Predicting fraud from quarterly conference calls: A small-sample study of scripted language

Research-in-Progress

Lee Spitzley
University of Arizona
lspitzley@email.arizona.edu

Abstract

This study introduces a method of detecting deceptive and fraudulent executives by using the information contained in quarterly conference calls and financial statements. I argue that executives in fraudulent companies will adhere more closely to a script as way to minimize the risk of disclosing negative or inconsistent information, and that the MD&A section of the corresponding 10-K or 10-Q statement is a valid proxy for this script. I use a small sample of companies from the financial services industry during the financial crisis of 2007-2008. Using tf-idf term weighting and cosine similarity, I compare the executives’ language in the conference calls to the corresponding MD&A statement. The results indicate that executives in fraudulent companies use quarterly conference call language that is more similar to the corresponding MD&A statement than those in non-fraudulent companies. This is a research-in-progress, so I discuss the next steps for this project.

Introduction

Corporate financial fraud is a serious infraction that damages the firms involved and investor trust in markets, and requires significant resources to identify. Modern research has used both financial and behavioral indicators to aid auditors and investors in the identification of potential fraudulent financial reporting. This work extends the literature in this area by drawing upon theory from preparation and deception.

In many publicly traded companies, a small panel of executives will take part in a quarterly conference call to discuss their financial results. These calls normally follow the same format. First, the executives will give a prepared statement about their earnings performance. When this is completed, a panel of analysts who follow the company ask the executives questions.

These calls provide an interesting environment to study the role of rehearsal and deception in a field setting. A firm’s legal team usually prepares the language of the presentation portion and some of the expected questions, likely to minimize the risk of negative perception by the market (Lee 2014). These calls follow the release of the company’s quarterly earnings statement, the 10-K (annual) or 10-Q (quarterly). In these statements, the company presents their financials and qualitatively discusses the firm’s operations and outlook in a section called the “Management Discussion and Analysis”.

Prior work in this area of fraud detection has explored the language of financial statements (Cecchini et al. 2010a; Goel and Gangolly 2012; Goel et al. 2010; Humpherys et al. 2011) and the verbal (Larcker and Zakolyukina 2012) and non-verbal (Hobson et al. 2012; Mayew and Venkatachalam 2012) behavior of executives in the conference calls. This paper is interested in integrating these two data sources to uncover instances of financial fraud. My research questions are:

- Do executives from fraudulent companies adhere more closely to a script than executives from non-fraudulent companies?
- Do executives in fraudulent companies give less relevant answers to analyst questions during conference calls?
To address these questions, I examine 24 conference calls from six financial services companies that were active during the financial crisis of 2007-2008. Using concepts from information retrieval, I compare these conference calls to their corresponding 10-K and 10-Q statements. I hypothesize that executives in fraudulent companies will use language that is more similar to their financial statements as a strategy to reduce the amount of information they disclose and reduce their chances of disclosing fraudulent activities. To my knowledge, this is the first study to use this method to detect financial fraud.

This also provides an interesting test bed for research on deceptive communication, because it includes instances of high-stakes deception in field setting, prepared and spontaneous communication, and a good estimator of the script the deceptive parties used to fool their audience. I discuss the relevant literature, the methods used to obtain and analyze the data, my findings, and proposed future research. Because this is still a research-in-process paper, not all hypotheses were tested.

**Literature review**

Publicly traded companies must file annual financial statements (10-K) with the SEC, and many companies file additional quarterly (10-Q) financial statements that track the interim performance of the company. Prior research has used a variety of methods to attempt to predict fraud from these statements. Many studies focus on anomalies in financial ratios as indicators of fraud (Abbasi et al. 2012; Beneish 1999; Cecchini et al. 2010b; Dechow et al. 2011). While financial indicators remain a popular and powerful way to detect fraud, new research has suggested that linguistic analysis of the management discussion and analysis (MD&A) section of 10-K or 10-Q financial statements contains meaningful information to market participants. This type of analysis also has the ability to predict fraud (Cecchini et al. 2010a; Glancy and Yadav 2011; Goel and Gangolly 2012; Goel et al. 2010; Humpherys et al. 2011). There is also meaningful information in news coverage about firms (Tetlock et al. 2008). Researchers have also started to analyze information contained in conference calls that discuss the financial statements.

**Conference calls and executive disclosure**

To supplement the release of a financial statement, many publicly traded companies also hold a conference call with analysts who follow the companies. These calls generally contain information above what is contained in the accompanying press release of the company (Matsumoto et al. 2011). Additional analyses of these calls includes voice patterns, such as cognitive dissonance (Hobson et al. 2012; Mayew and Venkatachalam 2012).

**Theory and hypotheses**

*Adherence to a script*

The first part of these calls typically follows a script. The question and answer portion of the calls is more difficult to prepare for, so I expect this to be less similar to the MD&A than the prepared portion of the call. This hypothesis is straightforward, but it is important to consider the differences between both parts of the calls.

\[ H1: \text{The formal presentation portion of conference calls will be more similar to the MD&A section of the corresponding financial statement than the Q&A portion of the call.} \]

*The role of rehearsal*

Executives rehearse conference calls extensively before the call takes places as a way to manage disclosure and avoid any lapses that observers may perceive negatively. Rehearsal is expected to make it more difficult for human judges to detect deceptive statements (deTurck and Miller 1990; Littlepage and Pineault 1985). One reason that investors follow conference calls is the disclosure of new information above press releases and financial statements (Matsumoto et al. 2011); however, Lee (2014) found that when scripting was used, firms disclosed less information. Because the deceptive executives are attempting to convince the audience that what they say is true, I believe they will stick closer to a script to maintain consistency and to avoid inadvertently disclosing damaging information.
**Responses to questions**

Deceivers who have time to prepare before delivering a deceptive message tend to give shorter messages (Sporer and Schwandt 2006). As part of a risk-mitigation strategy, I expect executives from fraudulent companies to give shorter answers to analysts' questions.

**H3:** Executives in fraudulent companies will give shorter responses to analyst questions than executives in non-fraudulent companies.

I also believe that executives will give less-relevant answers to analyst questions to avoid disclosing additional information that the market may treat negatively, or that regulators may notice. When deceivers have time to prepare, they tend to give answers that are similar to those of truth-tellers, but they provide less detail for unexpected questions (Shaw et al. 2013). I expect that fraudulent executives will give less relevant answers to analysts' question in general.

**H4:** Executives in fraudulent companies will give less-relevant answers to unexpected analyst questions than executives in non-fraudulent companies.

**Method**

This section describes the process of identifying my sample, collecting data collection and preparation, and how I analyzed the data.

**Sample selection**

For this study, I chose six companies in the financial services industry that were active during the financial crisis of 2007-2008. I used three companies with no evidence of fraud as the control group, while the Securities and Exchange Commission charged the other three businesses with fraud. Evaluating all of the earnings call transcripts from these companies will help me identify any differences between fraudulent and non-fraudulent companies.

Two main constraints bounded this sample. First, the historical coverage of conference calls in the Seeking Alpha (SA) transcript database only dates back to about 2006 or 2007, depending on the company. This limited my potential sample by excluding companies that committed fraud in periods that were not included in this dataset.

Second, the latency between fraud being committed and formal charges, documented in an Accounting and Auditing Enforcement Release (AAER) or shareholder lawsuits, is typically several years. Because of this, there could potentially be companies in the sample demographic that have committed fraud and an investigation has not completed or even started.

**Description of the companies**

The first fraudulent company in the sample is Countrywide Financial. The SEC charged executives in this company of misleading investors about their credit risk, and the CEO of insider trading (SEC 2009). The second company is Citigroup. Bondholders charged Citigroup with misleading investors from May 2006 to November 2008. In 2013, Citigroup agreed to pay $730 million to settle the lawsuit and the executives from that time period settled charges of withholding information (Alloway 2013). Executives from Bank of America also misled shareholders during this period. The company reached a $150 million settlement with the SEC, and executives have settled lawsuits totaling more than $60 million for failing to disclose negative information surrounding acquisitions during this period (Sterngold et al. 2014).

The use of these companies, in which the executives were directly implicated in misleading investors, helps to address concerns of Bloomfield (2012), who dutifully notes that accounting misstatements may not be the result of deception on the part of the executives.
I include three banks not implicated in any wrongdoing as a control group. Fifth Third Bank avoided lawsuits and remained solvent through the 2007-2008 banking crisis, and they still have the same CEO as they did in 2007-2008. I also use US Bancorp, which also avoided litigation and has retained their CEO from the crisis era. The third control company, PNC Financial, also avoided litigation and retained their CEO until his retirement in 2013. While it may be possible that these companies had committed fraud during this period, there is no evidence supporting this and therefore it is reasonable to assume these companies were legitimate during this period.

Data collection

After identifying the companies for my sample, I built a tool that could extract important metadata from each SA transcript and then separate each speaker-instance from the transcript. Conveniently, SA identified the speakers as executives or analysts prior to the beginning of the actual call (see Figure 1) which allowed me to extract each individual’s name, organization, and position, if an executive.

![Figure 1: Structure of SA meta-data](image)

An HTML “p” tag separated each speaker-instance (see Figure 2). For each paragraph, I compared the text to the list of speakers in the transcript. If it matched, I attributed the subsequent paragraphs to that individual, until reaching a paragraph with a new speaker. This required some minor fault tolerance to be included because of minor formatting variations between transcripts, but a manual check revealed that the program was able to handle this and produce accurate output for each transcript.

Once the program had all of the speaker-instances for a call, it output text files of each instance. To enable sorting of the transcripts, the filenames contained a unique ID number, information about the company and the period of the call, who was speaking, what part of the call the utterance occurred (prepared or Q&A), and whether the speaker was an executive or an analyst. Lastly, I separated the executive excerpts from the analysts and between fraud and non-fraud.

The financial statements came from the SEC EDGAR database. This database is publicly available and contains all reports required of publicly traded companies. I manually located the statements from each company that corresponded to the conference call data, downloaded them, and manually removed the tables from the documents.
Next, I discuss the processes of preparing data for classification.

**Linguistic analysis techniques**

Researchers have used various linguistic techniques in recent years to analyze qualitative market information. Dictionary-based approaches can extract tone and affective language from annual and quarterly statements (Li 2008). Research has also been able to use dictionaries to discriminate between deceptive and non-deceptive conference calls (Larcker and Zakolyukina 2012).

Prior research has also used techniques from the information retrieval domain. Most common is the term frequency-inverse document frequency (tf-idf) weighting algorithm and cosine similarity to measure the similarity between to documents (Manning et al. 2008 pp. 117–125).

Term frequency represents the number of times a term appears in a given document. This is represented as $tf_{t,d}$, which is a $t \times d$ matrix containing the number of times each term appears for each document. The idf formula is

$$idf_t = \log \frac{N}{df_t},$$

where $idf_t$ is the weighting of term $t$, $N$ is the number of documents in the corpus, and $df_t$ is the number of documents containing term $t$.

The multiplication of $tf \times idf$ yields the weighted matrix for the documents in the corpus. Each row vector in this matrix represents a document, and these vectors can be compared to each other using cosine similarity. Because these vectors only contain non-negative values, the value of the cosine similarity will be between 0 and 1, where 0 represents two completely orthogonal vectors (completely dissimilar) or 1, which represents the same vector (completely the same). One useful characteristic of this approach is that the length of the documents is not a factor. The formula for cosine similarity is

$$\cos \theta = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|},$$

where $\theta$ is the angle between vectors $v_1$ and $v_2$, the numerator is the dot product of the two vectors, and the denominator is the product of the vector lengths (i.e. $\|v_1\|$ represents the length of $v_1$).

Financial literature has used tf-idf and cosine similarity measure to compare MD&A sections for modifications over time (Brown and Tucker 2011). One of the primary benefits of tf-idf is that it gives more weight to rare terms in a corpus and greatly diminishes the importance of words like “the” or “and”. There are words that may have connotations in an everyday setting, such as “debt” or “gain”, but mean little in a financial setting, where these terms are frequent (Loughran and McDonald 2011); tf-idf can mitigate this issue.
Data preparation

The text needed to be converted to a quantitative format, so I used scikit-learn\(^1\) to compute the idf values, using the MD&A sections as the corpus of documents. To compute the tf-idf vectors of the conference calls, I first loaded the executives’ utterances (excluding those of the analysts) of each call and concatenated them into two strings, one for the prepared statements and the other for the Q&A portion of the call. I used the TfidfVectorizer function provided by scikit-learn for each of these strings to compute the tf-idf weightings for each call segment. I then created the vector for the MD&A statement.

Measuring similarity between the calls and financial statements

I used cosine similarity to measure the similarity of the documents. The scikit-learn package includes a function to compute cosine similarity. I used this to compare the tf-idf vectors of the prepared remarks and Q&A of each call to the corresponding MD&A section. This function returns a value between 0 and 1. For each combination, I output the similarity value and prepared it for analysis with R.

Results

To test H1 and H2, I used a mixed-design ANOVA, with prepared and Q&A as within subjects effects and fraud as between subjects effects. The dependent variable in these tests is the similarity score.

Testing the main effect of prepared vs Q&A (H1), which states that the prepared portion of the conference calls will be more similar to the MD&A than the Q&A is supported at \( p < 0.001 \). The mean similarity of the prepared portion is 0.8119, and the mean similarity of the Q&A portion is 0.6346. This supports H1.

The between subjects test, which tests for a main effect between fraud and non-fraud, is significant at \( p < 0.01 \). The mean similarity scores of the prepared portions are 0.7809 for non-fraud, and 0.8430 for fraud; the mean similarity scores of the Q&A portions are 0.5914 for non-fraud and 0.6779 for fraud. This supports H2, which asserts that fraudulent executives will stay more closely to their script. See Figure 3 for an interaction plot of these results.

\[ \text{Interaction Plot} \]

\[ \text{Mean Similarity Score} \]

\[ \text{Call Portion} \]

Figure 3: Visualization of ANOVA results

---

Proposed study of question responses

This is still a research in progress, so I have not yet tested H3 and H4. To test these hypotheses, I plan to continue with the information retrieval approach. Since the responses will be constrained to a specific domain, a more simple approach—such as word matching or weighted word matching—may be appropriate (Kolomiyets and Moens 2011). Another potential measure would be the number of times analysts need to repeat their question or request clarification, which would look for analyst utterances such as “can you quantify that?” or “let me just ask it a different way…,” (these were follow-up questions asked in the CFC calls). It will also involve removing phatic communication (i.e. “thank you” or “good afternoon”) from the sample.

Post-hoc analyses

I will use several classification algorithms in Weka (e.g. logistic regression and SVM) to test the ability of various measures to classify truthful vs. deceptive statements. Another interesting test I plan to conduct follows on the work of Brown and Tucker (2011), which finds that usefulness of MD&A statements decreased over time. Since these calls appear to have a lot in common with the MD&A, it seems reasonable that conference calls may also be decreasing in their usefulness as well.

Discussion

In this small sample study, I found that the executives of fraudulent banking corporations used language that was more similar to their corresponding financial statement than their peers in non-fraudulent companies. There were main effects for prepared vs. Q&A and between fraudulent and non-fraudulent companies. This supports my hypotheses; however, future research should explore the strategic factors behind the use of more similar language. The measures here are only correlations, so I cannot infer causality from this analysis.

Limitations and future research

There are several important limitations to this study. I assume that the executives in the presentation are aware of the financial statement manipulations. There is evidence that the executives in this sample are, and some were formally charged and convicted of crimes. Another limitation is the small sample size (six companies, 24 transcripts). Future research should address this by collecting a larger sample and using companies outside of the financial services sector. As one reviewer suggested, the inclusion of all voluntary disclosures (in addition to the financial statements) would improve this research.

Additionally, I was not able to explore the proposed causal mechanism behind the higher level of similarity: that this higher level of similarity is a strategic method to limit disclosure new information. A method to test this will be crucial to future of this research.

It would also be interesting to see how well this technique works in other areas where responses are scripted and deception may occur, and we have access to a script. Finance researchers may want to see how the market and analysts react to varying levels of scripted responses.

References


