compute the fraud score is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. The dependent variable is based on the firm's disclosure and varies by row as indicated. The Future AAER dummy is one if a future AAER action indicates that the firm was involved in fraudulent activity in the the year after the current year of the observation. See Table 1 for the description of our key variables. All regressions are estimated with year and industry fixed effects, and standard errors are clustered by firm. t-statistics are in parentheses.

Row	Dependent Variable	Future AAER Dummy	Operating Income /Sales	R&D /Sales	CAPX /Sales	Log Sales	Obs./ R^2
				Panel A: Entire Sample			
(1)	Fraud Profile Sim.	0.014	-0.003	0.005	-0.003	0.001	49,039
		(1.77)	(-1.27)	(6.56)	(-1.82)	(1.99)	0.004
(2)	Industry Herding Sim.	-0.000	-0.013	0.008	0.005	0.008	49,039
		(-0.05)	(-6.37)	(11.07)	(2.69)	(20.19)	0.200
				Panel B: Above Median Firm Size Only			
(3)	Fraud Profile Sim.	0.017	0.036	0.233	-0.007	0.004	24,523
		(2.10)	(6.72)	(7.73)	(-2.25)	(5.09)	0.057
(4)	Industry Herding Sim.	-0.008	-0.005	0.085	0.007	0.000	24,523
		(-1.14)	(-1.01)	(6.14)	(1.41)	(0.44)	0.202
				Panel C: Below Median Firm Size Only			
(5)	Fraud Profile Sim.	-0.002	-0.008	0.003	-0.002	0.003	24,516
		(-0.11)	(-3.14)	(4.01)	(-1.14)	(3.59)	0.006
(6)	Industry Herding Sim.	0.017	-0.027	0.009	0.004	0.018	24,516
		(1.48)	(-11.58)	(11.38)	(2.05)	(23.49)	0.242
				Panel D: Entire Sample (Out of Sample Years Only)			
(1)	Fraud Profile Sim.	-0.000	0.001	0.003	0.001	0.000	25,926
		(-0.02)	(0.21)	(3.65)	(0.56)	(0.69)	0.002
(2)	Industry Herding Sim.	-0.015	-0.011	0.007	0.005	0.007	25,926
		(-0.87)	(-4.51)	(8.68)	(1.94)	(15.69)	0.228
				Panel E: Above Median Firm Size Only (Out of Sample Years Only)			
(3)	Fraud Profile Sim.	0.049	0.046	0.203	-0.013	0.004	12,965
		(1.52)	(6.26)	(8.82)	(-2.69)	(5.47)	0.053
(4)	Industry Herding Sim.	-0.012	-0.010	0.067	0.013	0.000	12,965
		(-0.36)	(-1.43)	(5.03)	(1.73)	(0.39)	0.236
				Panel F: Below Median Firm Size Only (Out of Sample Years Only)			
(5)	Fraud Profile Sim.	-0.018	-0.002	0.001	0.001	0.001	12,961
		(-0.88)	(-0.82)	(1.46)	(0.68)	(0.96)	0.004
(6)	Industry Herding Sim.	-0.018	-0.023	0.008	0.004	0.015	12,961
		(-0.92)	(-8.15)	(9.03)	(1.62)	(16.84)	0.258

Table 8: Disclosure Outcome Regressions (Post-AAER Disclosures)

In Panels A to C, the table reports OLS regressions for our sample of 49,039 observations based on annual firm observations from 1997 to 2008. In Panels D to F, the table reports OLS regressions for our out of sample period including 26,061 annual firm observations from 2002 to 2008. These out of sample tests are out of sample because the base vocabulary used to compute the fraud score is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. The dependent variable is based on the firm's disclosure and varies by row as indicated. The Past AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the the year prior to the current year. See Table 1 for the description of our key variables. All regressions are estimated with year and industry fixed effects, and standard errors are clustered by firm. t-statistics are in parentheses.

Row	Dependent Variable	Past AAER Dummy	Operating Income /Sales	R&D /Sales	CAPX /Sales	Log Sales	Obs./ R^2
				Panel A: Entire Sample			
(1)	Fraud Profile Sim.	0.009	-0.003	0.005	-0.003	0.001	49,039
		(1.66)	(-1.27)	(6.57)	(-1.81)	(1.99)	0.004
(2)	Industry Herding Sim.	0.006	-0.013	0.008	0.005	0.008	49,039
		(1.13)	(-6.36)	(11.07)	(2.69)	(20.15)	0.200
				Panel B: Above Median Firm Size Only			
(3)	Fraud Profile Sim.	0.013	0.036	0.233	-0.007	0.004	24,523
		(2.21)	(6.73)	(7.73)	(-2.23)	(5.10)	0.057
(4)	Industry Herding Sim.	0.011	-0.005	0.084	0.007	0.000	24,523
		(1.85)	(-1.01)	(6.14)	(1.42)	(0.39)	0.202
				Panel C: Below Median Firm Size Only			
(5)	Fraud Profile Sim.	-0.008	-0.008	0.003	-0.002	0.003	24,516
		(-0.78)	(-3.15)	(4.01)	(-1.14)	(3.60)	0.006
(6)	Industry Herding Sim.	0.003	-0.027	0.009	0.004	0.018	24,516
		(0.33)	(-11.57)	(11.38)	(2.04)	(23.48)	0.241
				Panel D: Entire Sample (Out of Sample Years Only)			
(1)	Fraud Profile Sim.	0.007	0.001	0.003	0.001	0.000	25,926
		(1.25)	(0.22)	(3.65)	(0.56)	(0.66)	0.002
(2)	Industry Herding Sim.	0.004	-0.011	0.007	0.005	0.007	25,926
		(0.68)	(-4.51)	(8.68)	(1.94)	(15.65)	0.228
				Panel E: Above Median Firm Size Only (Out of Sample Years Only)			
(3)	Fraud Profile Sim.	0.011	0.046	0.203	-0.013	0.004	12,965
		(1.76)	(6.26)	(8.81)	(-2.60)	(5.43)	0.053
(4)	Industry Herding Sim.	0.006	-0.010	0.066	0.013	0.000	12,965
		(1.01)	(-1.43)	(5.02)	(1.73)	(0.36)	0.236
				Panel F: Below Median Firm Size Only (Out of Sample Years Only)			
(5)	Fraud Profile Sim.	-0.008	-0.002	0.001	0.001	0.001	12,961
		(-0.82)	(-0.83)	(1.46)	(0.67)	(0.98)	0.004
(6)	Industry Herding Sim.	0.001	-0.023	0.008	0.004	0.015	12,961
		(0.06)	(-8.15)	(9.03)	(1.62)	(16.83)	0.258

Table 9: Key words Driving Fraud Scores

The table lists the top 50 words (Panel A) and Topic Model Factors (Panel B) found to be important among firms involved in AAER actions as compared to firms not involved in AAER actions. Panel A is restricted to the top 50 most significant words and Panel B restricts attention to 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). This difference is based on industry adjusted word frequencies, and hence any difference in vocabulary is industry adjusted. The table reports the top fifty words (those with the highest t-statistic in a test examining whether the given word is used more by firms involved in AAERs as compared to firms that are not involved in AAERs) in addition to each word's t-statistic regarding how different its frequency is relative to non-AAER firms.

Panel A: Top 50 Words

Word Rank	Word	t-statistic	Word Rank	Word	t-statistic
1	FEASIBILITY	4.765	26	EXPOSED	3.668
2	RESTATEMENT	4.738	27	ATTENTION	3.661
3	LITIGATION	4.490	28	TIMELY	3.621
4	WHERE	4.489	29	POLITICAL	3.616
5	ACQUISITIONS	4.475	30	PRODUCTS	3.597
6	TECHNOLOGICAL	4.223	31	INTERESTS	3.588
7	UPDATE	4.222	32	SIGNIFICANT	3.585
8	WORDS	4.197	33	ASSERTIONS	3.560
9	CONSIDERED	4.172	34	DILUTIVE	3.545
10	POOLING	4.009	35	OFTEN	3.537
11	CALCULATIONS	4.003	36	PREVAILING	3.533
12	PUBLICLY	3.922	37	COMMERCIALLY	3.517
13	PUBLISHED	3.895	38	DUTIES	3.507
14	PREDICT	3.878	39	DEFECTS	3.490
15	CANNOT	3.839	40	INSTABILITY	3.484
16	BELIEVE	3.837	41	FOUNDED	3.480
17	SIMILAR	3.831	42	COMMITTEE	3.479
18	ACQUIRED	3.826	43	FOUNDRIES	3.471
19	ASSIMILATE	3.813	44	COSTEFFECTIVE	3.459
20	EARTHQUAKE	3.809	45	PURPORTED	3.453
21	COUNTRIES	3.804	46	INCORPORATE	3.388
22	UNDERLYING	3.789	47	BRAZIL	3.378
23	AFRICA	3.742	48	SELECTION	3.374
24	FLUCTUATING	3.736	49	STRAIN	3.363
25	UNEARNED	3.705	50	PUBLISH	3.322

Panel B: Topics from Ball, Hoberg and Maksimovic (2013)

Topic Commongrams	t-statistic
partially offset, primarily due, offset decrease, due primarily, decreased decrease	-6.12
sufficient meet, additional financing, sources liquidity, raise additional, additional funds	-5.05
acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	4.24
restructuring charge, restructuring charges, write downs, special charges, fourth quarter	3.71
continued growth, business strategy, growth strategy, business opportunities, core business	3.16
payments made, principal payments, payment dividends, pay dividends, dividends paid	-3.09
foreign currency, foreign exchange, north america, currency exchange, domestic international	2.88
compliance covenants, coverage ratio, restrictive covenants, leverage ratio, maintain minimum	-2.76
common stock, preferred stock, series preferred stock, shares common stock, convertible preferred stock	-2.68
marketing expenses, professional fees, salaries benefits, expenses related, related expenses	-2.39

Table 10: Key words Driving Fraud Scores (Pre-AAER Disclosures)

The table lists the top 50 words (Panel A) and Topic Model Factors (Panel B) found to be important among firms in the year before they are involved in AAER actions as compared to firms not involved in AAER actions. Panel A is restricted to the top 50 most significant words and Panel B restricts attention to 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). Panel B restricts attention to 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). This difference is based on industry adjusted word frequencies, and hence any difference in vocabulary is industry adjusted. The table reports the top fifty words (those with the highest t-statistic in a test examining whether the given word is used more by firms in the year before they are involved in AAERs as compared to firms that are not involved in AAERs) in addition to each word's t-statistic regarding how different its frequency is relative to non-AAER firms.

Panel A: Top 50 Words

Word Rank	Word	t-statistic	Word Rank	Word	t-statistic
1	DERIVATIVE	3.622	26	RANGING	2.061
2	RECORDS	2.736	27	HEDGING	2.049
3	INSTRUMENTS	2.673	28	UNDERLYING	2.047
4	DERIVATIVES	2.639	29	BA	2.043
5	REFORM	2.307	30	CURRENCIES	2.042
6	PRIVATE	2.268	31	NEARTERM	2.039
7	ACQUISITION	2.260	32	SIZES	2.036
8	OPPORTUNITIES	2.249	33	PLANNED	2.034
9	SUSTAIN	2.242	34	DO	2.031
10	HARBOR	2.210	35	BILLION	2.013
11	IMPACT	2.206	36	IDENTIFICATION	1.984
12	COMBINATION	2.196	37	INTERNATIONAL	1.975
13	COMMENCE	2.183	38	UNDERSTANDING	1.970
14	INVESTMENT	2.175	39	OLDER	1.969
15	ACCUMULATED	2.175	40	MOODYS	1.945
16	AMORTIZED	2.167	41	PACIFIC	1.932
17	CANADA	2.161	42	EFFECTIVENESS	1.932
18	INTERNALLY	2.152	43	ESTABLISHING	1.919
19	GROWTH	2.144	44	CORRESPONDING	1.915
20	DUPLICATE	2.138	45	NA	1.910
21	EXPOSED	2.134	46	PRESENTS	1.905
22	SAFE	2.105	47	PRUDENT	1.886
23	ACQUIRED	2.088	48	DOMESTICALLY	1.884
24	ACT	2.074	49	POORS	1.872
25	REPUTATION	2.063	50	INCORRECTLY	1.861

Panel B: Topics from Ball, Hoberg and Maksimovic (2013)

Topic Commongrams	t-statistic
payments made, principal payments, payment dividends, pay dividends, dividends paid	-2.44
acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	2.35
partially offset, primarily due, offset decrease, due primarily, decreased decrease	-2.22
foreign currency, foreign exchange, north america, currency exchange, domestic international	2.10

Table 11: Key words Driving Fraud Scores (Post-AAER Disclosures)

The table lists the top 50 words (Panel A) and Topic Model Factors (Panel B) found to be important among firms in the year after they are involved in AAER actions as compared to firms not involved in AAER actions. Panel A is restricted to the top 50 most significant words and Panel B restricts attention to 5% level significant topics from the set in Table IX of Ball, Hoberg and Maksimovic (2013). This difference is based on industry adjusted word frequencies, and hence any difference in vocabulary is industry adjusted. The table reports the top fifty words (those with the highest t-statistic in a test examining whether the given word is used more by firms in the year after they are involved in AAERs as compared to firms that are not involved in AAERs) in addition to each word's t-statistic regarding how different its frequency is relative to non-AAER firms.

Panel A: Top 50 Words

Word Rank	Word	t-statistic	Word Rank	Word	t-statistic
1	SEC	5.806	26	INVESTIGATIONS	3.405
2	RESTATEMENT	5.114	27	INQUIRY	3.397
3	COMMITTEE	4.594	28	COMMITTEES	3.378
4	FINDINGS	4.428	29	AMONG	3.331
5	LITIGATION	4.237	30	DISTRICT	3.317
6	LAWSUITS	4.189	31	INTERNAL	3.296
7	FORMAL	4.174	32	OUTCOME	3.296
8	AUDIT	4.158	33	AUDITORS	3.289
9	ACTION	4.151	34	RESTATED	3.249
10	CONTROLS	3.959	35	INFORMAL	3.202
11	THINGS	3.787	36	VIOLATIONS	3.201
12	ACTIONS	3.778	37	LEGAL	3.183
13	FILE	3.726	38	CORRECT	3.154
14	AGAINST	3.710	39	SURROUNDING	3.123
15	PROCEEDINGS	3.672	40	Q	3.123
16	OFFICERS	3.655	41	CIVIL	3.118
17	FILED	3.572	42	INFORMED	3.106
18	RELATING	3.557	43	FIDUCIARY	3.105
19	UNABLE	3.504	44	CERTAIN	3.042
20	FORMER	3.469	45	CONDUCTED	3.039
21	REVIEW	3.452	46	RESTATE	3.019
22	CONDUCTING	3.436	47	CONNECTION	3.015
23	FILING	3.430	48	INITIATED	3.015
24	INDEPENDENT	3.423	49	PENDING	2.988
25	REPORTS	3.413	50	APPOINTED	2.920

Panel B: Topics from Ball, Hoberg and Maksimovic (2013)

Topic Commongrams	t-statistic
restructuring charge, restructuring charges, write downs, special charges, fourth quarter	6.79
primarily attributable, increase attributable, increase primarily, attributable increase, increase increase	-4.60
legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	3.75
research development, research development expenses, product development, process research development, development stage	-3.28
common stock, preferred stock, series preferred stock, shares common stock, convertible preferred stock	-2.72

Table 12: Top 25 Representative Anti-Herding and Industry Herding AAERs

The table lists the 25 most representative AAER firms based on the magnitude of their fraud score (first four columns) and the magnitude of their industry herding score (latter four columns). For each list, we include the year the AAER firm achieved the high score, and we also report each firm's fraud score and its industry herding score. See Table 1 for the description of our key variables.

	Top 25 Fraud Score				Top 25 Industry Herding			
Row	Company Name	AAER Year	Fraud Score	Industry Herding Score	Company Name	AAER Year	Fraud Score	Industry Herding Score
1	SMARTDISK CORP	2000	0.336	0.672	EMBARCADERO TECHNOLOGIES INC	2005	-0.005	0.877
2	SMARTDISK CORP	2001	0.330	0.791	GENESCO INC	2001	0.054	0.865
3	TRIDENT MICROSYSTEMS INC	2004	0.326	0.786	QUEST SOFTWARE INC	2002	0.054	0.861
4	ENTERASYS NETWORKS INC	2001	0.320	0.853	ENTERASYS NETWORKS INC	2001	0.320	0.853
5	LERNOUT HAUSPIE SPCH PD	1999	0.290	0.696	UTSTARCOM HOLDINGS CORP	2004	0.050	0.851
6	TRIDENT MICROSYSTEMS INC	1999	0.285	0.812	INSPIRE PHARMACEUTICALS INC	2005	0.032	0.848
7	DAISYTEK INTL CORP	2002	0.280	0.784	I2 TECHNOLOGIES INC	2002	0.192	0.846
8	COMVERSE TECHNOLOGY INC	2001	0.272	0.729	EMBARCADERO TECHNOLOGIES INC	2002	0.192	0.843
9	TRIDENT MICROSYSTEMS INC	2000	0.270	0.781	MICHAELS STORES INC	2005	0.039	0.841
10	MARVELL TECHNOLOGY GROUP LTD	2000	0.270	0.822	MARVELL TECHNOLOGY GROUP LTD	2005	0.067	0.838
11	PEREGRINE SYSTEMS INC	2000	0.261	0.742	JUNIPER NETWORKS INC	2002	0.107	0.838
12	DAISYTEK INTL CORP	2001	0.260	0.766	PURCHASEPRO.COM	2001	-0.006	0.836
13	TRIDENT MICROSYSTEMS INC	1997	0.259	0.806	ASCENTIAL SOFTWARE CORP	1997	0.181	0.831
14	MERCURY INTERACTIVE CORP	1999	0.258	0.761	BROADCOM CORP	2002	0.210	0.831
15	BROADCOM CORP	1999	0.254	0.705	VALEANT PHARMACEUTICALS INTL	2002	-0.020	0.830
16	SPORT-HALEY HOLDINGS INC	2000	0.252	0.616	MERCURY INTERACTIVE CORP	2001	0.150	0.829
17	SYCAMORE NETWORKS INC	2000	0.252	0.749	ASCENTIAL SOFTWARE CORP	1998	0.244	0.828
18	UNIVERSAL CORP/VA	1997	0.251	0.612	CLEARONE COMMUNICATIONS INC	2001	0.226	0.827
19	MICROSTRATEGY INC	1998	0.249	0.824	BLUE COAT SYSTEMS INC	2005	0.011	0.826
20	TRIDENT MICROSYSTEMS INC	2005	0.246	0.802	MARVELL TECHNOLOGY GROUP LTD	2006	0.109	0.825
21	UNIVERSAL CORP/VA	2000	0.246	0.561	MICROSTRATEGY INC	1998	0.249	0.824
22	TENFOLD CORP	1999	0.245	0.805	ENDOCARE INC	2001	-0.064	0.824
23	ASCENTIAL SOFTWARE CORP	1998	0.244	0.828	ANALOG DEVICES	2001	0.173	0.823
24	QUEST SOFTWARE INC	2000	0.242	0.797	TRIDENT MICROSYSTEMS INC	1998	0.225	0.822
25	COMVERSE TECHNOLOGY INC	1999	0.240	0.761	MARVELL TECHNOLOGY GROUP LTD	2000	0.270	0.822

Figure 1: Empirical distribution of firm Fraud Scores. 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.

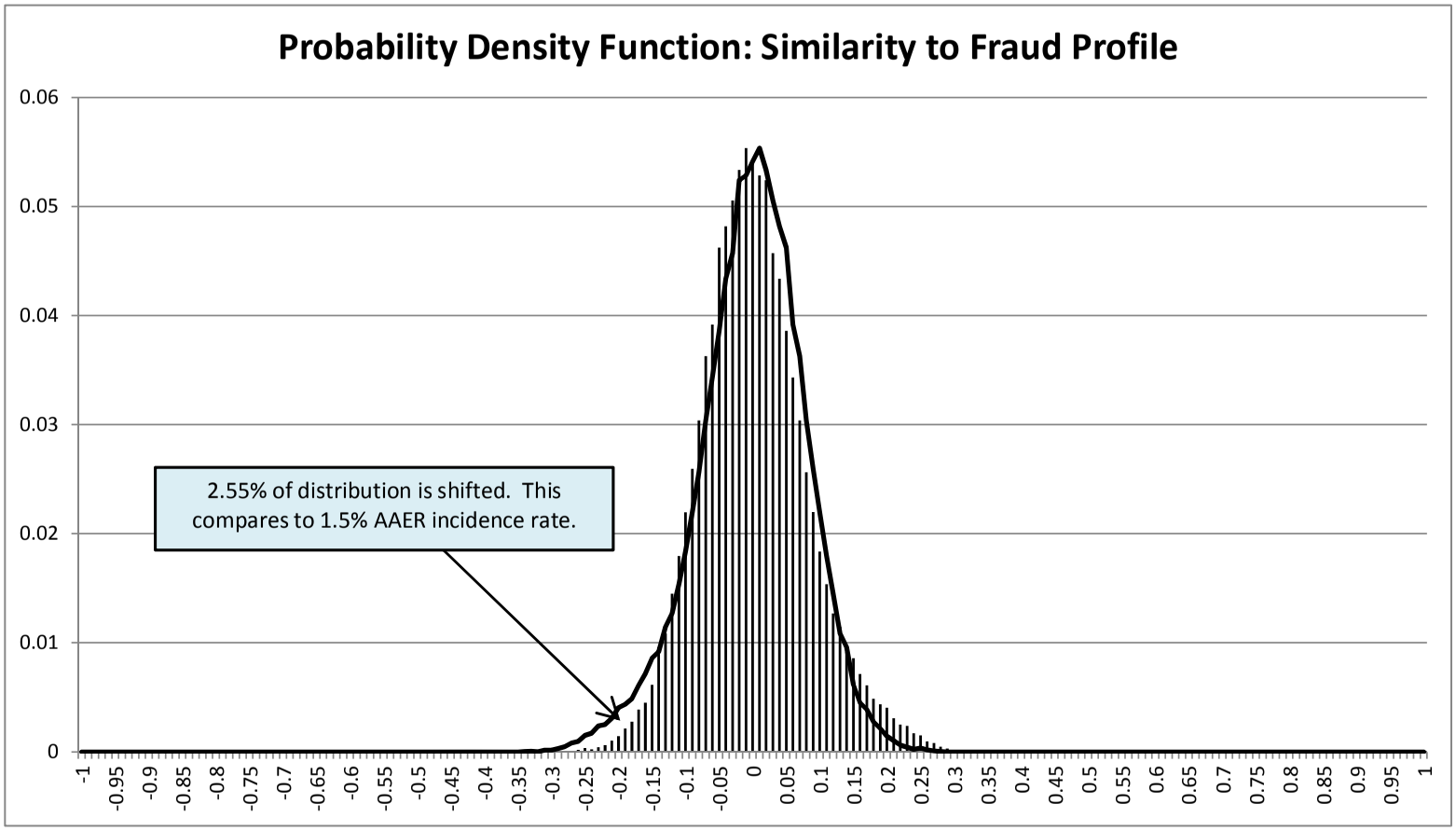
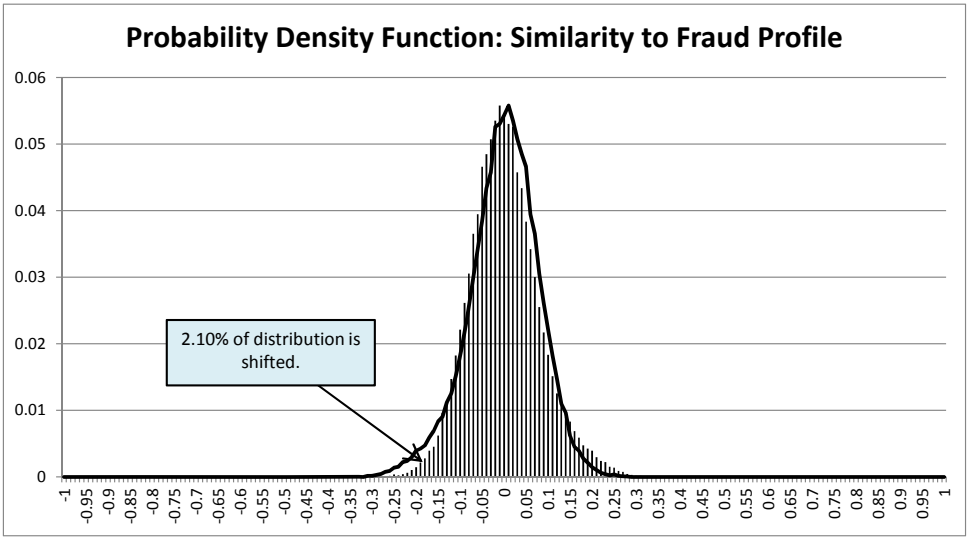


Figure 2: Empirical distribution of firm Fraud Scores for two subsamples. The upper figure's distribution is based on all firms in our sample excluding firm years involved in AAERs. The lower figure reports the fraud score distribution only for firms-years involved in AAERs. In both figures, the actual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



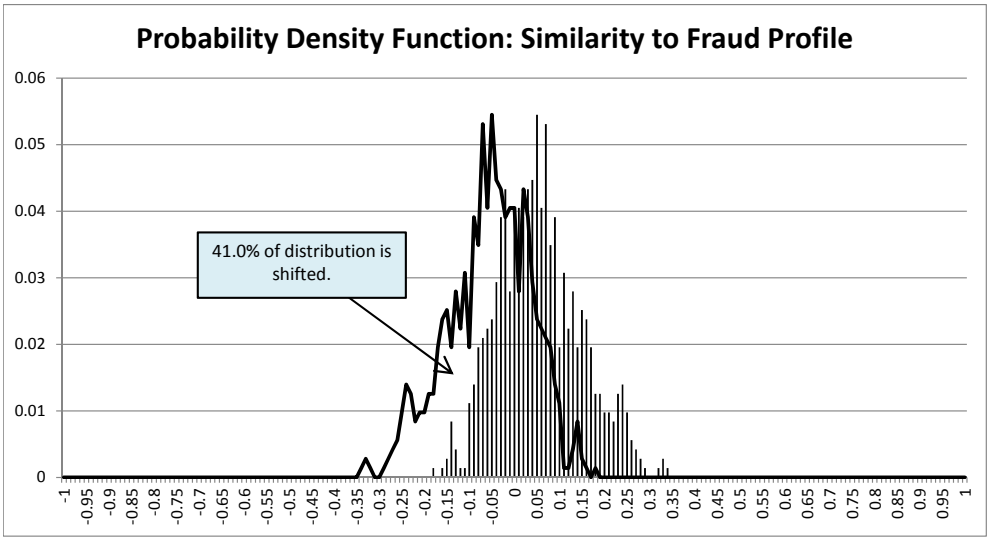


Figure 3: Average Fraud Scores over time for firms involved in AAERs. The figure displays the average fraud score (localized anti-herding) and the industry similarity score (industry herding) during the period of time that the AAER alleges fraud occurred, and also during the period of time preceding 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.