

Information Content of Public Firm Disclosures and the Sarbanes-Oxley Act

Preliminary and Incomplete

Shimon Kogan

Red McCombs School of Business, University of Texas at Austin
Austin, TX 78712, USA
shimon.kogan@mcombs.utexas.edu
Tel: 512.232.6839

Bryan R. Routledge

Tepper School of Business, Carnegie Mellon University
Pittsburgh, PA 15213, USA
rout@andrew.cmu.edu
Tel: 412.268.7588

Jacob S. Sagi

Owen Graduate School of Management, Vanderbilt University
Nashville, TN 37203, USA
Jacob.Sagi@Owen.Vanderbilt.edu
Tel: 615.343.9387

Noah A. Smith

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
nasmith@cs.cmu.edu
Tel: 412.268.4963

October, 2009

Abstract

We find evidence that public firm disclosure, in the form of Management Discussion and Analysis (Sections 7 and 7a of annual reports), is more informative about the firm's future risk following the passage of the Sarbanes-Oxley Act of 2002. Employing a novel *text regression*, we are able to predict, out of sample, firm return volatility using the Management Discussion and Analysis section from annual 10-K reports (which contains forward-looking views of the management). Using the relative performance of the text model as a proxy for the informativeness of reports, we show that the MD&A sections are significantly more informative after the passage of SOX. We further show that this additional information is associated with a reduction in share illiquidity, suggesting that the information divulged was new to investors. Finally, we find that the increase in informativeness of MD&A reports is most pronounced for firms with higher costs of adverse selection.

1. Introduction

In the summer of 2002, the accounting scandals and subsequent bankruptcies of major US blue chip companies, along with the growing trend in accounting restatements, prompted Congress to pass the Sarbanes-Oxley Act (SOX). The Public Company Accounting Reform and Investor Protection Act of 2002, as it was otherwise known, did not enact new legislation as much as it attempted to reinforce existing laws that were not being satisfactorily enforced. This was done by requiring companies and their auditors to disclose and certify internal controls for the prevention of financial fraud and materially misleading financial statements, and by making auditors and officers of the companies they audited directly accountable for the accuracy of such disclosures. Various other provisions of the Act included the creation of a supervisory body to monitor auditors, the mandatory inclusion of a ‘financial expert’ on the audit committee, ‘whistle blower’ protection, and new regulations concerning conflict of interest between auditors and firms.

While, according to some studies, post-SOX mandatory disclosures of weakness in internal controls appear to affect stock prices (Hammersley, Myers, and Shakespeare, 2008), suggesting that such disclosures are material to investors, it is not clear whether SOX has contributed to more informative disclosure *outside* the narrow scope mandated by the Act.¹ We argue in this paper that the impact of the SOX legislation went beyond the intended prevention of fraud and misrepresentation by public firms and their auditors. We are able to demonstrate that, post-SOX, annual reports contain information allowing one to better forecast firms’ return volatility. Moreover, this additional forecasting power is not apparently related to governance and reporting issues that are the intended purview of SOX. We employ a novel *text regression*, discussed in Kogan, Levin, Rutledge, Sagi, and Smith (2009), to forecast firm return volatility using the Management Discussion and Analysis (MD&A) section from annual 10-K reports (which contains forward-looking views of the management). Our text-based forecast of volatility significantly improves after 2002.

By examining the cross-section, we explore various explanations for this apparent SOX effect. We find that firms in which information production is costly, such as firms with low market

¹Ogneva, Raghunandan, and Subramanyam (2007) find no evidence that the types of disclosures discussed in Hammersley, Myers, and Shakespeare (2008) are associated with a change in the cost of capital of a firm. Bhattacharya, Groznik, and Haslem (2007) found no evidence that CEO certification impacted share prices.

capitalization, high book-to-market ratio, low analyst coverage, or high analysts forecast dispersion, experience the largest improvement in the post-SOX performance of the text-based model. It appears that after the adoption of SOX, firms where asymmetric information might be highest divulged more information (in their MD&A reports) that is useful in predicting future return volatility. Indeed, we also find a significant relationship between improvement in the informational content of MD&A reports and a reduction in share illiquidity.

The paper proceeds as follows. Section 2 provides a literature review. Section 3 first discusses the research questions we explore in the paper, outlines the methodology of text regression, and describes the dataset used in our empirical analysis. Section 4 reports the various results from our text regressions and the cross-sectional analysis vis à vis the hypotheses in Section 3. Section 5 concludes with a discussion about whether firms divulged more information because they learned more or whether it was because managers decided to reduce expected personal liability.

2. Background

This paper investigates the informational content in textual disclosures pre- and post-SOX. As such, we overlap two distinct literatures. The first deals with the impact of SOX on US corporations. The second is concerned with reducing text-based information into cross-sectional and time-series variables that can be of use in financial economics.

2.1. The Sarbanes-Oxley Act of 2002

The Sarbanes-Oxley Act of 2002 was enacted on July 30, 2002, as a response to a number of accounting scandals (e.g., Enron). The act contains eleven “titles” that pertain to auditors, the firm executives, and analysts, and each title is divided into multiple sections. The sections most relevant for this paper are:

- *Section 302: Corporate responsibility for financial reports* — requiring that CEOs and CFOs certify that they have read the company’s financial reports, and that the latter depict a fair representation of the company’s financial situation and do not contain untrue or misleading

information. This section also asserts that the signing executives are responsible for establishing and maintaining effective internal controls for fraud prevention, and reporting on the effectiveness of these controls.

- *Section 401: Disclosures in periodic reports* — requiring that firms disclose off-balance sheet transactions and attest to the completeness and accuracy of their pro forma financial statements.
- *Section 404: Management assessment of internal controls* — requiring annual reports to include a discussion on internal controls, highlighting any material weaknesses, over financial reporting. The discussion must be accompanied by an auditor attested assessment of the effectiveness of the company's internal controls.
- *Section 906: Corporate responsibility for financial reports* — outlines the criminal penalties for executives who fail to comply with the requirements set forth by the act.

A number of academic studies have examined the fallout from SOX. Piotroski and Srinivasan (2008) find no evidence that SOX imposed sufficient net costs on foreign firms to dissuade them from listing in the United States. This is corroborated by Doidge, Karolyi, and Stulz (2008). Another line of literature questions whether the net costs imposed by SOX have caused delisting within the US or have negatively impacted the value of public firms.² Leuz (2007) argues that evidence for these hypotheses is inconclusive.³ The evidence on how SOX has affected firms' values is mixed.⁴ In terms of theoretical work on this topic, Goldman and Slezak (2006) state that when managers' compensation has an equity-stake component, disclosure requirements may lead to an increase in manipulation. This is somewhat orthogonal to our question because we are examining an increase in non-mandatory disclosure.

²A 2005 survey by Charles River Associates suggests that Fortune 1000 companies spent about \$6 million in their first year of compliance with internal controls requirements of the Act.

³Leuz (2007) criticizes two other papers (published earlier in the same journal) that claimed to find significant negative repercussions to SOX. Engel, Hayes, and Wang (2007) claim that SOX caused firms to delist, while Zhang (2007) claimed that the market value of firms subsequent to the passage of SOX suffered a major loss.

⁴See the studies cited in the Introduction: Bhattacharya, Groznik, and Haslem (2007); Hammersley, Myers, and Shakespeare (2008); Ogneva, Raghunandan, and Subramanyam (2007).

2.2. *Textual analysis in financial economics*

The majority of research in finance exclusively employs quantitative information about firm attributes, usually gathered from annual reports or other public sources, and consolidated into databases such as CRSP and COMPUSTAT. Recently, various researchers in financial economics have explored the wealth of information available in text format. The most common approach is to reduce textual material, such as a single report or an article, to a univariate measure by calculating the number of ‘positive’ words and subtracting from that the number of ‘negative’ words.⁵ Such a measure has been shown to contain forecasting power for earnings, stock returns, return volatility, and trading volume (see Koppel and Shtrimberg, 2004; Tetlock, 2007; Tetlock, Saar-Tsechansky, and Macskassy, 2008; Gaa, 2007; Engelberg, 2007). In a similar spirit, Li (2005) measures the association between frequency of words related to risk and subsequent stock returns. In a different application of textual analysis, Das and Chen (2001) and Antweiler and Frank (2004) examine whether message board postings predict stock performance. Other researchers have found relationships between the attributes of textual data and initial public offerings, investor sentiment and investor opinions (see Weiss-Hanley and Hoberg, 2008; Pang, Lee, and Vaithyanathan, 2002; Wiebe and Riloff, 2005; Lerman, Gilder, Dredze, and Pereira, 2008). Finally, Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, and Allan (2000b) and Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, and Allan (2000a) modeled influences between text and time series financial data (stock prices) using language models.

One criticism of approaches to textual data that pre-select words according to some valence criteria is that they potentially ignore other information that might be informative. In this regard, our approach is somewhat unique in this budding field because we allow for a higher dimensionality of textual information and, in that respect, let the data ‘speak for itself’. The type of regression technique we introduce builds on the Support Vector Regression literature pioneered by Drucker, Burges, Kaufman, Smola, and Vapnik (1997). Our approach follows considerable research in information retrieval and text categorization in representing the entirety of a document as a vector of (transformed) word counts, sometimes called a “bag of words.” While this is an obvious sim-

⁵Words are typically assigned a positive or negative association based on a dictionary compiled by psychologists (e.g., the ‘Harvard-IV-4 dictionary’).

plification of the meaning of a piece of text, it is known to work quite well for many applications. Further, the same learning method can be applied with text representations that aim to take more knowledge into account (e.g., by filtering words or emphasizing certain phrases) or deeper linguistic structure (e.g., predicate-argument relations between words or discourse markers). Indeed, the task of making a quantitative forecast (e.g., predicting volatility) can be seen as a new application of techniques in natural language processing for automatic understanding of text.

3. Methodology and data

3.1. Research questions

The annual reports that public firms in the United States are required to file with the SEC typically contain a section dealing with Management's discussion and analysis of financial conditions and results of operations, and a section containing quantitative and qualitative disclosures about market risk. These are, respectively, Sections 7 and 7a, and often contain forward looking information about a firm's financial and operational situation.

As Table 1 indicates, MD&A reports have nearly doubled in the years following SOX, growing from an average of roughly 5,000 words per document between 1996 and 2001 to over 10,000 per document between 2002 and 2006.⁶ Part of this could be explained by an additional ruling in 2003 by the Securities and Exchange Commission (SEC) concerning SOX (Section 401) that mandates companies report off balance sheet transactions and obligations in Section 7 of the annual report.⁷ However, it seems unlikely that such a new mandate would lead to a doubling in the size of MD&A.

There are other rationales for the growth in report sizes. As noted by Wagner and Dittmar (2006),

...Sections 302 and 404...require CEOs and CFOs to attest personally to the effectiveness of internal control over financial reporting, and Section 906...makes willful failure to portray the true condition of the company's operations and finances a crime.

⁶The difference in a two-means test yields a *t*-statistic of approximately 6.0.

⁷See <http://www.sec.gov/news/press/2003-10.htm>.

Year	(1) No. of documents	(2) Ave. words/doc.	(3) Med words/doc.	(4) Max words/doc.
1998	2,357	4,913	3,972	29,721
1999	2,391	5,921	4,912	36,273
2000	2,323	5,691	4,578	30,260
2001	2,419	6,108	5,083	33,600
2002	2,652	8,220	6,963	102,145
2003	3,410	10,045	8,589	56,940
2004	3,500	11,191	9,672	107,783
2005	3,420	12,255	10,642	92,198
2006	3,280	11,857	9,905	207,049
Pre-SOX	12,142	6,221	5,044	102,145
Post-SOX	13,610	11,332	9,674	207,049

Table 1: Characteristics of the reports (Sections 7 and 7a from annual reports) used in this paper.

Because there may be several ways to interpret the increase in the size of Sections 7 and 7a beyond what might be required by Section 401, we are led to ask the following questions:

1. *Is there any evidence that the increase in size was accompanied by an increase in informativeness?* The cynical view of the effect of SOX is that some or all of the increase in liability has simply caused management to pad Sections 7 and 7a with legalese. This view would suggest that post-SOX annual reports do not contain information that is useful *outside* of the narrow scope mandated by the Act (i.e., useful information that is not associated with internal controls and/or off balance sheet transactions).
2. *Is there any evidence that information new to the MD&A reports was new to investors?* Related to the previous question, one might suspect that “useful” additional information in MD&A sections that is not directly related to SOX is also not new to investors (i.e., such information might have been available elsewhere before SOX). Rather, in order to avoid the possibility of legal action, managers include such information whereas they did not bother to do so before SOX.
3. *What type of firms chose to disclose more information?* Sections 401 and 404 effectively require that a greater degree of senior managerial resources be devoted to identifying and

understanding various risks. Such an enterprise could lead to a greater understanding by managers of their companies' operational and financial risks that are not directly linked to internal controls. If investors expect managers to know more about their companies as a result of SOX, failure to disclose more information could lead to an adverse selection problem. Firms facing more internal opacity might stand to learn more from compliance with SOX and, to avoid an adverse selection problem, might choose to increase the informativeness of their reports.

On the other hand, Sections 302 and 906 required officer certification for the accuracy of statements made in the MD&A reports (and elsewhere), and threatened officers with criminal liability for failure to comply. This might have prompted managers of firms in which there is more asymmetric information between investors and insiders to improve post-SOX disclosure in MD&A sections.

These observations suggest investigating the cross-sectional variation of MD&A informativeness with respect to proxies for internal opacity and asymmetric information.

Our goal is to attempt to address these three questions. To do so, we first have to identify a measure of informativeness. The strategy we employ is to focus on a firm-specific variable that can be forecast using textual data. We choose the firm's return volatility for various reasons. First, because in a previous investigation MD&A text was shown to have predictive power for volatility (Kogan, Levin, Routledge, Sagi, and Smith, 2009). Second, firms' return volatility was not directly targeted by the SOX legislation or its SEC implementation (i.e., the goal of the legislation is to reduce fraud and not to improve investors' ability to predict a firm's return volatility, *per se*). Finally, while volatility forecasting is known to be of great importance to market participants, the possibility of finding forecasting power for volatility in publicly available data is economically uncontroversial.⁸ As such, the ability of MD&A text to predict return volatility allows us to ask whether MD&A informativeness has increased post-SOX, whether such an increase is associated with the narrow requirements set forth by SOX, whether or not increased informativeness appears

⁸Our study is about the information content in textual data, and not about efficient markets, *per se*. Thus, we wish to avoid variables for which market forces tend to work against attempts to forecast their direction (e.g., equity return forecasting and the "efficient markets hypothesis").

new to (and therefore impacts) investors, and whether or not it is associated with a reduction in costs of asymmetric information.

In the next subsection we review the technique we employ to predict volatility from textual data. We follow that with a description of our data sources and an outline of the empirical tests we adopt to answer the questions above.

3.2. *Text Regressions*

Our approach in constructing a measure of the informativeness of an MD&A report is to use the textual data in the MD&A of each firm to forecast a firm-related attribute (e.g., return volatility). The technique we use, introduced in Kogan, Levin, Routledge, Sagi, and Smith (2009) is a *text regression*, in which textual data from each MD&A report is first converted into a vector whose elements are frequencies (or a function of the frequency) of words or word combinations.⁹ We then estimate an attribute of the firm (e.g., return volatility) as a linear function of this vector of frequencies. The weightings chosen for the ‘best’ estimate minimize an objective function that balances goodness of fit against the propensity to ‘over-fit’ (a procedure known as regularization), and using a large training set of reports (e.g., all available MD&A reports from the years 1996 and 1997). The weightings thus obtained can subsequently be used to produce forecasts of the firm attribute *out of sample*. The squared error in the out-of-sample forecasts furnishes us with a measure of informativeness that we can benchmark to either non-textual forecasts or to prior forecasts (e.g., comparing squared errors before and after SOX).

Let $d_{i,t}$ denote a portion of text associated with firm i at date t . In our specific setting, each portion of text corresponds to Sections 7 and 7a of firm i ’s annual report—which we collectively refer to as the ‘MD&A’ section. Let M be the number of terms in a lexicon that contains all of the relevant terms in all the MD&A reports. One can uniquely associate any term that appears in a report with a number, j , between 1 and M . Let $f_j(d_{i,t})$ be the frequency of the j^{th} word in text $d_{i,t}$. In many instances, $f_j(d_{i,t})$ will be zero because the j^{th} term does not appear in the text $d_{i,t}$. We ignore punctuation, digits, and letter cases in calculating $f_j(d_{i,t})$. One can, and in some instances

⁹See also Ghose, Ipeiritos, and Sundararajan (2007).

we do, include word combinations as well as single words as terms in the lexicon, in which case M corresponds to the size of the set of all admissible one or two word combinations.¹⁰

Suppose one wishes to forecast a firm-specific attribute, $y_{i,t+1}$ using $f_j(d_{i,t})$, $j = 1, \dots, M$. In particular, we consider the following forecasting model:¹¹

$$\text{LOG1P} \quad \hat{y}_{i,t+1} = b + \sum_{j=1}^M w_j \log(1 + f_j(d_{i,t})). \quad (1)$$

where the w_j 's correspond to weights.

To estimate the weights, we follow Drucker, Burges, Kaufman, Smola, and Vapnik (1997) and minimize the following function with respect to a training set of documents:

$$\min_{\substack{w_j: j=1, \dots, M \\ t: T_1, \dots, T_2}} \left\{ \frac{1}{2} \sum_{j=1}^M w_j^2 + \frac{C}{N} \sum_{i=1}^N \underbrace{\max(0, |y_{i,t} - \hat{y}_{i,t}| - \epsilon)}_{\epsilon\text{-insensitive loss function}} \right\} \quad (2)$$

where N is the number of documents in the training set, the y_i 's correspond to *realized* value of the variable to be forecasted, and the estimation is performed using a subsample (the training set). Roughly, the first sum serves to 'regularize' the estimated weights so as to prevent over-fitting. The regularization is controlled by C . When $\epsilon > 0$, words whose frequency has little relation to the forecasted variable will tend to be assigned zero weight. We used a freely available implementation of this procedure, called $\text{SVM}^{\text{light}}$ (Joachims, 1999).¹² In our estimation, we set C using the default

¹⁰Combinations of two words are referred to as *bigrams*.

¹¹We also experimented with the following forecasting models:

$$\begin{aligned} TF \quad \hat{y}_{i,t+1} &= b + \sum_{j=1}^M w_j \frac{f_j(d_{i,t})}{|d_{i,t}|}, \\ \text{TFIDF} \quad \hat{y}_{i,t+1} &= b + \sum_{j=1}^M w_j \frac{f_j(d_{i,t})}{|d_{i,t}|} \log\left(\frac{N}{Q_j}\right). \end{aligned}$$

where $|d_{i,t}| = \sum_j f_j(d_{i,t})$; while N is the number of documents in a 'training set' (used to estimate the weights) and Q_j is the number of documents in the training set that contain at least one instance of the j^{th} word. Overall, the LOG1P model performed the best (in sample).

¹²Available at <http://svmlight.joachims.org>.

choice in SVM^{light} , while ϵ is set at 0.1.¹³

3.3. Data description

The textual data consists of 40,865 MD&A sections (corresponding to Sections 7 and 7a) taken from the annual reports of 8,393 publicly traded U.S. firms over the years 1996-2006. The MD&A text was collected automatically using a Perl script that attempts to identify the beginning of Sections 7, 7a, and 8, and subsequently captures the text between Sections 7 and 8 if it consists of more than 1000 words. From each captured MD&A section, we remove all HTML ‘mark-up’ code (e.g., “<P>” and “<A HREF=”). We also remove all punctuation (commas, etc.). All words are converted to lowercase (e.g., “Enron” becomes “enron”). All numbers are collapsed to the character ‘#’ (e.g., “\$5,000” and “\$150” both become “\$#”).¹⁴ This procedure, referred to as tokenization of the text, is standard and was performed with a short Perl script.¹⁵ We do not employ ‘stemming’ (e.g., collapsing words like “looking” and “looks” to the same stem word, “look”.) Not all annual reports downloaded from the SEC pass the filter imposed by our Perl scripts to yield an MD&A report. In some instances, the reports were too short, while in others they were only available separately (i.e., not in the main body of the annual report available online). We estimate that over 80% of the annual reports available from the SEC for the years studied pass the Perl script filter.

We further restrict the set of firms in this study to include only those for which return data is available from CRSP and firms with market capitalization higher than \$10M (i.e., we exclude ultra micro-cap firms). Table 1 summarizes the text documents, in their final (tokenized) format, that we include in this study.

In addition to the textual data, we also use the CRSP daily return database, COMPUSTAT for accounting data, IBES for analyst coverage and forecasts, and the RiskMetrics Governance Data for the corporate governance index.¹⁶ Share liquidity information in the form of the Amihud-

¹³Experimenting with other values did not yield significant gains with in-sample forecasting.

¹⁴We eliminate numbers from the reports because our focus is on information content in *textual* data. We take the view that better sources for quantitative data exist in consolidated formats (e.g., the COMPUSTAT datasets, etc.).

¹⁵The Perl script used to extract and tokenize the text data, as well as the original, extracted, and tokenized data, are available at <http://www.ark.cs.cmu.edu/10K>.

¹⁶We wish to thank Ilan Guedj, Jennifer Huang, and Johan Sulaeman for providing us with the data on firms’ Herfindahl index and analyst data.

ratio (Amihud, 2002) and Hasbrouck’s Gibbs measure (Hasbrouck, 2006) are taken from Joel Hasbrouck’s publicly available database.¹⁷

3.4. *A description of the empirical methodology and summary statistics*

We begin by applying the model in Eq. (1) to predicting the log-volatility of each firm and using this forecast to construct a measure of the informativness of an MD&A section. Specifically, take $y_{i,t} = \log(\sigma_{i,t})$, where $\sigma_{i,t}$ is the daily return volatility for firm i over 12 months (where the last month is the month prior to the publication of the annual report in year t). To estimate the coefficients in (2) for the purpose of forecasting the volatility of firms at year t , we use training data from the two previous years. To assess the forecasting power, we calculate the mean of the squared error (MSE) over all firms in a given year:

$$\text{MSE}_t = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{i,t} - \hat{y}_{i,t})^2, \quad (3)$$

where n_t is the number of firms for which data exists at date t .

The MSE obtained from the text model can be compared to other log-volatility forecasts. We use the past year’s realized log-volatility as a forecast of the following year’s volatility. This is consistent with the acknowledged persistence of realized volatility (i.e., past volatility contains information about future volatility) and serves as a benchmark forecast.¹⁸ Following Kogan, Levin, Routledge, Sagi, and Smith (2009), we document in Table 2 the MSEs in the years 1998-2006 for the benchmark forecast and for the text-based forecast.

Table 2 establishes several key things. First, keeping in mind that the forecasts are out of sample, it is apparent that the text-based forecasts are comparable to forecasts based solely on past volatility. This provides evidence that MD&A sections contain information about future stock return volatility. Thus, such a text-based forecast may be a suitable instrument for investigating changes in the informativeness of MD&A sections pre- and post-SOX. Second, the average

¹⁷See <http://pages.stern.nyu.edu/jhasbrou/>.

¹⁸We also used a GARCH(1, 1) model (Engle, 1982; Bollerslev, 1986) and found that, at the horizon of one year, it was not significantly better than past realized volatility at forecasting.

Year	(1) Mean daily vol.	(2) Mean log daily vol.	(3) Historical vol. MSE	(4) Text model MSE
1998	3.71%	-3.20	0.234	0.229
1999	4.61%	-3.14	0.154	0.164
2000	4.89%	-3.08	0.153	0.162
2001	5.34%	-3.23	0.175	0.212
2002	4.45%	-3.29	0.157	0.200
2003	4.04%	-3.60	0.188	0.178
2004	3.19%	-3.72	0.143	0.143
2005	2.70%	-3.78	0.135	0.128
2006	2.53%	-3.83	0.147	0.145
Pre-SOX	4.60%	-3.19	0.174	0.194
Post-SOX	3.12%	-3.73	0.153	0.148

Table 2: Summary statistics of daily volatility and Mean Squared Errors (MSEs) for the benchmark volatility forecast and for the text-based forecast. The benchmark forecast is the past year’s realized volatility. The text-based forecast is described in Eq. (1). Volatilities are quoted as daily (a daily volatility of 3% corresponds to roughly 50% annualized volatility).

volatility and the benchmark forecast MSE have decreased from the pre- to the post-SOX period. One may therefore be concerned that a decrease in the post-SOX MSE of the text-based volatility forecast may be due to something other than increased informativeness of the MD&A reports.

To control for the possibility that volatility might be easier to forecast in some periods, for reasons other than better disclosure, we define a *relative* measure of MD&A informativeness, which we henceforth refer to as “informativeness”, as follows: For firm i at date t ,

$$\mathcal{J}_{i,t} \equiv (y_{i,t} - y_{i,t-1})^2 - (y_{i,t} - \hat{y}_{i,t})^2. \quad (4)$$

The first term in Eq. (4) is the squared error from using the benchmark forecast, while the second term is the squared error from using the text-based forecast. The measure $\mathcal{J}_{i,t}$ increases as the text-based forecast improves relative to the benchmark forecast. Thus, $\mathcal{J}_{i,t}$ captures the informativeness in an MD&A report net of that present in past volatility. It should be clear that $\mathcal{J}_{i,t}$ will be noisy. Ideally, one would want to take an expression such as the right side of Eq. (4) and average over a long time-series in order to reduce the noise. Because we do not have a long time-series or a

balanced panel, we instead opt for a more noisy measure, hoping that the large size of the cross-section will allow for statistical discrimination.

Using $\mathcal{J}_{i,t}$ as a measure of informativeness, we can address the the questions posed in Section 3.1. To answer the first question we compare the informativeness pre- and post-SOX. We also examine the words with the highest weight magnitudes to see whether the text-based model is driven primarily by SOX related words. Finally, we compare the change in informativeness (pre- and post-SOX) for firms that were required to comply with all of the provisions set out by SOX versus firms that had minimal compliance requirements.

To address the second question in Section 3.1, we look to see whether increases in informativeness were accompanied by a reduction in asymmetric information. To do this, we examine the relationship between changes in the illiquidity measures of publicly traded shares and changes in the informativeness measure. As proxies for share illiquidity, we select the Amihud ratio, a measure of price impact (Amihud, 2002) and the Gibbs measure, a measure of the bid-ask spread (Hasbrouck, 2006).

The third question in Section 3.1 essentially asks about the role of asymmetric information in prompting firms to divulge more information. To explore this, we identify a set of variables that may proxy for the presence of asymmetric information and investigate the relationship between these variables and changes in MD&A informativeness in the pre- versus post-SOX periods.

4. Empirical analysis

4.1. *Did the informativeness of MD&A sections improve after SOX?*

Consider the simple panel regression with firm random effects,

$$\mathcal{J}_{i,t} = \text{const} + b_i \text{SOX}_t + c_i \text{LnMDA}_t + \epsilon_{i,t}. \quad (5)$$

where SOX_t is a dummy variable taking the value of 0 before 2003, and the value of 1 in the years 2003 and beyond. The variable LnMDA is the log-frequency of the words in the MD&A

document. The t -stat on the dummy variable in this regression is positive and highly significant ($t \sim 11$), indicating that, in aggregate, the informational content in MD&A reports has increased in the post-SOX period along with the size of the reports.¹⁹ In fact, we find that the text model underperforms the historical volatility model pre-SOX but significantly outperforms it post-SOX. Our proxy for informativeness, $\mathcal{J}_{i,t}$, is indeed noisy as evidenced by the low regression R^2 (0.003).

Is the change due to disclosure requirements specified by the legislation? We address this question by looking at SOX related words. Table 3 shows the frequency of occurrence of different SOX related words. As expected, we see that the usage of these words increases dramatically after 2002. This is also the case for words that do not include “Sarbanes” or “Oxley” explicitly, such as “Internal Control”, “Compliance”, “Fraud”, and “Off Balance Sheet”.

At the same time, as Table 4 shows, these words do not appear to be particularly important in forecasting volatility. The table lists the words associated with the highest magnitude weights resulting from the text model with bigrams.²⁰ As the table suggests, SOX-related words are not strongly weighted either before or after the legislation.

Further, we show that the frequency of SOX related words does not explain the performance of the text model. Table 5 reports a series of panel regressions using the log-frequency (i.e., $\log(f_j(d_{i,t}))$) of the SOX-related words as independent variables and $\mathcal{J}_{i,t}$ as the dependent variable. We include a SOX dummy and, for some specifications, a variable corresponding to the logarithm of the number of words in a report and/or the logarithm of the total number of SOX-related words in the report. While the frequency of these words explains some of the improvement in the forecasting power of the text-based model, the SOX dummy remains strongly significant. Further, the explanatory power of the regressions increases only marginally when including the frequencies of the SOX related words. This suggests that much of the increase in informativeness is not directly related to the reporting requirements of the legislation, narrowly defined.

Finally, we use a natural experiment imposed by the SEC implementation of SOX to show that even firms which were not required to include new information in their reports were affected

¹⁹The results of this regression are reported in the first column of Table 5 along with the results of other regressions discussed in this section.

²⁰Recall that positive (negative) weights are associated with higher (lower) forecasted volatility

Year	Sarbanes Oxley	Internal Control	Officer Certification	Deficiency	Weakness	Balance Sheet	Compliance	Fraud	Off Balance Sheet
1998	0	1,403	0	70	106	585	1,508	56	124
	0	3,322	0	133	148	1,141	4,409	109	258
1999	1	1,964	0	84	127	878	2,087	62	170
	1	6,910	0	141	195	1,704	10,077	110	360
2000	0	1,451	0	82	122	921	1,390	60	196
	0	3,579	0	165	186	1,822	3,645	116	430
2001	2	1,116	0	69	142	1,107	1,003	64	187
	2	2,511	0	139	206	2,324	2,121	113	389
2002	24	1,383	0	101	330	1,517	1,360	113	495
	34	3,511	0	170	531	3,919	3,392	223	835
2003	196	1,979	0	198	525	2,253	2,003	200	1,028
	313	5,375	0	380	974	6,751	5,811	462	1,891
2004	633	2,286	0	216	546	2,572	2,237	254	1,957
	1,028	7,001	0	407	1,019	8,664	7,195	710	3,925
2005	2,231	3,011	0	273	629	2,583	2,625	350	2,138
	4,042	13,681	0	541	1,280	8,861	9,849	1,008	4,291
2006	2,086	2,674	1	217	402	2,509	2,486	287	2,268
	3,582	11,117	1	471	726	8,732	8,957	765	4,645

Table 3: Frequencies of SOX-related words appearing in the MD&A reports. The first row corresponds to the total number reports (in a given year) where the words appear, and the second row corresponds to the total number of occurrences. We use “_” to denote a whitespace between two words. The column labeled “Sarbanes Oxley” adds up the frequencies of the terms “sarbanes”, “oxley”, “sarbanesoxley”, “sarbanes_oxley” and “sox”; the column labeled “Internal Control” adds up the frequencies of the terms “internal”, “controlsystem”, and “internalcontrol”; the column labeled “Officer Certification” adds up the frequencies of the terms “officer_certification” and “officercertification”; the column labeled “Deficiency” adds up the frequencies of the terms “deficiency” and “materialdeficiency”; the column labeled “Weakness” adds up the frequencies of the terms “weakness” and “materialweakness”; the column labeled “Balance Sheet” adds up the frequencies of the terms “balancesheet”, “balance_sheet”; the column labeled “Off Balance Sheet” adds up the frequencies of the terms “offbalance”, “offbalancesheet”, “offbalance_sheet”, and “off_balancesheet”.

	Informativeness (J)				
SOX	0.029*** [0.003]	0.030*** [0.003]	0.029*** [0.003]	0.031*** [0.004]	0.030*** [0.004]
LnMDA	0.007** [0.003]		0.003 [0.004]	0.006 [0.004]	0.005 [0.004]
LnSOX		0.006*** [0.002]	0.005* [0.002]		0.018*** [0.006]
sarbanes				0.096* [0.050]	0.097* [0.050]
oxley				-0.016 [0.120]	-0.012 [0.120]
sarbanesoxley				0.002 [0.005]	0.000 [0.005]
sarbanes_oxley				-0.080 [0.116]	-0.091 [0.116]
internal				0.005* [0.003]	-0.004 [0.004]
controlsystem				-0.001 [0.015]	-0.003 [0.015]
officer_certification				0.018 [0.319]	0.003 [0.319]
deficiency				0.009 [0.008]	0.006 [0.008]
weakness				-0.008 [0.006]	-0.012** [0.006]
balancesheet				-0.018 [0.059]	-0.022 [0.059]
balance_sheet				0.002 [0.003]	-0.005 [0.004]
compliance				0.000 [0.002]	-0.009** [0.004]
fraud				-0.001 [0.007]	-0.002 [0.007]
sox				-0.038*** [0.014]	-0.037*** [0.014]
internalcontrol				-0.002 [0.006]	-0.002 [0.006]
materialweakness				-0.021 [0.016]	-0.017 [0.016]
materialdeficiency				-0.098 [0.165]	-0.090 [0.165]
offbalance				-0.006 [0.008]	-0.010 [0.008]
offbalancesheet				0.004 [0.012]	0.001 [0.012]
offbalance_sheet				-0.005 [0.004]	-0.010** [0.004]
off_balancesheet				0.034 [0.110]	0.032 [0.110]
Constant	-0.033*** [0.004]	-0.058* [0.031]	-0.058* [0.031]	-0.077** [0.033]	-0.079** [0.033]
Number of Obs.	25,752	25,752	25,752	25,752	25,752
R ²	0.003	0.003	0.003	0.004	0.004

Table 5: The table reports panel regressions (with firm random effects) using MDA informativeness as the dependent variable. “SOX” is a dummy variable taking the value of 1 for the post-SOX years (2003 and on). “LnMDA” is the log frequency of MDA words, “LnSOX-Words” is the log frequency of SOX related words in the MDA. All other independent variables are computed as the log frequencies of the corresponding words.

by the passage of the act. Until 2007, the SEC exempted smaller public companies (those with market cap lower than \$75M) from having to provide an assessment of their internal controls over financial reporting.²¹ Moreover, to our knowledge, an auditor's attestation for this assessment of internal controls is still not required for smaller companies. In effect, the only impact of SOX on smaller companies has been the requirement that their executive officers certify that the financial statements are not misstated. Thus, narrowly read, the act only requires substantial additional disclosure in the MD&A report from larger companies. If firms only responded to the letter of the act, one would expect to find more post-SOX improvement in the informativeness of MD&A reports of larger companies.

To test this, we construct a dummy variable that equals one whenever a firm's market capitalization is above \$75 million and zero otherwise.²² We then regress each of the log MD&A document size (word count) and $J_{i,t}$ on the a SOX dummy, a larger firm dummy, the interaction between these two dummies, and the firm's log market capitalization. The results are presented in Table 6.

First, we find that even firms that were not required to include new information in their reports ended up with larger reports post-SOX. Second, while larger firms are unconditionally more likely to submit more informative reports, smaller firms experienced the largest post-SOX improvement in informativeness. Thus, the firms with the least onerous compliance requirements were also those that exhibited the greatest improvement in their informativeness.

In summary, the data suggest that the informativeness of MD&A reports did improve after SOX, and that much of that improvement is not directly related to the information disclosure mandated by SOX.

²¹For reference, see: <http://www.sec.gov/info/smallbus/404guide.pdf>

²²To interpret what defines a "smaller public company," consult <http://www.sec.gov/rules/interp/2007/33-8810.pdf>. This, in turn, refers to a report in <http://www.sec.gov/info/smallbus/acspc/acspc-finalreport.pdf>. From the latter, it seems that one can safely interpret \$75M in market capitalization as a cutoff (the cutoff between accelerated and non-accelerated filers—see footnote 34 in the last link).

	Ln(MD&A Size)	Informativeness (J)
SOX	0.467*** [0.011]	0.043*** [0.006]
Above \$75M	-0.079*** [0.012]	0.017*** [0.006]
SOX x Above \$75M	0.098*** [0.013]	-0.016** [0.007]
Ln(Market Cap)	0.064*** [0.004]	-0.002 [0.002]
Constant	7.844*** [0.041]	-0.007 [0.019]
Number of Obs.	24,893	24,893
R^2	0.246	0.003

Table 6: The table reports panel regressions (with firm random effects) of MDA word count and informativeness as the dependent variables. “SOX” is a dummy variable taking the value of 1 for the post-SOX years (2003 and on), “Above \$75” takes on the value of 1 if the firm had a market capitalization of more than \$75M and was therefore required to comply with all of SOX requirements.

4.2. *Did the increase in MD&A informativeness affect investors?*

The second question we pose raises the possibility that the additional information in the post-SOX MD&A reports was not new information to investors (i.e., this information was available elsewhere). For example, it is possible that the firm disclosed this information in other sections of their reports (not the MD&A) or through analyst calls. To investigate this, we ask whether improvements in our informativeness measure are related to a reduction in a proxy for asymmetric information. Such a reduction could be interpreted as evidence that the additional information in the MD&A reports was new to investors. Because share illiquidity is theoretically linked to adverse selection, we focus on the Amihud ratio, which is a measure of price impact (Amihud, 2002), and the Gibbs measure, which is a measure of the bid-ask spread (Hasbrouck, 2006).

Table 7 reports the results of panel regressions testing to see whether, everything else being equal, changes in informativeness predict changes in illiquidity. The dependent variables are the changes in the Gibbs and Amihud illiquidity measures, respectively (i.e., increases in the dependent variable are associated with deterioration in liquidity). We attempt to control for alternative

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Δ Gibbs Illiquidity Measure				Δ Amihud Illiquidity Measure			
$\Delta J_{i,t}$	-0.074*** [0.021]	-0.029 [0.023]	-0.023 [0.020]	-0.028 [0.023]	-0.858** [0.344]	-0.325** [0.152]	-0.262 [0.160]	-0.322** [0.152]
Δ Book-to-Market		0.042*** [0.006]		0.042*** [0.006]		0.317*** [0.038]		0.319*** [0.038]
Δ Market Cap			-0.117*** [0.010]	-0.106*** [0.011]			-0.803*** [0.082]	-0.606*** [0.073]
SOX	-0.132*** [0.011]	-0.105*** [0.011]	-0.117*** [0.010]	-0.106*** [0.011]	-1.504*** [0.177]	-0.603*** [0.073]	-0.803*** [0.082]	-0.606*** [0.073]
Constant	0.018** [0.008]	0.008 [0.009]	0.019** [0.008]	0.00800 [0.009]	1.188*** [0.147]	0.490*** [0.071]	0.651*** [0.074]	0.491*** [0.071]
Number of Obs.	12,149	9,190	11,531	9,190	12,149	9,190	11,531	9,190
R^2	1.40%	1.97%	1.34%	1.99%	0.71%	2.00%	1.04%	2.01%

Table 7: The table reports panel regression results (with firm random effects) of (negative) changes in the Gibbs and Amihud illiquidity measures on changes in informativeness, change in book-to-market, and change in market cap. All changes are measured year-over-year.

explanations for changes in liquidity by including changes to a firm's book-to-market ratio and changes to its market capitalization, which are known to be related to liquidity. Overall, we can reject the hypothesis that changes to informativeness are unrelated to changes in liquidity. Instead we find that improvement in the informativeness, as measured by the relative performance of the text model, are associated with improvement in liquidity. This lends support to the conjecture that the improvement in MD&A informativeness corresponds to new information revelation.

4.3. Did firms with higher agency costs increase their MD&A informativeness?

To address this question, we introduce a set of cross-sectional variables that might proxy for the presence of internal opacity and asymmetric information (see the discussion in Section 3.1):

1. **Market capitalization:** Smaller firms may be subject to higher agency costs (e.g., because fixed monitoring costs might be too high relative to the benefits monitoring brings). Alternatively, smaller firms might be more opaque.
2. **Herfindahl Index and number of reporting segments:** These variables proxy for the complexity of the firm (although they are also related to firm size). These measures could be correlated with both internal opacity and asymmetric information.

3. **Book-to-market:** Companies with a low book-to-market ratio may be associated with more intangible assets or products and thus subject to higher adverse selection costs. This prompts many studies to include book-to-market as a proxy for the presence of asymmetric information. On the other hand, firms with extremely high book-to-market might be in financial distress and, given the unusual circumstances even to insiders, subject to greater internal opacity.
4. **Governance:** SOX also stipulated on governance issues. Although many of these stipulations echoed what was already required from listed companies by the NYSE and NASDAQ (e.g., a majority of independent directors, financial literacy on the part of at least one audit committee member), SOX did add to this marginally by requiring that audit committee members of listed companies be ‘independent’ and control the firm-auditor relationship.²³ Thus, an improvement in governance could conceivably be linked to a decrease in asymmetric information. We use the G-Index of corporate governance (Gompers, Ishi, and Metrick, 2003) to proxy for the overall strength of corporate governance.
5. **Financial and operating risk:** Firms with more risky operations might experience greater internal opacity. We proxy for operational risk with financial leverage (debt-to-asset ratio) and operating leverage (year-over-year change in EBIT divided by year-over-year change in sales). On the other hand, these variables could also be correlated with size.
6. **Analyst coverage:** One might expect less information to be available for firms with lower analyst coverage, thus associating such firms with greater asymmetric information. This measure could be correlated with size.
7. **Analyst dispersion:** Likewise, greater disagreement between experts concerning a firm’s prospects may signal greater potential for asymmetric information.
8. **Idiosyncratic volatility:** One might expect that private information about a firm would be of an idiosyncratic nature. Firms whose volatility is largely driven by idiosyncratic noise

²³SOX introduced a ban on loans to executive officers, but this does not preclude close substitute such as an increase in short-term compensation coupled with a decrease in long-term compensation.

might therefore afford more scope for the presence of asymmetric information. On the other hand, greater idiosyncratic volatility may create more scope for internal opacity.

We emphasize that the relationship between these variables, adverse selection costs, and internal opacity is not always clear. What guides us to select these variables is more a suspicion that they may be related to internal opacity and/or asymmetric information. By including them in a cross-sectional regression we hope to learn more about which firms might have been more likely to experience an improvement in their post-SOX informativeness.

The summary statistics for the cross-sectional variables and the liquidity measures we use are reported in Table 8. The table suggests that compared with the pre-SOX period, firms in the post-SOX period are larger, have lower book-to-market, are somewhat less leveraged, have wider analyst coverage, exhibit less idiosyncratic volatility (though not in relation to total volatility which also declines), and are more liquid.

One measure of asymmetric information that is theoretically important and is readily observable is share illiquidity. Rather than guess at the relationships between asymmetric information and the cross-sectional variables listed above, it might be useful to let the data suggest a relationship. To this end, we investigate how the cross sectional variables are related to illiquidity, before and after the passage of SOX, by regressing the illiquidity measures on a dummy indicating whether each of these measures fall into the top or bottom quintile of a given cross-sectional variable, a SOX dummy, and the interaction of the two.²⁴ Table 9 reports the results. Regardless of the measure used, it appears that illiquidity declined in the post-SOX period, because the SOX dummy is negative and significant in all the regressions. In addition, it appears that save for the Herfindahl Index (second column in Table 9), all of the variables have some significant relationship with one or both share illiquidity measures, and this relationship is the same across the two illiquidity measures (whenever significant). It is also clear from the table that not all variables are related to the illiquidity measures in the manner suggested by our earlier discussion. Specifically, the book-to-market ratio and analyst forecast dispersion have the opposite signs relative to what we expected. Finally, it appears that, to the extent that the relationship between illiquidity and each of the cross-sectional

²⁴Each year, we assign firms into quintiles based on that year's distribution. Both illiquidity measures assign higher values to less liquid firms.

	Market Cap in (\$1,000)	H Index	Number of Segments	Book-to Market	G Index	Debt-to Assets	Operating Leverage	Analyst Coverage	Analyst Forecast Disp	Idiosync. Volatility	Gibbs Illiquidity	Amihud Illiquidity
Mean	1,399,812	0.361	2.28	0.87116	8.549	0.194	3.12	3.97	0.25	0.042	1.1285	2.843
Standard Dev.	8,682,356	0.350	0.55	1.41822	2.654	0.233	92.48	5.34	1.31	0.024	1.0587	16.103
25% percentile	53,554	0.030	2.00	0.29213	7.000	0.003	0.72	0.00	0.02	0.026	0.4796	0.012
50% percentile	168,478	0.237	2.00	0.56732	8.000	0.109	0.98	2.00	0.05	0.038	0.7808	0.119
75% percentile	611,490	0.670	2.00	1.03622	10.000	0.320	1.21	6.00	0.17	0.054	1.3730	1.069
Pre-SOX												
Mean	1,951,613	0.369	2.34	0.75522	8.939	0.173	2.67	4.97	0.21	0.029	0.5133	1.074
Standard Dev.	8,100,837	0.357	0.62	0.84886	2.498	0.221	56.95	6.02	1.14	0.018	0.4736	9.134
25% percentile	104,630	0.026	2.00	0.33121	7.000	0.001	0.81	0.00	0.02	0.017	0.2439	0.002
50% percentile	334,266	0.248	2.00	0.56710	9.000	0.094	1.00	3.00	0.05	0.024	0.3728	0.019
75% percentile	1,189,965	0.718	3.00	0.91570	11.000	0.271	1.23	7.00	0.14	0.036	0.5993	0.212
Post-SOX												

Table 8: Summary statistics of the different cross-sectional proxies for information asymmetry in the pre-SOX period (1998-2002) and the post-SOX period (2003-2006).

characteristics changed after SOX, the relationship weakens (the sign of the “High” coefficient is generally opposite to the sign “SOX x High” coefficient). Specifically, for each of size, book-to-market, governance, financial leverage, analyst coverage and analyst dispersion, the difference in liquidity between the high and low quintile shrinks after SOX.

Next, we ask what firm characteristics are associated with greater disclosure. As posited earlier, firms with greater asymmetric information might choose to disclose more. This could be because they learned more through the implementation of SOX and investors, anticipating this learning process, expect greater informativeness. Alternatively, managers might be more concerned about the greater liability introduced by SOX and therefore seek to reduce asymmetric information. We create a dummy variable, *High*, which takes the value of 1 if a firm is in the top quintile of the proxy (in a given year), the value of 0 if the firm is in the bottom quintile of the proxy, and treated as ‘missing’ otherwise. Then, to test whether the post-SOX improvement in forecasting power of MD&A reports is related to the proxy, we run the following panel regressions with firm random effects:

$$J_{i,t} = \alpha + \beta_1 SOX_t + \beta_2 H_{i,t} + \beta_3 SOX_t H_{i,t} + \epsilon_{i,t}, \quad (6)$$

where $H_{i,t}$ is the cross-sectional variable being investigated. Panel A in Table 10 reports the results from this regression and provides an indication of how our measure of MD&A informativeness varies with the proxies.²⁵ The first row in Panel B reports the average informativeness across firms in post-SOX years (2003-2006). The coefficient is positive, consistent with our earlier finding that the text-based model forecasts are better than the past year’s volatility in predicting future volatility after SOX. The second and third rows in Panel B report the average informativeness in the top and bottom quintiles, respectively, of the associated cross-sectional variable. The fourth and fifth rows test to see whether the informativeness of the top and bottom quintiles are significantly different from the unconditional informativeness. Thus Panel B is another way to test for a post-SOX relationship between the cross-sectional variables and MD&A informativeness and can be viewed as a non-parametric “test” of the robustness of Panel A.

²⁵Given the time trends in some of these variables (e.g., market capitalization) we sort observations into quintiles separately for each year.

	Market Cap in (\$1,000)	H Index	Number of Segments	Book-to Market	G-Index	Debt-to Assets	Operating Leverage	Analyst Coverage	Analyst Forecast Disp
Panel A: Gibbs Illiquidity Measure									
SOX	-1.102*** [0.023]	-0.483*** [0.049]	-0.472*** [0.024]	-0.433*** [0.024]	-0.271*** [0.013]	-0.543*** [0.021]	-0.609*** [0.029]	-0.837*** [0.023]	-0.281*** [0.016]
High	-1.552*** [0.030]	0.091 [0.073]	-0.254*** [0.100]	0.342*** [0.030]	-0.176*** [0.025]	-0.174*** [0.032]	-0.055* [0.029]	-1.125*** [0.033]	0.349*** [0.018]
SOX $\times$ High	0.835*** [0.032]	0.035 [0.069]	0.12 [0.100]	-0.208*** [0.034]	0.085*** [0.022]	0.082** [0.033]	0.01 [0.040]	0.545*** [0.036]	-0.193*** [0.023]
Constant	2.087*** [0.021]	0.980*** [0.055]	0.994*** [0.022]	0.994*** [0.022]	0.642*** [0.014]	1.222*** [0.022]	1.279*** [0.025]	1.748*** [0.021]	0.650*** [0.014]
Number of Obs. R^2	7,209 48.60%	1,563 10.20%	3,183 11.00%	5,802 16.00%	2,639 18.10%	7,170 13.20%	4,709 14.20%	8,392 29.50%	4,630 26.40%
Panel B: Amihud Illiquidity Measure									
SOX	-4.564*** [0.370]	-2.814*** [0.969]	-1.417*** [0.505]	-0.330** [0.167]	-0.003 [0.044]	-1.542*** [0.391]	-1.745*** [0.649]	-4.068*** [0.535]	-0.119* [0.062]
High	-7.461*** [0.393]	0.476 [1.537]	-1.43 [2.052]	1.688*** [0.233]	-0.096 [0.094]	-1.657** [0.654]	-0.01 [0.675]	-7.933*** [0.637]	0.399*** [0.076]
SOX $\times$ High	4.551*** [0.522]	1.626 [1.381]	0.87 [2.063]	-0.821*** [0.237]	-0.024 [0.074]	0.907 [0.608]	-0.62 [0.909]	4.013*** [0.820]	-0.270*** [0.092]
Constant	7.482*** [0.272]	3.279*** [1.157]	1.290*** [0.192]	1.290*** [0.192]	0.158*** [0.054]	3.938*** [0.481]	4.006*** [0.528]	8.076*** [0.409]	0.323*** [0.062]
Number of Obs. R^2	7,208 7.24%	1,562 0.39%	3,182 0.50%	5,803 2.60%	2,637 0.23%	7,169 0.36%	4,704 0.60%	8,487 3.32%	4,621 2.15%

Table 9: We presents panel regression results (with firm random effects) where the dependent variables are measures of illiquidity (the Gibbs measure (multiplied by 100) in Panel A and the Amihud measure in Panel B). The independent variables are the SOX dummy, a dummy for a cross sectional variable's top-quintile, and the interaction between the two. The cross-sectional dummy takes on the value of 1 if the observation is in the top quintile, 0 if the observation is in the bottom quintile, and missing value otherwise. We estimate the model for one cross-sectional measure at a time.

	Market Cap in (\$1,000)	H Index	Number of Segments	Book-to Market	G-Index	Debt-to Assets	Operating Leverage	Analyst Coverage	Analyst Forecast Disp	Gibbs Illiquidity	Amihud Illiquidity
						Panel A					
SOX	0.045*** [0.007]	0.019 [0.015]	0.027*** [0.007]	0.017** [0.007]	0.028*** [0.009]	0.039*** [0.006]	0.023*** [0.007]	0.048*** [0.006]	0.01 [0.007]	0.021*** [0.007]	0.027*** [0.007]
High	0 [0.009]	0.004 [0.017]	-0.00 [0.026]	-0.024*** [0.009]	-0.005 [0.019]	0.015* [0.008]	0.00 [0.008]	0.020** [0.008]	-0.003 [0.007]	0.031*** [0.007]	0.01 [0.007]
SOX $\times$ High	-0.024** [0.010]	-0.001 [0.021]	-0.00 [0.029]	0.025** [0.010]	0.002 [0.016]	-0.010 [0.009]	-0.01 [0.010]	-0.031*** [0.010]	0.019** [0.010]	0.011 [0.010]	-0.003 [0.010]
Constant	-0.025*** [0.006]	-0.023** [0.012]	-0 [0.006]	-0 [0.006]	-0.022** [0.011]	-0.037*** [0.006]	-0.012* [0.006]	-0.037*** [0.005]	-0.010* [0.005]	-0.045*** [0.005]	-0.024*** [0.005]
Number of Obs.	9,954	1,932	3,863	7,210	3,497	10,018	6,171	11,770	6,073	7,567	7,567
R^2	0.32%	0.13%	0.23%	0.37%	0.08%	0.45%	0.26%	0.40%	0.45%	1.12%	0.30%
						Panel B					
Unconditional	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]	0.024*** [0.003]
Top quintile	0.013** [0.006]	0.015 [0.014]	-0.00 [0.026]	0.038*** [0.008]	0.013 [0.015]	0.028*** [0.007]	0.019** [0.008]	0.014** [0.006]	0.032*** [0.008]	0.032*** [0.007]	0.020*** [0.008]
Bottom quintile	0.037*** [0.008]	0.014 [0.016]	0.021*** [0.007]	0.00500 [0.007]	0.012 [0.011]	0.031*** [0.006]	0.024*** [0.007]	0.040*** [0.007]	0.008 [0.006]	0.017** [0.007]	0.024*** [0.005]
Top quintile = Unconditional	0.070	0.567	0.430	0.054	0.428	0.533	0.497	0.076	0.298	0.2460	0.572
Bottom quintile = Unconditional	0.067	0.351	0.679	0.005	0.282	0.237	0.947	0.001	0.003	0.2615	0.9918

Table 10: In panel A, we presents panel regression results (with firm random effects) where the dependent variable is our informativeness measure. The independent variables are the SOX dummy, a dummy for the cross sectional dummy, and the interaction between the two. The cross-sectional dummy takes on the value of 1 if the observation is in the top quintile, 0 if the observation is in the bottom quintile, and missing value otherwise. We estimate the model for one cross-sectional measure at a time. Panel B reports the average informativeness in the post-SOX period unconditionally and in the top and bottom quintile of the cross-sectional variable. The last two rows report the p values associated with testing the difference between the unconditional average informativeness and average in each of the quintiles.

In Panel A, we find that firms with low book-to-market ratios, high financial leverage, and high share illiquidity were unconditionally more likely to have informative MD&A reports. On the other hand, we find that firms with small market value, high book-to-market ratio, low analyst coverage, and high analyst forecast dispersion experienced the greatest increase in informativeness post-SOX. What is striking about these results is that in every case, the significant coefficient has the opposite sign as the analogous coefficient in Table 9. In other words, a significant unconditional predictor of informativeness is also a significant unconditional predictor of share liquidity. Likewise, a significant predictor of post-SOX improvement in informativeness is a significant unconditional predictor of improvement in liquidity. This is consistent with our findings from Table 7, and lends further support to the notion that the degree of asymmetric information is related unconditionally to the degree of informativeness.

As described earlier, the level of a firm's idiosyncratic volatility can proxy for both the potential of asymmetric information and internal opacity. To see whether idiosyncratic volatility and informativeness are related, we regress our measure of informativeness on total volatility, idiosyncratic volatility, and systematic volatility while separating the sample into pre- and post-SOX years. We compute systematic volatility with respect to the CRSP Value Weighted Index using an annual correlation estimate that is robust to infrequent stock trading (see Dimson (1979)). Table 11 shows the results for leading (columns 1-4) and contemporaneous (columns 5-8) dependent variables. Pre-SOX, there is no relation between past volatility (either total, systematic, or idiosyncratic) and subsequent informativeness. However, post-SOX we see that past total volatility is positively related to subsequent model improvement. Moreover, this relationship appears to reside entirely with the idiosyncratic component of volatility, consistent with the hypothesis. Similar results are obtained when we examine the relation between model improvement and contemporaneous volatility.

5. Conclusions

By now, a vast literature exists about the effects on firms of the Sarbanes-Oxley Act (SOX) of 2002. It is a fact that mandatory disclosure documents have ballooned in size since the passage of SOX. What is not clear is whether this increase in disclosure document size has led to more transparency

	J_{t+1}				J_t			
	Pre-SOX	Post-SOX	Pre-SOX	Post-SOX	Pre-SOX	Post-SOX	Pre-SOX	Post-SOX
Total Volatility	0.004 [0.03]	0.690 [3.48]**			0.445 [1.92]	3.623 [9.02]**		
Systematic Volatility			-0.038 [0.13]	0.319 [1.30]			-0.224 [1.11]	0.129 [0.30]
Idiosyncratic Volatility			0.049 [0.30]	0.593 [2.48]*			0.563 [2.20]*	3.731 [7.79]**
Constant	-0.009 [1.15]	-0.015 [2.17]*	-0.011 [1.54]	-0.013 [1.97]*	-0.045 [4.00]**	-0.111 [9.11]**	-0.045 [4.11]**	-0.107 [9.53]**
Observations	6433.000	11101.000	6312.000	10928.000	12142.000	13610.000	11887.000	13411.000
R^2	0.001	0.012	0.001	0.011	0.002	0.062	0.002	0.061

Table 11: The dependent variable is informativeness and independent variables include total firm Volatility, systematic volatility, and idiosyncratic volatility. Panel regressions with firm random effects.

(or alternatively, has been associated with more information being disclosed). We provide evidence that annual reports have become more informative in the post-SOX era. Our evidence suggests that this is not related to the degree or type of compliance with SOX. The information disclosed does appear to be new in the sense that it is associated with a reduction in illiquidity. Moreover, examining how the increased informativeness is related to firm-specific characteristics suggests that the increased informativeness is driven by the degree of information asymmetries—firms characterized by more asymmetric information appear to disclose more in the post-SOX period than they did in the pre-SOX period.

There might be several reasons why adverse selection costs might induce firms to divulge more information in the post-SOX period. Firms might have learned more through the implementation of SOX provisions, or because of the increased liability to senior executives and auditing firms, might be paying more attention to financial and operational aspects of their company. Those firms that learn more would disclose more, lest investors believe that the absence of an increase in disclosure signals something bad that was learned about the firm. This rationale presupposed that SOX somehow induced firms to learn more.

An alternative hypothesis is that managers became frightened by the increased liability, and those in companies with higher adverse selection costs had greater reason to be frightened. Such firms might not have learned anything about their operations or financial situation through SOX

but the fear might have induced disclosure that spilled over beyond the narrow transparency requirements of SOX.

While we are unable to clearly separate the causes for disclosure (because our proxies may be correlated with both internal opacity and adverse selection costs), there appears to be some suggestive evidence that “fear”, and not just “learning” played an important role in the increased informativeness post-SOX. Specifically, the cohort of firms exhibiting the largest increase in informativeness is the group to which SOX applied minimally (firms with market capital under \$75M). Moreover, post-SOX improvement in informativeness of MD&A reports appears to be related to post-SOX improvement in share liquidity.

References

- Amihud, Y., 2002, “Illiquidity and stock returns: cross-section and time-series effects,” *Journal of Financial Markets*, 5(1), 31–56.
- Antweiler, W., and M. Z. Frank, 2004, “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards,” *Journal of Finance*, 59, 1259–1294.
- Bhattacharya, U., P. Groznik, and B. Haslem, 2007, “Is CEO certification of earnings numbers value-relevant?,” *Journal of Empirical Finance*, 14(5), 611–635.
- Bollerslev, T., 1986, “Generalized Autoregressive Conditional Heteroskedasticity,” *Journal of Econometrics*, 31, 307–327.
- Das, S., and M. Chen, 2001, “Yahoo for Amazon: Extracting Market Sentiment from Stock Message Boards,” in *Proc. of Asia Pacific Finance Association Annual Conference*.
- Dimson, E., 1979, “Risk Measurement when Shares are Subject to Infrequent Trading,” *Journal of Financial Economics*, 7(2), 197–226.
- Doidge, C., G. A. Karolyi, and R. M. Stulz, 2008, “Why Do Foreign Firms Leave U.S. Equity Markets? An Analysis of Deregistrations Under SEC Exchange Act Rule 12h-6,” NBER Working Papers 14245, National Bureau of Economic Research, Inc.
- Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, 1997, “Support Vector Regression Machines,” in *Advances in NIPS 9*.
- Engel, E., R. M. Hayes, and X. Wang, 2007, “The Sarbanes-Oxley Act and firms’ going-private decisions,” *Journal of Accounting and Economics*, 44(1-2), 116–145.
- Engelberg, J., 2007, “Costly Information Processing: Evidence from Earnings Announcements,” .
- Engle, R. F., 1982, “Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation,” *Econometrica*, 50, 987–1008.

- Gaa, C., 2007, “Media Coverage, Investor Inattention, and the Market’s Reaction to News,” .
- Ghose, A., P. Ipeirotis, and A. Sundararajan, 2007, “Opinion Mining using Econometrics: A Case Study on Reputation Systems,” in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 416–423, Prague, Czech Republic. Association for Computational Linguistics.
- Goldman, E., and S. L. Slezak, 2006, “An equilibrium model of incentive contracts in the presence of information manipulation,” *Journal of Financial Economics*, 80(3), 603–626.
- Gompers, P., J. Ishi, and A. Metrick, 2003, “Corporate Governance and Equity Prices,” *Quarterly Journal of Economics*, 118(1), 107–155.
- Hammersley, J. S., L. A. Myers, and C. Shakespeare, 2008, “Market reactions to the disclosure of internal control weaknesses and to the characteristics of those weaknesses under section 302 of the Sarbanes Oxley Act of 2002,” *Review of Accounting Studies*, 13, 141–165.
- Hasbrouck, J., 2006, “Trading Costs and Returns for US Equities: Estimating Effective Costs from Daily Data,” .
- Joachims, T., 1999, “Making Large-Scale SVM Learning Practical,” in *Advances in Kernel Methods - Support Vector Learning*. MIT Press.
- Kogan, S., D. Levin, B. R. Routledge, J. S. Sagi, and N. A. Smith, 2009, “Predicting Risk from Financial Reports with Regression,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 272–280, Boulder, Colorado. Association for Computational Linguistics.
- Koppel, M., and I. Shtrimberg, 2004, “Good news or bad news? Let the market decide,” in *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.
- Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, 2000a, “Language Models for Financial News Recommendation,” in *Proc. of CIKM*.

- , 2000b, “Mining of concurrent text and time series,” in *Proc. of KDD*.
- Lerman, K., A. Gilder, M. Dredze, and F. Pereira, 2008, “Reading the markets: Forecasting public opinion of political candidates by news analysis,” in *COLING*.
- Leuz, C., 2007, “Was the Sarbanes-Oxley Act of 2002 really this costly? A discussion of evidence from event returns and going-private decisions,” *Journal of Accounting and Economics*, 44(1-2), 146–165.
- Li, F., 2005, “Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?,” .
- Ogneva, M., K. Raghunandan, and K. Subramanyam, 2007, “Internal Control Weakness and Implied Cost of Equity: Evidence from SOX Section 404 Disclosures,” *The Accounting Review*, 82(5), 1155–1197.
- Pang, B., L. Lee, and S. Vaithyanathan, 2002, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proc. of EMNLP*.
- Piotroski, J. D., and S. Srinivasan, 2008, “Regulation and Bonding: The Sarbanes-Oxley Act and the Flow of International Listings,” *Journal of Accounting Research*, 46(2), 383–425.
- Tetlock, P. C., 2007, “Giving Content to Investor Sentiment: The Role of Media in the Stock Market,” *Journal of Finance*, 62(3), 1139–1168.
- Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, 2008, “More Than Words: Quantifying Language to Measure Firms’ Fundamentals,” *Journal of Finance*, 63(3), 1437–1467.
- Wagner, S., and L. Dittmar, 2006, “The Unexpected Benefits of Sarbanes-Oxley,” *Harvard Business Review*, pp. 133–140.
- Weiss-Hanley, K., and G. Hoberg, 2008, “Strategic Disclosure and the Pricing of Initial Public Offerings,” .

Wiebe, J., and E. Riloff, 2005, “Creating Subjective and Objective Sentence Classifiers from Unannotated Texts,” in *CICLing*.

Zhang, I. X., 2007, “Economic consequences of the Sarbanes-Oxley Act of 2002,” *Journal of Accounting and Economics*, 44(1-2), 74–115.