Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach

Allen Huang
Hong Kong University of Science & Technology
Department of Accounting

Reuven Lehavy
Stephen M. Ross School of Business
University of Michigan

Amy Zang
Hong Kong University of Science & Technology
Department of Accounting

Rong Zheng
Hong Kong University of Science & Technology
Business Statistics and Operations Management

Ross School of Business Working Paper
Working Paper No. 1229
June 2015

Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach

Allen Huang
Department of Accounting
Hong Kong University of Science and Technology
allen.huang@ust.hk

Reuven Lehavy
Ross School of Business
University of Michigan
rlehavy@umich.edu

Amy Zang
Department of Accounting
Hong Kong University of Science and Technology
amy.zang@ust.hk

Rong Zheng
Department of Information Systems, Business Statistics and Operations Management
Hong Kong University of Science and Technology
rzheng@ust.hk

June 2015

Allen Huang, Amy Zang and Rong Zheng would like to thank the financial support provided by HKUST. Reuven Lehavy would like to thank the financial support of the Harry Jones Endowment Fund. We would also like to thank Weining Zhang, Bin Zhu, Qiang Cheng, Lian Fen Lee, seminar participants at the HKUST, Nanyang Business School, Singapore Management University, Southern Methodist University, Boston College, Tel Aviv University, Tsinghua University, Shanghai University of Finance and Economics, and Chinese University of Hong Kong, and the participants at the 2014 Workshop on Internet and BigData Finance (WIBF), 2014 China Summer Workshop on Information Management (CSWIM), and 2014 MIT Asia Conference in Accounting for helpful comments.

Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach

Abstract

This study examines the analyst information discovery and interpretation roles using a textual analysis of analyst research reports and management discussions during earnings conference calls. Empirically, we employ an advanced topic modeling methodology from computational linguistic research to compare the thematic content of a large sample of analyst reports issued promptly after earnings conference calls with the content of the calls themselves. Our results show that analysts meaningfully engage in both discovery and interpretation in their written reports. Additionally, we find that the proportion of new information in analyst reports relative to that in conference calls increases when firms face a higher level of competition, have a greater litigation risk, operate in a more volatile information environment, or have bad news to convey in their conference calls. We also find that the analyst interpretation function is magnified when information processing costs increase, specifically when conference calls include more uncertain or qualitative statements, firm operations are more complex, or firms convey bad news. Finally, we provide evidence that investors value both the information discovery and information interpretation roles. Overall, our results enhance our understanding of the important information intermediary role analysts play in capital markets.

1. Introduction

Previous research has documented the important information intermediary role that analysts play in capital markets (e.g.,Frankel, Kothari, and Weber 2006; Ramnath, Rock, and Shane 2008). Specifically, Ivkovic and Jegadeesh (2004), Chen, Cheng and Lo (2010), and Livnat and Zhang (2012), among others, have identified the value of analysts to investors in both discovering new information and interpreting existing information. In their information discovery role, analysts provide value to investors by collecting, processing, and providing information that is otherwise not readily available. Analysts may obtain this information from a variety of public and private channels such as store visits to collect information on traffic, customer surveys to evaluate customer satisfaction or product quality, supplier evaluations to assess potential input shocks, and competitor research to understand a firm's competitive advantage. In addition to obtaining information, analysts provide value to investors through their ability to interpret information. In their information interpretation role, analysts clarify publicly available information, offer their opinions on issues raised through public disclosures, compare information with objective benchmarks, and quantitatively assess management's subjective statements.

Evidence in the extant literature suggests that investors value both analyst information roles. Ivkovic and Jegadeesh (2004) and Chen et al. (2010) find that investors value the ability of analysts to discover new information more than their ability to interpret earnings announcements. In their study of corporate disclosures beyond the earnings announcement, Livnat and Zhang (2012) find that investors value the interpretation role more than the discovery role. Finally, Francis, Schipper, and Vincent (2002) and Frankel et al. (2006) find that analyst research and

earnings announcements complement each other in terms of informativeness, suggesting that investors value the analyst information interpretation role.

While the extant literature provides insight regarding investor perceptions of the analyst information discovery and interpretation roles, it does not examine the economic determinants that drive analysts to serve in either role nor does it examine what types of information they choose to provide. This study fills this gap in the literature by providing an in-depth analysis of the information embedded in the textual content of analyst research reports relative to the information content of management discussions during earnings conference calls. Specifically, we use a large sample of analyst reports issued on the same day and the day following the quarterly earnings conference call (hereafter, prompt reports) to investigate three research questions: (1) Do analyst reports issued promptly after conference calls provide incremental information (both new and interpretive) related to the discussions in these calls? (2) When do analysts play an information discovery role and when do they play an information interpretation role? And (3) do investors value the analyst information interpretation and discovery roles?[1]

Quarterly earnings announcements and their related conference calls are arguably the most important corporate disclosures made by a firm. Their importance is reinforced by the overwhelming number of sell-side analyst research reports issued immediately following these events. Indeed, we find that 46.5% of analyst revision reports within a given year are issued either on the day of or the day after an earnings conference call. Despite their concentration around the event window, however, relatively little is known about the information content of these prompt reports. One reason for this lack of attention is the difficulty in separating the

[1] We focus on analyst reports issued in this short window to avoid the potentially confounding effects of other information that may have been released between the time of the conference call and the issuance of the analyst reports.

content of the reports from the potentially confounding effect of the earnings disclosure event in close proximity[2]. To mitigate this difficulty, we focus on the specific information attributes of the narratives analysts provide during this short window relative to management disclosures in the conference call. This approach helps us better understand how analysts convey information to capital markets, and ultimately their value to investors.[3]

Our investigation into the economic determinants of analyst information interpretation and discovery roles is important because it provides insight into how managers, analysts, and investors together shape a firm's information environment. Specifically, prior research finds that managers make voluntary disclosure decisions based on how they anticipate capital market information intermediaries will react (Dutta and Trueman 2002). Thus, managers consider how well analysts can provide new information and interpret existing information immediately after a corporate disclosure event. Which ability a manager values more depends on the firm's circumstances. For example, a firm facing higher litigation risk may need to rely more on the analyst information discovery role to reduce the information asymmetry between insiders and investors. Conversely, a firm disclosing more complex or uncertain news may need to rely more on the analyst interpretation role to reduce potential misunderstanding and the consequent excess stock price volatility caused by diverse opinions.

[2] Studies that rely on market reactions often exclude these reports from their samples (e.g., Chen, Cheng, and Lo 2010; Green, Jame, Markov, and Subasi 2014; Loh and Stulz 2011).

[3] Analysts are not expected nor assumed to be able to discover new information, write their complete reports and provide new and interpretive information to their clients within hours of the conference call. A more realistic view is that analysts continuously engage in information acquisition and analysis on the stocks they cover throughout the quarter (e.g., industry analysis, analysis of competitors, conducting store visits, and discussions with management). Once earnings are announced and the conference call is held, they put the "final touches" on their report (among other, decide on the extent of new and interpretive information to include) and release the reports to their clients (who, of course, demand the reports as soon as possible). That, empirically, the overwhelming number of analyst reports are issued immediately after the conference call supports this view.

Investors may also be concerned with whether analysts are capable of promptly responding to varying economic conditions and meet their information demands. For example, investors may be concerned with whether analysts generate more private information if managers withhold information. Or, they may be concerned with whether analysts use their expertise to process publicly available information if managers provide more complex corporate disclosures. Our study contributes to our understanding of these questions by examining the economic determinants of analyst information roles. In doing so, our paper heeds the call for such insight from a number of studies, including Ramnath, Rock, and Shane (2008), Bradshaw (2011), and Brown et al. (2014).

To empirically measure analyst information roles, we conduct a large-scale comparison of the thematic content of prompt analyst reports (denoted AR) to that of manager narratives in the earnings conference calls (denoted CC).[4] Specifically, we employ a topic modeling approach called *Latent Dirichlet Allocation* (LDA) to construct novel measures of the economically meaningful topics discussed in analyst reports and conference calls. Developed by Blei, Ng, and Jordan (2003), LDA is an advanced textual analysis technique grounded in computational linguistics that calculates the statistical correlations among words in a large set of documents to identify and quantify the underlying topics in these documents. LDA can be thought of as a dimensionality-reduction technique, similar to cluster analysis or principle components analysis, but designed specifically for clustering words in a given text.

Within the context of our study, we use LDA to identify economically interpretable topics within analyst research reports and conference call transcripts. LDA provides several advantages

[4] Conference call narratives include both manager discussions in the presentation and the Q&A (question and answer) parts of the conference call. Analyses based on only the presentation part of the conference calls yield similar results.

for our study. First, the LDA algorithm is able to handle a massive collection of documents that would otherwise be impossible for human coders to process. Second, as an unsupervised statistical learning method, LDA does not require any pre-specified set of topics or labeling effort from researchers. These features make LDA well-suited for analyzing the thematic content of large volumes of textual data within fields from political science to psychology to economics.[5]

We base our analysis on a sample of 17,750 conference call transcripts and over 160,000 prompt analyst reports. In our sample, we find that management discussions in conference calls contain an average of 29 topics while analyst reports issued promptly after these calls contain an average of 26 topics. However, we find that 83% of the content of the conference calls (analyst reports) is devoted to discussions of 10 (9) economically meaningful topics. Using LDA and standard validation procedures, we find that managers and analysts routinely discuss economically meaningful topics related to growth, financial performance (current and outlook), business outlook, cash flow and financing, and valuation. In addition, we find that managers and analysts discuss industry-specific topics such as drilling in the energy industry, Internet advertising in the software industry, and drug trials in the health care industry.

To validate our approach, we also analyze subsamples in the banking and telecommunication industries during the period 2003 to 2012. For each subsample, we visually examine whether topic trends correspond to key economic developments in these industries (a similar analysis is conducted by Quinn et al. (2010) in their study of the text of the U.S. Senate Congressional Record). A plot of the temporal variation in the proportion of discussions devoted to key topics for our sample visually confirms that manager and analyst discussions closely track economic developments in the respective industries. For example, during the financial crisis, we find that

[5] See for example, Quinn et al. (2010), Grimmer (2010), Atkins, Rubin, Steyvers, Doeden, Baucom, and Christensen (2012), Bao and Datta (2012), Griffiths and Steyvers (2004) and Kaplan and Vakili (2013).

managers and analysts in the banking sector spend more time discussing mortgage-related issues and deteriorating financial performance and less time discussing mortgage origination and loan growth. Similarly, we find that managers and analysts in the telecommunications sector gradually shift their discussions from landline services in the early part of our sample period to the smartphone business in the latter part of our sample period. We interpret these validation tests as support for the use of LDA to meaningfully discover the thematic content in conference call transcripts and analyst research reports.

After validating our empirical approach, we turn our attention to the different information roles played by analysts compared to the content in the immediately preceding earnings conference calls. We classify analyst discussions of topics beyond those raised by managers during the earnings conference calls as information discovery activity because such discussions are more likely to capture analyst private research and information generation. For example, in a Morgan Stanley report issued on August 12, 2011, immediately after J.C. Penney's conference call, an analyst notes: "*The top reason consumers say they shop JCP is due to 'low prices, great discounts' (as per our most recent consumer survey).*" This example reflects private information generated by the analyst.[6] Likewise, we find that banking industry analysts are more likely than firm managers to discuss topics such as "risk and loss," "comparison to peers," and "loan quality" for our banking sample observations surrounding the financial crisis (2008-2009). Overall, we find that analysts, on average, spend 27% of their discussion on exclusive topics that receive little attention by managers, suggesting that analysts serve an information discovery role.

[6] According to the LDA algorithm, this sentence in the analyst report is classified as related to a "consumer survey and brand" topic of the retailing industry. This topic was not discussed by managers during the corresponding conference call.

By contrast, we classify analyst discussions of topics raised by management in conference calls as information interpretation activity, as these discussions are more likely to reflect the analyst ability to process, discern and evaluate management statements.[7] Empirically, we measure analyst interpretation activity by comparing the respective vocabularies used by analysts and managers to discuss the key conference call topics. When the textual characteristics of analyst discussions of the CC topics are significantly different from those of management discussions of these topics, we consider this as evidence of an interpretation role. Our empirical analyses indicate that the vocabulary used by analysts to describe the key CC topics statistically differs from that of management in 59% of our sample observations. This difference is positively correlated with the difference between CC and $AR's$ percentage of uncertain and quantitative words, readability, and tone.

Given that we find evidence of both an information discovery and interpretation role for analysts in our sample, we next examine the economic determinants that drive analysts to serve in either information role. Specifically, we hypothesize that analysts will respond to investor demand for additional, new information when managers withhold information by increasing the amount of private information disclosed in their analyst reports. We further hypothesize that analysts will respond to investor demand to clarify information by increasing their efforts to interpret management information when the cost of processing information is high. Evidence from our empirical tests supports each prediction. First, we find that the amount of new information in an analyst reports increases when firms face greater competition, higher litigation

[7] Analysts may need to use some private information to interpret an existing disclosure provided by the manager. However, as discussed in Section 5, as long as our topic-modeling algorithm statistically indicates that analysts and managers discuss the same economic topic, we consider this scenario as information interpretation by the analyst. By contrast, if the analyst discussion is sufficiently different from the manager's disclosure such that LDA classifies these discussions as relating to exclusive topics, we consider this as analyst information discovery. Our empirical methods aim at operationalizing the concepts of information discovery and interpretation but do not assume that these two information roles are mutually exclusive.

risk, or more intense information volatility as well as when managers are delivering bad news. Second, we find that the amount of report information related to analyst interpretation increases with several empirical proxies for the information processing cost, including a greater number of uncertain and qualitative statements in the calls, greater complexity in firm operations, and more information complexity due to bad news.

Finally, we examine how investors value the information discovery and interpretation roles. Our findings indicate that investors value each role. We further find that investor reactions to these information roles are incremental to their reaction to the firm's earnings news, other information provided by managers during the conference calls, and other research outputs in the analyst reports such as the revisions of earnings forecasts, stock recommendations and target prices.

Our study provides several contributions to the literature. First, we contribute to the literature by providing further insight into the information intermediary function of analysts, including the determinants of their information discovery and information interpretation roles. This provides important evidence within the growing body of literature on the relative value of analyst report text. Second, our study contributes to the literature by applying a computational linguistic methodology to our development of novel measures of the information content in textual content. Notably, we develop measures of information content based on the semantic discussions of economically meaningful topics in analyst reports and manager disclosures. Our measures based on topic identification are in contrast to those of other studies that measure disclosure content through the immediate market reaction to the release of these disclosures. By focusing on the topics in analyst reports, we are able to avoid any confounding effect of other disclosures inherent in the market reaction approach. Finally, our study contributes to the emerging area of

textual analysis in the accounting and finance literature. Much of this research focuses on the textual characteristics (e.g., readability and tone) of corporate financial disclosures (e.g., MD&A in 10-K and S-1). Our topic modeling methodology provides another avenue through which researchers can expand their analysis of the textual content of corporate financial disclosures from "*how* texts are being said" to "*what* is being said" in these disclosures.

The rest of our paper is organized as follows. Section 2 reviews the related literature. We introduce our empirical methodology in Section 3. Section 4 discusses our sample selection procedure. Section 5 presents the summary statistics for our topics in AR and CC. Sections 6 and 7 describe the key empirical measures, empirical tests and results. We conclude the study in Section 8.

2. Review of related literature

Extant research examines the relative importance of the analyst information discovery and interpretation roles (see Ramnath et al. 2008 for a review). For example, several early studies (e.g., Dempsey 1989 and Shores 1990) find evidence that supports the analyst information discovery role. Specifically, these studies find that the market reaction to an earnings announcement decreases with analyst coverage of a firm. In another set of studies, Francis et al. (2002) and Frankel et al. (2006) find support for an analyst interpretation role. These studies find that the information content of analyst reports complements that of earnings announcements. In an examination of the relative importance of analyst discovery and interpretation roles, Ivkovic and Jegadeesh (2004) find that the discovery role dominates. Specifically, they find that the market reaction to analyst revisions during the week after an earnings announcement is weaker compared to other periods. Likewise, Chen et al. (2010) conclude that the discovery role dominates, based on their finding of a negative association between the market reaction to

earnings announcements and that of analyst reports before and after earnings announcements, with the exception of the first week after an earnings announcement.[8] However, Livnat and Zhang (2012) find that the interpretation role dominates when other types of public corporate disclosure and analyst reports issued within three trading days after the public disclosure are included in the analysis. Specifically, they find that the majority of analyst reports fall into the category of "prompt reports" and that these reports trigger a greater market reaction (measured as the three-day abnormal returns centered on the report date) than do non-prompt reports. This conclusion is in contrast to those in Chen et al. (2010) and Ivkovic and Jegadeesh (2004), but is consistent with those of Francis et al. (2002) and Frankel et al. (2006).

In each of the above studies, the information content of analyst reports is measured through the immediate market reaction to the issuance of these reports. However, as mentioned, one potential limitation of this measure is that market reactions may be influenced by concurrent and adjacent disclosures. The market reaction measure also makes it difficult to disentangle coexisting information discovery and interpretation roles. For example, Ivkovic and Jegadeesh (2004), Chen et al. (2010), and Livnat and Zhang (2012) all assume that prompt analyst reports focus on interpretation while non-prompt reports focus on discovery. However, it is possible that an analyst revision issued promptly after a public disclosure may contain both the analyst's own private information and the analyst's interpretation of the firm's public disclosures. Similarly, a non-prompt revision may reflect both private information and a belated interpretation of previously-disclosed public information. Further, Chen et al. (2010) and Francis et al. (2002) interpret a positive (negative) relation between the market reaction to earnings announcements

[8] Chen et al. (2010) measure information content as the absolute value of the abnormal stock return on the announcement date, and exclude a large number of analyst reports issued on days [-1, +1] relative to the earnings announcement dates to mitigate any potential confounding effect of the two information events.

and the reaction to analyst reports as evidence for an information interpretation (discovery) role. However, it is possible that a positive correlation between the market reaction to earnings announcements and the reaction to analyst reports can be driven by both analyst interpretation and discovery.

Within the area of textual analysis, Asquith, Mikhail, and Au (2005) study report tone by manually categorizing the content of 1,126 reports issued by 56 *Institutional Investor* All-American "First Team" analysts into 14 justification variables. They find that the market reacts to report tone conditional on earnings forecasts, stock recommendations, and target prices. In another study, Kothari, Li, and Short (2009) use a dictionary method based on the General Inquirer to classify analyst report text as positive or negative. They find an insignificant relation between analyst report content and the cost of capital, suggesting that analysts might respond to market events after the events have taken place. Using a naïve Bayes machine learning approach, Huang, Zang, and Zheng (2014) classify the textual opinions in over 360,000 analyst reports and find that the incremental information content of analyst reports is economically significant, with its cross-sectional variation explained by report characteristics. Finally, DeFranco, Hope, Vyas and Zhou (2014) examine the readability of analyst reports and find that a more readable report triggers greater stock trading volume. Overall, these studies underscore the importance of examining the textual content of analyst reports to better understand their information intermediary role in capital markets. [9]

In addition to its contribution to research on the role of analysts and the content of analyst reports, our study relates to research exploring different applications of the topic-modeling

[9] In related literature, Bushee, Core, Guay and Hamm (2010) find that the business press provides value both by disseminating information and by creating new information beyond the firm disclosure. In this study, we view analyst information dissemination role as a special case of a low level of information interpretation (or zero at the extreme) when analysts simply 'parrot' management discussions.

methodology. Topic modeling, or LDA, has been used extensively in a variety of fields to analyze the textual content of large volumes of linguistic data. Examples of influential studies within this area range from Quinn et al. (2010), who use LDA to analyze legislative speech and political attention, to Griffiths and Steyvers (2004), who use it to analyze the abstracts from the Proceedings of the National Academy of Sciences to identify "hot topics." While LDA has been used in a number of different fields to discover the thematic content in linguistic data, its use in accounting and finance research has been fairly limited. One study in this area that uses LDA is that of Ball, Hoberg, and Maksimovic (2014), who extract topics from the MD&A part of corporate 10-K filings to measure corporate disclosure quality. In another study, Bao and Datta (2014) use a variation of the LDA model to summarize the risk-related topics contained in the risk disclosure section (section 1A) of corporate 10-K filings. Finally, Lang and Stice-Lawrence (2014) employ LDA to examine the impact of IFRS adoption on the topics disclosed in annual reports. Our research adds to this stream of research by using LDA to study the economic determinants of the analyst discovery and interpretation roles.

3. Empirical methodology

3.1. Topic modeling and Latent Dirichlet Allocation

In this section, we describe our methodology. We obtain our empirical measures of analyst information discovery and interpretation roles by comparing the textual narratives contained in AR to those in CC at both the topic level and the word level. To obtain our topic measures, we use a computational linguistic technique to uncover the thematic structure of linguistic data by automatically analyzing the semantic content in large collections of this data (Blei 2012). These topic modeling algorithms provide a topic annotation of documents by uncovering a set of hidden topics and assigning individual words to specific topics. Topic modeling is similar to

other dimensionality-reduction techniques, such as cluster analysis or principle component analysis, but is designed for use with text.

Topic modeling has several desirable features for our analysis. First, it is capable of processing a massive collection of documents that would be impossible for human coders to process. Second, it does not require training or topic pre-specification, implying that the entire procedure is consistent and replicable. Finally, the resulting topics, presented as sets of coherent words, are typically interpretable. With this feature, we can discern the economic interpretation of the identified topic (Blei 2012; Quinn et al. 2010). In short, topic modeling allows us to analyze textual data at the topic level.

Latent Dirichlet Allocation (LDA)

Introduced by Blei et al. (2003), LDA has become the most widely used topic-modeling algorithm. LDA uses a statistical procedure to imitate the process of how a human author writes a document. Specifically, the algorithm assumes that the author writes each word in a document in two steps (see Appendix I.a for illustration). First, the author selects a topic from the distribution of all available topics. While all documents share the same set of topics, each document has its own topic distribution, i.e., some topics are more likely in certain documents than in other documents. Second, given the topic, the author selects a word from the distribution of all words representing this topic. Note that, while there is only one common vocabulary, each topic has its own word distribution (i.e., some words are more common in certain topics than in other topics). Given these assumptions, LDA implements a Bayesian procedure to find the model that best fits the textual data. The procedure determines the model parameters based on word co-occurrences. If two words appear frequently in the same document, there is a higher likelihood that LDA will assign them to the same topic (see Appendix I.b for a detailed technical

description of the LDA estimation process). The output from the LDA algorithm comprises a matrix of word frequencies in each topic. The probability of a word appearing in a given topic is determined from its frequency in that topic divided by the total frequency of all words in that topic.

As described in Appendix I.c, prior to applying the LDA algorithm, we applied several standard preprocessing steps to the conference call transcripts and analyst reports and set the number of topics to 60 based on the output of the Perplexity Score.[10] We then perform the LDA analysis at the industry level by estimating the topic distribution in the combined set of available conference call transcripts and analyst reports in each industry. Based on the LDA output, we then annotate the content of each document, d, by constructing the topic vector of AR (CC) in which each element describes the fraction of AR (CC) that is dedicated to a discussion of each topic. Formally:

$$Topic\ vector\ of\ document\ d = T_d = (S_{d1}, S_{d2}, \ldots S_{d60}), \quad (1)$$

where S_{dk} represents the fraction of the discussion in document, d, that is devoted to the topic, k.

3.2. Validation tests of the LDA output

To test the validity of the LDA choice of topics, we follow the procedure in Quinn et al. (2010), Atkins et al. (2012), and Bao and Datta (2012), and manually read the high-probability words in key topics and their respective sentences. This procedure allows us to discern the underlying economic content of the topic. Table 1 presents the results of applying this validation technique to our sample. The table reports the top 20 words in each of the top ten topics as well

[10] We compared LDA results based on 30, 60, and 100 topics. Based on our comparison, we conclude that the LDA results with 60 topics outperform the other specifications in identifying intuitively important topics without generating an undue number of uninterpretable topics. We repeat our main analysis using LDA results based on 30 and 100 topic specifications and find qualitatively similar results for each specification.

as our inferred topic labels. We present the results for the five largest industries represented in our sample (ranked by the total number of conference calls).

[Insert Table 1 here]

Overall, the results in Table 1 validate the effectiveness of the LDA algorithm in identifying distinct, economically meaningful topics in conference calls and analyst reports. Specifically, our manual analysis verifies that the words assigned by the LDA algorithm to a specific topic appear semantically related. For example, the frequent appearance of the words "multiple," "target," "price," "valuation," "eps," and "PE," in a topic in the Capital Goods industry suggests that this topic is related to valuation models and target price. Similarly, the frequent appearance of the words "drug," "trial," "announce," "clinical," and "phase," in a topic in the Health Care Equipment & Services industry suggests that this topic relates to drug trials. We also find that LDA effectively uncovers general topics related to a firm's financial performance as well as industry-specific topics. For example, among the top ten topics, all industries contain discussions of growth- and performance-related topics. In addition, the LDA algorithm accurately identifies industry-specific topics such as offshore drilling in the energy industry, enterprise software and IT services in the software industry, and steel production in the materials industry. Finally, our results verify that the LDA algorithm recognizes the polysemy or contextual nature of words by assigning the same word to multiple topics. For example, its classification of the word "price" in both "Valuation" and "Raw Materials and Input Price" in the Capital Goods industry reflects the contextual definition of the word. Overall, the evidence in Table 1 suggests that the output from the LDA model provides a reliable delineation of economically meaningful topics for the analyst reports and conference call transcripts.

In addition to conducting a manual validation of the LDA algorithm, we compare the temporal variation in the amount of discussion dedicated to key topics with important contextual events. This comparison allows us to validate our methodology even further. The relation between temporal variation and contextual events is seen in a study by Quinn et al. (2010) who find that the proportion of key political topics in the Congressional Record tracks exogenous events such as the 9/11 attack and the Iraq War. In our study, we visually examine whether the temporal variation in the weight assigned to key topics is related to changes in industry and economy-wide conditions. We depict this relation in Figure 1, which illustrates the proportion of key topics in earnings conference calls and analyst reports for the banking and telecommunication industries from 2003 to 2012, and the performance of their respective sector indices (Financial Sector SPDR – XLF and iShares US Telecommunications – IYZ index, respectively). We select these two industries based on the turmoil in the banking industry and the technology evolution in the telecommunication industry during the period of our study.

Panel A of Figure 1 presents visual evidence of a reliable relation between the temporal variation in the distribution of key topics and economic performance in the banking industry. For example, from 2003 to 2006, management and analyst discussions are devoted primarily to the topics of "Growth" (mostly in loans and deposits) and "Mortgage Origination." However, the discussion of these topics declines substantially in 2007, with the advent of the financial crisis, while that of "Real Estate Loans" and "Performance and Losses" increases. Not surprisingly, after the approval of the Troubled Asset Relief Program, or TARP, in October 2008, we see an increase in discussions of the topic "Equity Issuance and TARP."

Panel B of Figure 1 depicts the relation between technological developments and topic discussion for the telecommunications industry. Here, we see that landline-related topic

discussions (e.g., DSL technology) decrease during our sample period while topics labeled "Smartphone Business," and "Wireless Subscribers" increase. Taken together, we interpret the evidence of the validation tests presented in Table 1 and Figure 1 as supporting the effectiveness of LDA to qualitatively identify and quantitatively measure economically meaningful contextual topics in the earnings conference calls and analyst reports in our study.

4. Sample selection

Our sample is comprised of quarterly earnings conference calls transcripts and analyst reports issued on the same day or the day following these conference calls for all S&P 500 firms from 2003 to 2012.[11] To begin, we obtain our sample of conference call transcripts from the Thomson Reuter Streetevent Database. We begin with 2003 as the database coverage of conference calls prior to 2003 is incomplete.[12] Table 2 describes our sample selection criteria. As shown in Panel A, to obtain our final sample, we first identify 18,607 earnings conference call transcripts. To verify these are earnings conference calls, we match them with earnings announcements from I/B/E/S. This matching reduces our sample to 18,236 conference calls that occurred during days [0, +1] relative to the earnings announcement dates (this is the sample used in the LDA model). We next require each conference call to be accompanied by an analyst report. This requirement yields a final sample of 17,750 earnings conference calls with matched analyst reports.[13]

[Insert Table 2 here]

[11] Our sample firms constitute, on average, about 72% of the total U.S. market capitalization, or 77% of the total U.S. firms covered by analysts. We acknowledge that our findings based on a sample S&P 500 firms might not directly apply to smaller firms that receive less analyst coverage.

[12] There are only 270 conference calls in 2001 and 1,379 conference calls in 2002 for S&P 500 firms in the Thomson Reuter Streetevent Database. For comparison, in 2003-2012, the database contains around 1,900 to 1,950 conference calls for S&P 500 firms.

[13] Thomson Reuters Streetevent Database provides tickers of firms hosting the conference calls. We manually match the conference calls to Compustat's S&P 500 list using these tickers. For analyst reports, we extract firms' ticker from each analyst report and manually match the report to Compustat's S&P 500 list using its ticker.

We obtain our sell-side analyst reports from the Investext Database. As reported in Panel B of Table 2, we include all reports issued for S&P 500 firms during 2003-2012. This yields an initial sample of 476,633 reports, which we use to perform our LDA analysis. We then exclude reports not issued on the day of or the day following an earnings conference call. We also exclude reports issued on the day of but prior to the start time of a call. We impose these criteria to avoid any potential confounding effects of new information issued between the end of the CC and the issuance of the AR. Our final sample is comprised of 159,210 analyst reports. Panel B of Table 2 shows that the proportion of analyst reports issued for S&P 500 firms on the day of or the day after a quarterly conference call constitutes 33% of the entire population of analyst reports (or 47% if we only consider revision reports); these statistics reinforce the importance of the conference call and the analyst reports as an important corporate disclosure event.

The statistics in Panel C of Table 2 show that the number of conference calls increases steadily from 1,605 in 2003 to 1,886 in 2012. We also see that the number of prompt analyst reports and the number of reports per call dips in 2008 to 13,368 and 7.46, respectively, but reaches a high of 22,343 and 11.85 in 2012. Over the entire sample period, an average of nine analysts issue reports in the two-day window after a quarterly conference call. Since our focus is on the information role of analysts in aggregate, we treat all analyst reports issued during this two-day window as a single report and denote it as AR.

Finally, Panel D presents the GICS industry composition for the firms in our sample. These statistics show that the industries with the largest number of earnings conference calls in our sample are capital goods, energy, software and services, materials, and health care equipment and services.

5. The distributions of topics discussed in earnings conference calls and analyst reports

Number of topics in prompt analyst reports and conference calls

Table 3 reports the summary statistics for the number of topics identified by the LDA algorithm in our earnings conference calls and analyst reports. From Panel A, we see that the earnings conference call management discussion (CC) contains an average of 29 distinct topics. These topics consist of an average of 22 topics in the presentation portion (CCP) and 23 topics in the management answers to analyst questions (CCA). By comparison, we see that the set of analyst reports issued promptly after a given call (AR) contains an average of 26 topics; however, the relatively high standard deviation in the number of topics in AR suggests a greater variation in the thematic content of these reports than in the conference calls.

Panel B of Table 3 provides the summary statistics for the number of topics whose weight in a given document exceeds 2.5% of the entire length of the document. These statistics show that the CC (AR) in our sample discusses an average of 10.5 (9.6) key topics with a standard deviation of around 2; the combined length of these key topics accounts for over 83% of the entire discussion in the CC (AR). Overall, the summary statistics reported in Table 3 indicate that managers and analysts devote most of their discussion to ten key topics. Accordingly, in our subsequent empirical analysis, we focus on the management discussion of the top-ten topics in the CC.[14]

[Insert Table 3 here]

[14] We consider the top ten topics in a conference call as key topics. In total, these topics account for about 81% of the CC discussion. In untabulated results, we find that, on average, managers spend only 0.37% of their total narrative on non-top ten topics (median being 0.00%). Therefore, we do not include these topics in our empirical tests as they might introduce noise. As a robustness check, we rerun our empirical tests with the top eight and top 12 topics of CC and find similar results.

Difference in the topic distributions of conference calls and prompt analyst reports

In our first set of tests, we examine the difference between the respective topic proportions in the analyst and manager narratives. To do so, we conduct a Pearson's chi-square test for the homogeneity of the distribution of topics discussed in each AR and CC pair (i.e., we test the null that $T_{CC} = T_{AR}$; see equation 1 in Section 3.1). Panel C of Table 3 presents the results of these tests. The results in Panel C show that the mean (median) value of the chi-square statistic across all 17,750 pairs of AR and CC is 145.9 (137.4), indicating that the homogeneity between the topic distribution in these documents is rejected 90.8% of the time (significant at the 10% level). That is, in 90.8% of the AR-CC pairs, managers and analysts devote different proportion of narratives to each topic.

To put these results in context, we conduct a benchmark test comparing the *AR* topic distribution to the respective topic distributions in the following: the analyst comments during the Q&A portion of the call (CCQ), the manager presentation portion of the call (CCP), and the manager answers during the Q&A portion of the call (CCA). These results show that the AR topic distribution is most similar to that of the analyst comments during the Q&A portion. We also find the value of the chi-square statistic of the topic distribution difference between CCQ and CCA is only 29.1, indicating that analyst questions and manager answers have similar topic distributions (homogeneity is rejected for only 0.06% of the time at the 10% significance level). This finding provides further validation for our topic measures.

6. Empirical measures, tests and results

6.1. *Measuring the analyst information discovery role*

We use the output from the LDA analysis to develop our definition for the analyst discovery role. We define this role as the proportion of an analyst report discussion dedicated to exclusive

topics. Specifically, we measure analyst information discovery as the proportion of AR discussion devoted to economically meaningful topics that receive little or no attention by managers during the CC. This definition assumes that exclusive topics are likely to reflect analyst effort to present private information to investors. We denote this variable $Discovery$. The statistics in Table 4 show that analysts spend an average of 27.3% of their discussion on $Discovery$ topics. Appendix II provides excerpts from conference call transcripts and analyst reports that illustrate topics classified as analyst information discovery.

6.2. Measuring the analyst information interpretation role

We next use a statistical comparison of the words used by managers and analysts to discuss the most important CC topics to develop our definition of the information interpretation role. Analysts' efforts to process an existing management disclosure, explain or transform it to a more meaningful narrative should manifest itself in a different word usage from that of managers. Accordingly, we define the analyst interpretation role by the extent to which analysts use different words than managers to discuss economically meaningful topics.

Empirically, we construct vectors of word usage for each topic discussed in the AR and the CC:

$$\begin{aligned} Word\ vector\ of\ topic\ k\ in\ CC &= W_{CC,k} = (v_{1k}, v_{2k}, \dots v_{Nk})\ ; \\ Word\ vector\ of\ topic\ k\ in\ AR\ &= W_{AR,k} = (w_{1k}, w_{2k}, \dots w_{Nk}); \end{aligned} \tag{2}$$

where each element of these vector (v_{wk}) is the frequency of word w in the discussion of topic k in the respective document (N is the total number of unique words in the corpus).[15]

[15] In an untabulated analysis, we use the Pearson's chi-square to test the null that $W_{AR,k} = W_{CC,k}$ in Eq. (2) for each of the top ten CC topics. We find that the homogeneity between the distribution of words used to describe these topics in CC and AR is rejected (at the 10% level) for 59.0% of the sample, suggesting that promptly after the CC, analysts provide meaningful interpretation for an average of six of the top ten CC topics.

To measure the extent to which prompt analyst reports provide the analyst's interpretation of the information contained in the CC, we compute the average difference between $W_{AR,k}$ and $W_{CC,k}$ across the top ten CC topics (denoted $Interpret$). We calculate the difference between the word vectors as one minus the cosine similarity between $W_{AR,k}$ and $W_{CC,k}$.[16] Cosine similarity captures the textual similarity between the two vectors. This similarity is neatly bounded by [0, 1], and is widely used in information retrieval research to compare textual documents (e.g., Singhal 2001; Hanley and Hoberg 2010; and Brown and Tucker 2011). With this definition, low values of $Interpret$ mean that the analysts and the manager use similar words, while high values mean that the analysts use different word choices. Formally, we represent our definition as follows:

$$\begin{aligned} Interpret &= \frac{1}{10}\sum_{k=1}^{10}\left(1 - cosine\ similarity\ between\ W_{AR,k}\ and\ W_{CC,k}\right) \\ &= \frac{1}{10}\sum_{k=1}^{10}\left(1 - \frac{\sum_{j=1}^{N}\left(w_{jk}\cdot v_{jk}\right)}{\sqrt{\sum_{j=1}^{N}\left(w_{jk}\right)^2}\cdot\sqrt{\sum_{j=1}^{N}\left(v_{jk}\right)^2}}\right), \end{aligned} \tag{3}$$

where, w_{1k} is word 1's frequency in the discussion of topic k in the AR; v_{1k} is word 1's frequency in the discussion of topic k in the CC; N is the total number of unique words in the corpus; and k is one of the top ten topics discussed in the CC. The results in Table 4 show that

[16] Cosine similarity is computed as the dot product of the two vectors normalized by their vector length, and captures the textual similarity between two vectors of an inner product space using the cosine angle between them. Two vectors with the same orientation (i.e., two exact same topic vectors) have a cosine similarity of one; two orthogonal vectors have a similarity of zero. To illustrate, we assume there are two topics and two documents. In one document, 30% of the sentences relate to topic 1 and 70% relate to topic 2. In the second document, 60% of the sentences relate to topic 1 and 40% of the sentences relate to topic 2. The cosine similarity of their topic distributions is: $(0.3 \times 0.6 + 0.7 \times 0.4)/\sqrt{(0.3^2 + 0.7^2) \times (0.6^2 + 0.4^2)} = 0.8376$.

the analyst reports in our sample have an average $Interpret$ level of 0.51, consistent with the existence of their interpretation role immediately after earnings conference calls.

Appendix II provides two illustrative examples with high and low levels of $Interpret$ from excerpts of conference calls transcripts and analyst reports. To further validate our choice of $Interpret$, we examine the correlation between $Interpret$ and analyst/manager differences in the following linguistic features of their respective narratives: the percentage of uncertain words (Loughram and McDonald 2013), the percentage of quantitative words (Huang et al. 2014), the narrative readability level as measured by the Fog Index (Li 2008) and the Flesch Index (Smith and Smith 1971), and the tone of the analyst discussion (Huang et al. 2014). Our results, untabulated, show that $Interpret$ increases with the difference (measured as the linguistic attribute in CC less that in AR) in the percentage of uncertain words and the percentage of quantitative words, and with the difference in the readability level and the absolute difference in the tone. These findings suggest that when serving interpretation role, analysts supplement the information contained in a CC topic by providing different opinions on the topic, and that they facilitate investors' understanding of a topic by using fewer uncertain or quantitative words. Interestingly, the results also show that when $Interpret$ level is high, analyst narratives become less readable relative to the manager's discussion. We interpret this finding as an indication that the function of interpretation is reflected in longer, more informative sentences, resulting in a slightly lower readability score.[17] Taken together, these findings suggest that the analysts in our sample demonstrate an interpretation role in their reports issued immediately after earnings conference calls.

[17] Bushee, Gow and Taylor (2013) find that the lower readability of analyst statements indicates informativeness because it reflects analyst incentives to reveal information. Accordingly, we interpret the lower readability of analyst reports as an indication of as their effort to enrich the existing information provided by corporate disclosures.

6.3. Cross-sectional determinants of the analyst information discovery role

We conjecture that analysts may be more likely to play an information discovery role in cases where managers withhold value-relevant information from investors. In such cases, analysts are likely to respond to investors' demand for alternative sources of information and to supplement managers' limited disclosure with information obtained through their private research efforts. Prior literature on voluntary disclosure has identified several situations in which managers are more likely to withhold information. In our study, we examine the following determinants of voluntary disclosure.

Proprietary costs

Managers may choose to withhold information in cases where disclosure would reveal proprietary information about a firm's prospects to competitors. Numerous studies on the proprietary cost of disclosure find that such costs represent a significant consequence that prevents managers from being forthcoming (see reviews in Verrecchia 2001; Dye 2001; Healy and Palepu 2001). For example, managers may withhold information on research and development related to an innovative product or a new drug. In this case, analysts may choose to exert private research effort, such as communicating with the company's employees, researching the company's patent filing, investigating the company's suppliers, and attending company-hosted or industry conferences to collect value-relevant information that they then provide to investors. To test whether the analysts in our study increase their discovery role when managers face greater proprietary costs of disclosure, we follow Li, Lundholm, and Minnis (2013) and measure the proprietary cost of disclosure (denoted as $Competition$) as the percentage of competition references (i.e., occurrence of words related to competition) in the CC transcripts of

all firms in a given industry-year.[18] Li et al. (2013) argue that this measure reflects manager perceptions of competition and thus does not rely on industry boundaries or comprehensive identification of all sources of competition (e.g., competition from private firms, foreign firms, and potential new entrants).[19]

Litigation risk

Another factor that has been shown to impact disclosure is the amount of litigation risk faced by a firm (Frost and Pownall 1994; Johnson, Kasznik, and Nelson 2001; Baginski, Hassell, and Kimbrough 2002). For example, Rogers and Van Buskirk (2009) find that, despite the protection of the Safe Harbor provision of the 1995 Private Securities Litigation Reform Act, firms that have been subject to disclosure-related shareholder lawsuits are more wary about providing information to investors. Consistent with the results in these studies, Hollander et al. (2010) find that managers are less likely to answer participant questions during earnings conference calls when litigation risk is high. Based on these findings, we predict that analysts should be more likely to respond to higher investor demand for additional information for firms facing higher litigation risk. We measure litigation risk through an indicator variable that identifies industries that are subject to a high risk of private securities class action lawsuits (denoted $LitigRisk$). This proxy is proposed by Francis, Philbrick and Schipper (1994a, 1994b), who show that firms in

[18] Following Li et al. (2013), we consider a number of competition references: "competition," "competitor," "competitive," "compete," and "competing." We include words with an "s" appended, and remove phrases that contain negation, such as "less competitive," and "few competitors." We also scale the number of counts by the total number of words in the document. Although Li et al. (2013) construct their measure using the MD&A section of 10-K filings, we capture managers' perceptions of competition from our CC. We examine 100 randomly selected competition references from our sample and find that they highly resemble the examples provided in Appendix A of Li et al. (2013).

[19] As Berger (2011) and Beyer, Cohen, Lys, and Walther (2010) point out, other measures of competition in the existing literature (e.g., industry concentration measures based on Compustat data or on U.S. Census data as in Ali et al. 2009) suffer from these limitations because they fail to capture competition from private firms, non-U.S. companies, or potential entrants. In addition, they are bound by industry definitions. These limitations may result in unreliable measures of product market competition.

biotechnology, computers, electronics, and retail are more likely to be sued than those in other industries.[20]

Volatile information environment

The third factor we examine that may impact firm disclosure is the volatility of a firm's information environment (Dye 1985 and 1988; Jung and Kwon 1988). The more volatile an environment is, the more difficult it is for investors to discern if managers are withholding potentially bad news (Dye 1985; Jung and Kwon 1988; Jorgensen and Kirschenheiter 2003; Hughes and Pae 2004). The volatile information environment also makes it more difficult for investors to evaluate the valuation implication of a disclosure; that is, managers prefer not to disclose private information when an information environment is uncertain as they do not know how investors will respond (Suijs 2007; Dye 1988; Dutta and Trueman 2002). Consistent with these arguments, Chen, Matsumoto, and Rajgopal (2011) show that firms that choose to stop providing earnings guidance cite a "lack of predictability /uncertainty" as the reason behind this decision. They further show that these firms experience an increase in stock volatility. Based on these findings, we predict that analysts will be more likely to provide new information when a firm's information environment is more volatile. To capture information environment volatility, we use two measures: earnings persistence ($Persistence$) and stock return volatility prior to the conference call ($StockVol$). When earnings are persistent, the underlying business model is more predictable and the implication of earnings for future value is less uncertain. Stock return volatility is based on the finding that firms with more volatile stock prices are less transparent

[20] Prior studies using this proxy for litigation risk include Kim and Skinner (2012), Brown and Tucker (2011), Jayaraman and Milbourn (2009), Matsumoto (2002), Johnson, Kasznik, and Nelson (2001), and Jones and Weingram (1996).

and face greater uncertainty with the impact of disclosure on the firm's market value (Kothari, Shu and Wysocki 2009; Chen et al. 2011).

Bad news

Finally, we examine how the existence of bad news impacts a firm's disclosure choices and analyst provision of new information. Theoretical models generally predict that disclosure increases with firm performance (e.g., Dye 1985; 1986; Jung and Kwon 1988; Verrecchia 1983). Thus, if a manager has bad news to deliver, the manager may choose to withhold this information due to career concerns, as such news may decrease the manager's human capital and reputation (Verrecchia 2001; Nagar 1999). Empirical studies generally support these theories (e.g., Lang and Lundholm 1993; Miller 2002; Schrand and Walther 2000; Chen, Matsumoto and Rajgopal 2011). Therefore, we expect that analysts will be more likely to provide new information for firms that deliver bad news during their conference calls. We measure bad news with an the indicator variable of whether a firm's earnings have missed the analyst consensus forecast (denoted as $Miss$).[21]

6.4. Cross-sectional determinants of the analyst information interpretation role

In addition to examining the determinants of the analyst discovery role, we examine when analysts may be more likely to play an interpretation role in their reports. Previous research has shown that earnings conference calls may entail high information processing costs if manager statements are unstructured, ambiguous, subjective, or qualitative (Frankel, Johnson, and Skinner

[21] We also use an alternative proxy of bad news based on managers' tone in their CC narrative, measured as the percentage of positive sentences less the percentage of negative sentences in the CC. This proxy is based on survey evidence in Graham, et al. (2005) that "if the company fails to meet the guided number, the tone of the conference call becomes negative. The focus shifts to talking about why the company was unable to meet the consensus estimate" as opposed to talking about the firm's future prospects. We classify the tone of each sentence in the CC or AR as positive, neutral, or negative, following Huang et al. (2014). The empirical results, using manager tone as our measure, are similar to those based on $Miss$; that is, we find that the analyst discovery role increases when managers deliver bad news (i.e., when their tone is more negative). We do not include both manager tone and $Miss$ in the regression because of the high correlation between them.

1999; Brochet, Naranjo, and Yu 2013). Prior research also documents that the demand for information from analysts increases when investor understanding of corporate disclosures requires high processing costs (Lehavy, Li, and Merkley 2011). Accordingly, we expect that analysts will be more likely to serve an interpretation role when the information disclosed during the conference call is difficult for investors to process.

To evaluate the cost required to process conference call information, we use four measures. First, we follow Loughran and McDonald (2013) and measure the percentage of uncertain words contained in a CC ($Uncertain$).[22] Specifically, when managers use words like "may," "assume," "possibly," and "approximately," it is difficult for investors to judge the quality of the information (see also Epstein and Schneider 2008). Consistent with this argument, Loughran and McDonald find that a greater number of uncertain words in Form S-1 filings increases the volatility in the valuation of the IPO. Second, we follow Huang et al. (2014) and measure the extent to which qualitative vocabulary is used to discuss firm performance in the CC ($Qualitative$).[23] We calculate $Qualitative$ as one minus the percentage of sentences that contain "$" or "%." Third, we follow Frankel et al. (2006) and measure the complexity of a firm's operations by measuring the number of firm segments ($\#Segments$). Firms with more complex operations are likely to provide more complex information during the CC. Fourth, we follow Bloomfield (2002, 2008) and Li (2008) who demonstrate that managers have an incentive to obfuscate unfavorable information and that bad news is inherently more difficult to describe and understand than good news, and measure whether the firm conveys bad news by using an

[22] The complete list of uncertain words is available at http://www3.nd.edu/~mcdonald/Word_Lists.html.
[23] Huang et al. (2014) demonstrate that qualitative and subjective language is harder to process relative to quantitative information.

indicator variable that measures whether a firm's actual earnings in the current period miss the analyst consensus forecast ($Miss$).

6.5. Control variables

In our cross-sectional tests, we control for several conference call, analyst report, and firm characteristics. First, it is possible that analyst involvement in the conference call Q&A may be affected by the amount of disclosure as well as by whether manager statements require clarification. To control for analyst involvement during the conference call, we include the number of analyst questions during the Q&A session ($\#Questions$, measured as the natural log of one plus the number of questions raised by analysts in the conference call's Q&A session). We expect analysts' information roles in the prompt reports immediately after the calls to decrease with their involvement during the calls. Next, it is possible that the analyst information role is related to the magnitude of the earnings news. To control for this possibility, we include the absolute value of the earnings surprise (ABS_EPS_Surp). In addition, Brown and Tucker (2011) find that measures based on cosine similarity are positively correlated with document length. To mitigate this possibility, we control for the length of the combined prompt analyst reports (AR_Length). We also control for a number of firm characteristics such as firm size ($Size$), growth opportunities (book-to-market ratio, $BtoM$) and analyst following (i.e., the number of analysts issuing reports within the [0, 1] window relative to the conference call date, $\#Analysts$) as these characteristics may impact a firm's information environment (Lang and Lundholm 1993). Finally, we include both year and industry fixed effects to control for any common effect across all firms in a year or in an industry; our estimated coefficients are based on standard errors clustered at the firm and year levels.

6.6. Descriptive statistics

Table 4 reports the descriptive statistics for the variables used in our cross-sectional analyses. The statistics in Table 4 show that the mean of $Competition$ is 0.068 words per hundred words in a CC, which is comparable to the sample mean of 0.058 in Li et al. (2013). That is, an average CC in our sample contains a median of four competition-related words. We also see that 28% of our sample observations are from firms in industries subject to high litigation risk. Further, we see that the firms in our sample have an average stock price volatility of 8.6% and an earnings persistence of 0.53.The mean value of $Miss$ indicates that 22.2% of our sample conference calls contain information about earnings that have missed the consensus forecast. Regarding the text in our sample, we find that the mean value for $Uncertain$ is 0.836 words per one hundred words in the CC; which corresponds to an average of around 72 uncertain words in a CC. As a benchmark, the mean value for $Uncertain$ reported in Loughran and McDonald (2013) for their sample of S-1 filings is 1.41. Our mean value for $Qualitative$ indicates that, on average, 80.7% of the sentences in our CC are qualitative. Regarding firm characteristics, we find that the mean number of business segments for our sample firms ($\#Segments$) is two (the natural log of which is 0.751). Among the control variables, our sample calls on average raise 26 analyst questions during the Q&A session (as evidenced by the median value of $\#Questions$ of 3.3). The mean (median) length of the combined prompt analyst reports (AR_Length) is 411 (366) sentences, reflected across an average of 9 reports ($\#Analysts$).

[Insert Table 4 here]

6.7. Regression results

Table 5 reports the regression results for our cross-sectional determinants of analyst information roles. The dependent variable in columns (1) and (3) is $Discovery$; the dependent

variable in columns (2) and (4) is $Interpret$. The results in column (1) are consistent with our prediction that analysts increase their information discovery role when managers have greater incentives to withhold relevant information during conference calls. First, we find that the coefficient estimate on the proprietary cost measures ($Competition$) is positive and significant at the 5% level, suggesting analysts increase their private information and research efforts for firms that operate in a highly competitive environment. Second, we find that the estimated coefficient on our measure of litigation risk ($LitigRisk$) is positive and significant at the 1% level. This result suggests that analysts increase their information discovery role for firms that face greater risk of future litigation. Third, we find a significant and positive coefficient on $StockVol$. This result indicates that analysts increase their discovery role for firms with more volatile information environments. However, interestingly, we do not find a significant coefficient for earnings persistence, our other proxy for volatile information environment. Lastly, we find a positive and significant (at the 5% level) coefficient for $Miss$. This result suggests that analysts engage in more information discovery when managers deliver bad news during a conference call. Overall, we interpret these results as supporting the prediction that analysts choose to increase their information discovery role under conditions that are conducive to management withholding information from investors.

[Insert Table 5 here]

We next examine our results for the determinants of the analyst interpretation role. Specifically, column (2) of Table 5 reports the estimation results from regressing $Interpret$ on its determinants and control variables. These results show that all four measures of information processing costs ($Uncertain$, $Qualitative$, $\#Segments$ and $Miss$) are significant at the 5% level or better in the predicted direction. These results support our prediction that analysts

increase their interpretation role when the respective conference call contains information that is more difficult for investors to process (that is, when managers' statements are more uncertain and qualitative), when firm operation complexity increases, and when managers deliver bad news in the conference call.

To examine whether the analyst information discovery and interpretation roles are driven by similar or different economic conditions, we next estimate our $Discovery$ and $Interpret$ regressions based on an alternative specification that includes the determinants of both information roles. These results are reported in columns (3) and (4) of Table 5. As we can see from the results in column (3), the coefficient for $Uncertain$ is significant and negative. This result suggests that analysts shift effort from information discovery to information interpretation when managers' disclosure is harder to process. The coefficient for $\#Segments$ is insignificant, and the coefficient for $Qualitative$ is significant and positive. The results in column (4) indicate that none of the determinants of $Discovery$ loads in the $Interpret$ regression. Overall, these results suggest that the analyst information discovery and interpretation roles measured in our setting capture distinct aspects of their efforts under different economic circumstances. One interesting finding of note is that the results for $\#Questions$ yield a significantly negative coefficient in both the $Discovery$ and $Interpret$ regressions (as seen in columns (3) and (4), respectively). We interpret this finding as an indication that analysts embark on their information roles during the Q&A session of the earnings conference calls by asking questions. This involvement in turn preempts the level of information discovery and interpretation they exhibit in their prompt reports. This finding is further consistent with the evidence shown by Matsumoto, Pronk, and Roelofsen (2011) that the information content of earnings conference calls increases with analyst involvement.

7. Investor responses to analyst information discovery and interpretation roles

By comparing the textual narratives provided by analysts in their prompt reports with those of managers during earnings conference calls, we find evidence that analysts serve both information discovery and interpretation roles immediately after the calls. In this section, we examine how investors respond to each of these roles. Specifically, we use the information contained in the conference calls and prompt reports, as well as other control variables, to explain the market reaction during [0, 1] relative to the earnings conference call date (where $CAR[0,1]$ is the cumulative market-adjusted return during [0, 1]).[24] Because CAR is directional, we follow Huang et al. (2014) and Davis, Ge, Matsumoto, and Zhang (2012) and use the tone of the narratives (i.e., the percentage of positive sentences less the percentage of negative sentences) contained in the analyst reports and in the earnings conference calls to explain CAR, after controlling for other information signals released contemporaneously. Specifically, we estimate the following regression:

$$\begin{aligned} CAR[0,1] = {} & \alpha_1 Tone\_Discovery + \alpha_2 Tone\_Discovery \times Discovery + \alpha_3 Discovery \\ & + \beta_1 Tone\_CC + \beta_2 Tone\_CC \times Interpret + \beta_3 Interpret + \gamma_1 EF\_Rev \\ & + \gamma_2 REC\_Rev + \gamma_3 TP\_Rev + \gamma_4 EPS\_Surp + \gamma_5 Miss + \gamma_6 Prior\_CAR \\ & + \gamma_7 Size + \gamma_8 BtoM + \gamma_9 \#Analysts + \sum_t \delta_t I_t + \varepsilon, \end{aligned} \tag{4}$$

where, $Tone\_Discovery$ is the favorableness of the analyst opinions contained in the discovery topics in the prompt reports and $Tone\_CC$ is the favorableness of the manager tone during the conference call. Our control variables are as follows: 1) other research outputs contained in the analyst reports, including the revision of stock recommendations (Rec_Rev), earnings forecasts

[24] This return window encompasses all earnings announcements, conference calls, and analyst reports in our sample. We conduct robustness checks using market reactions during longer windows of [-1, 1] and [-1, 2] to capture investor reaction to these information events more fully. Results are similar using these longer return windows.

(EF_Rev), and target prices (TP_Rev); 2) earnings news, including earnings surprises (EPS_Surp), a dummy variable indicating whether a firm's earnings have missed the most recent analyst consensus forecast ($Miss$), and recent news or events captured by the abnormal returns during the ten trading days prior to the report date ($Prior\_CAR$); and 3) firm characteristics including firm size ($Size$), book-to-market ratio ($BtoM$), number of analyst reports being considered ($\#Analysts$), and year fixed-effects (I_t). Standard errors are estimated with a two-way cluster control at the firm and year level.

In the above regression, we would expect α_1 to be positive if investors value the information discovery role, as the textual opinions contained in the discovery topics ($Tone\_Discovery$) should trigger incremental market reactions beyond other information signals released contemporaneously. Furthermore, we would expect the interaction term of $Tone\_Interpret \times Interpret$, β_2, to be positive if investors value the information interpretation role, as greater interpretation should trigger a more intense investor reaction.

[Insert Table 6 here]

Table 6 reports our regression results for equation (4). Column (1) presents the results of a baseline regression excluding the interaction terms of $Tone\_Discovery \times Discovery$ and $Tone\_CC \times Interpret$. These results show that market reacts to both the tone of the manager discussion in the conference call ($\widehat{\beta_1}$ is positive and statistically significant at the 0.01 level) and the tone of the discovery topics in the analyst prompt reports ($\widehat{\alpha_1}$ is positive and statistically significant at the 0.01 level), after controlling for other contemporaneous information signals. The positive and significant coefficient on $Tone\_Discovery$ supports the prediction that investors value the information discovery role.

Column (2) presents the results when we add the interaction term of $Tone\_Discovery \times Discovery$ to the baseline model. Here, we again find a positive and significant coefficient (at the 0.01 level), suggesting that investors react more to the analyst discovery tone when the amount of discovery topics increases. Column (3) presents the results when we add the interaction term of $Tone\_CC \times Interpret$ to the baseline model. Again we find a positive and significant coefficient (at the 0.1 level), indicating that investor reactions to manager discussions increase with the extent of analyst interpretation. Finally, column (4) reports the estimation results for Eq. (4). Here, we see that both coefficients on $Tone\_Discovery$ and $Tone\_CC \times Interpret$ are positive and significant (at the 0.1 level or better). Note that investor reactions to these information roles are incremental to their reactions to earnings news (i.e., EPS_Surp and $Miss$), other research outputs in the analyst reports (i.e., EF_Rev, Rec_Rev and TP_Rev), and other firm characteristics and controls that might explain market reactions (i.e., $Prior\_CAR$, $Size$, $BtoM$, and $\#Analysts$). Taken together, the results in Table 6 suggest investors value both the information discovery and information interpretation roles.

8. Conclusion

In this study, we examine the information content embedded in the text of analyst reports issued immediately after earnings conference calls to understand the role analysts play in discovering and interpreting information for investors. To do so, we use algorithmic analyses of the topics discussed in the textual data of the conference calls and analyst reports to develop novel measures of the information content of this data. Using this methodology, we find that analyst reports issued promptly after earnings conference calls contain substantial amounts of discussion on exclusive topics not referred to in the conference calls. We also find that analyst discussions of conference call topics frequently entail different vocabulary than that used by

managers in their discussions. We interpret these two findings as support for analyst information discovery and interpretation roles, respectively.

We also extend our study and find that, cross-sectionally, analysts respond to investor demand for their services and play a greater information discovery role when managers have stronger incentives to withhold information during their conference calls; that is, when firms have greater proprietary costs, higher litigation risk, a more volatile information environment, or bad news. Our findings also show that analysts provide more information interpretation when the information processing costs are high (calls with more ambiguous and uncertain language), when firms operate in more complex environments, and when firms have greater information complexity related to bad news. Finally, through an examination of market reactions, we show that investors value both the information discovery role and the information interpretation role.

Our study advances the literature by contributing to our understanding of the different information roles that analysts play as well as the determinants of these roles. It does so by explicitly quantifying the thematic content of analyst research reports and contrasting it with the manager discussions during earnings conference calls. Additionally, we contribute to the literature by introducing a methodology for examining the information content of textual disclosures that does not rely on equity market reactions to the release of these disclosures. This methodology can mitigate the potential effect of confounding events when using measures based on market reactions. Finally, our study provides greater insight into how to use topic modeling to significantly expand the application of textual analysis to corporate financial disclosures beyond an understanding of "*how* texts are being said" to a broader understanding of "*what* is being said" in these texts.

References

Alexander, J., 1991. Do the merits matter? A study of settlements in securities class actions. *Stanford Law Review* 3, 497-598.

Ali, A., Klasa, S., Yeung, E., 2009. The limitations of industry concentration measures constructed with Compustat data: Implications for finance research. *Review of Financial Studies* 22, 3839-3871.

Asquith, P., Mikhail, M.B., Au, A.S., 2005. Information content of equity analyst reports. *Journal of Financial Economics* 75, 245-282.

Atkins, D. C., Rubin, T. N., Steyvers, M., Doeden, M. A., Baucom, B. R., Christensen, A., 2012. Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology* 26(5), 816-827.

Ball, C., Hoberg, G., Maksimovic, V., 2013. Disclosure and firm separation: A Text-based examination. Working paper, Maryland University.

Baginski, S. P., Hassell, J. M., Kimbrough, M. D., 2002. The effect of legal environment on voluntary disclosure: Evidence from management earnings forecasts issued in US and Canadian markets. *The Accounting Review* 77 (1), 25-50.

Bao, Y., Datta, A., 2012. Summarization of corporate risk factor disclosure through topic modeling. Proceeding of International Conference on Information Systems.

Barron, O.E., Byard, D., Kim, O., 2002. Changes in analysts' information around earnings announcements. *The Accounting Review* 77, 821-846.

Berger, P.G., 2011. Challenges and opportunities in disclosure research—A discussion of 'the financial reporting environment: Review of the recent literature'. *Journal of Accounting & Economics* 51, 204-218.

Beyer, A., Cohen, D.A., Lys, T.Z., Walther, B.R., 2010. The financial reporting environment: Review of the recent literature. *Journal of Accounting & Economics* 50, 296-343.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993-1022.

Blei, D., 2012. Probabilistic topic models. *Communications of ACM* 55(4), 77-84.

Bloomfield, R. J., 2002. The "incomplete revelation hypothesis" and financial reporting. *Accounting Horizons* 16 (3), 233-243.

Bloomfield, R., 2008. Discussion of "annual report readability, current earnings, and earnings persistence". *Journal of Accounting & Economics* 45 (2), 248-252.

Bradshaw, M. T., 2011. Analysts' forecasts: what do we know after decades of work?. Working paper. Boston College

Brochet, F., Naranjo, P.L., Yu, G., 2013. Capital market consequences of linguistic complexity in conference calls of non-US Firms. Working paper, Harvard Business School.

Brown, L. D., Call, A. C., Clement, M. B., Sharp, N. Y., 2014. Inside the "black box" of sell-side financial analysts. *Journal of Accounting Research* 53 (1), 1-47.

Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research* 49, 309-346.

Bushee, B. J., Core, J. E., Guay, W., Hamm, S. J., 2010. The role of the business press as an information intermediary. *Journal of Accounting Research* 48 (1), 1-19.

Bushee, B. J., Gow, I. D., Taylor, D. J., 2014. Linguistic Complexity in Firm Disclosures: Obfuscation or Information?. Working paper. University of Pennsylvania?

Chen, X., Cheng, Q., Lo, K., 2010. On the relationship between analyst reports and corporate disclosures: Exploring the roles of information discovery and interpretation. *Journal of Accounting & Economics* 49, 206-226.

Chen, S., Matsumoto, D., Rajgopal, S., 2011. Is silence golden? An empirical analysis of firms that stop giving quarterly earnings guidance. *Journal of Accounting & Economics* 51 (1), 134-150.

Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L., 2012. The effect of managerial "style" on the tone of earnings conference calls. In CAAA Annual Conference 2012.

De Franco, G., Hope, O. K., Vyas, D., Zhou, Y., 2011. Ambiguous language in analyst reports. Working paper, University of Toronto.

Dempsey, S.J., 1989. Predisclosure information search incentives, analyst following, and earnings announcement price response. *The Accounting Review*, 748-757.

Dutta, S., Trueman, B., 2002. The interpretation of information and corporate disclosure strategies. *Review of Accounting Studies* 7 (1), 75-96.

Dye, R. A., 1985. Disclosure of nonproprietary information. *Journal of Accounting Research*, 123-145.

Dye, R. A., 1988. Earnings management in an overlapping generations model. *Journal of Accounting Research*, 195-235.

Dye, R.A., 2001. An evaluation of "essays on disclosure" and the disclosure literature in accounting. *Journal of Accounting & Economics* 32, 181-235.

Epstein, L.G., Schneider, M., 2008. Ambiguity, information quality, and asset pricing. *The Journal of Finance* 63, 197-228.

Field, L., Lowry, M., Shu, S., 2005. Does disclosure deter or trigger litigation?. *Journal of Accounting and Economics* 39 (3), 487-507.

Francis, J., Schipper, K., Vincent, L., 2002. Earnings announcements and competing information. *Journal of Accounting & Economics* 33, 313-342.

Frankel, R., Johnson, M., Skinner, D.J., 1999. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research* 133-150.

Frankel, R., Kothari, S.P., Weber, J., 2006. Determinants of the informativeness of analyst research. *Journal of Accounting & Economics* 41, 29-54.

Frost, C. A., Pownall, G., 1994. Accounting disclosure practices in the United States and the United Kingdom. *Journal of Accounting Research*, 75-102.

Graham, J. R., Harvey, C. R., & Rajgopal, S., 2005. The economic implications of corporate financial reporting. *Journal of Accounting & Economics* 40 (1), 3-73.

Green, T. C., Jame, R., Markov, S., & Subasi, M., 2014. Access to management and the informativeness of analyst research. *Journal of Financial Economics* 114 (2), 239-255.

Griffiths, T., Steyvers, M., 2004. Finding scientific topics. Proceedings of the National Academy of Sciences 101, 5228-5235.

Grimmer, J., 2010. Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis* 18,1-35.

Hanley, K.W., Hoberg, G., 2010. The information content of IPO prospectuses. *Review of Financial Studies* 23, 2821-2864.

Healy, P.M., Palepu, K.G., 2001. Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting & Economics* 31, 405-440.

Hollander, S., Pronk, M., Roelofsen, E., 2010. Does silence speak? An empirical analysis of disclosure choices during conference calls. *Journal of Accounting Research* 48 (3), 531-563.

Huang, A., Zang, A., Zheng, R., 2014. Evidence on the information content of text in analyst reports. *The Accounting Review*, 89 (6), 2151-2180

Hughes, J. S., Pae, S., 2004. Voluntary disclosure of precision information. *Journal of Accounting and Economics* 37 (2), 261-289.

Ivković, Z., Jegadeesh, N., 2004. The timing and value of forecast and recommendation revisions. *Journal of Financial Economics* 73, 433-463.

Jayaraman, S., Milbourn, T. T., 2011. The role of stock liquidity in executive compensation. *The Accounting Review* 87 (2), 537-563.

Johnson, M. F., Kasznik, R., Nelson, K. K., 2001. The impact of securities litigation reform on the disclosure of Forward-Looking information by high technology firms. *Journal of Accounting Research* 39 (2), 297-327.

Jones, C. L., Weingram, S. E., 1996. The determinants of 10b-5 litigation risk. Working paper. George Washington University.

Jorgensen, B. N., & Kirschenheiter, M. T., 2003. Discretionary risk disclosures. *The Accounting Review* 78 (2), 449-469.

Jung, W. O., Kwon, Y. K., 1988. Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 146-153.

Kaplan, S., Vakili, K., 2013. Studying breakthrough innovations using topic modeling: A test using nanotechnology patents. Working paper, University of Toronto.

Kim, O., Verrecchia, R.E., 1994. Market liquidity and volume around earnings announcements. *Journal of Accounting & Economics* 17, 41-67.

Kim, O., Verrecchia, R.E., 1997. Pre-announcement and event-period private information. *Journal of Accounting & Economics* 24, 395-419.

Kothari, S., Li, X., Short, J.E., 2009. The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: A study using content analysis. *The Accounting Review* 84, 1639-1670.

Kothari, S. P., Shu, S.,Wysocki, P. D., 2009. Do managers withhold bad news?. *Journal of Accounting Research* 47 (1), 241-276.

Lang, M., Lundholm, R., 1993. Cross-sectional determinants of analyst ratings of corporate disclosures. *Journal of Accounting Research*, 246-271.

Lang, M., Stice-Lawrence, L., 2014. Textual analysis and international financial reporting: Large sample evidence. Working paper, The University of North Carolina at Chapel Hill

Lehavy, R., Li, F., Merkley, K., 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review* 86, 1087-1115.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. *Journal of Accounting & Economics* 45 (2), 221-247.

Li, F., Lundholm, R., Minnis, M., 2013. A measure of competition based on 10-K filings. *Journal of Accounting Research* 51, 399-436.

Livnat, J., Zhang, Y., 2012. Information interpretation or information discovery: Which role of analysts do investors value more? *Review of Accounting Studies* 17, 612-641.

Loh, R. K., Stulz, R. M., 2011. When are analyst recommendation changes influential?. *Review of Financial Studies* 24 (2), 593-627.

Loughran, T., McDonald, B., 2013. IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics* 109, 307-326.

Matsumoto, D. A., 2002. Management's incentives to avoid negative earnings surprises. *The Accounting Review* 77 (3), 483-514.

Matsumoto, D., Pronk, M., Roelofsen, E., 2011. What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review* 86, 1383-1414.

Mayew, W.J., 2008. Evidence of management discrimination among analysts during earnings conference calls. *Journal of Accounting Research* 46 (3), 627-659.

Mayew, W.J., Sharp, N.Y., Venkatachalam, M., 2013. Using earnings conference calls to identify analysts with superior private information. *Review of Accounting Studies* 18 (2), 386-413.

Miller, G. S., 2002. Earnings performance and discretionary disclosure. *Journal of Accounting Research* 40 (1), 173-204.

Mozes, H., 2003. Accuracy, usefulness and the evaluation of analysts' forecasts. International Journal of Forecasting, 19(3), 417-434.

Nagar, V., 1999. The role of the manager's human capital in discretionary disclosure. *Journal of Accounting Research*, 167-181.

Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R., 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54, 209-228.

Ramnath, S., Rock, S., Shane, P., 2008. The financial analyst forecast literature: A taxonomy with suggestions for future research. *International Journal of Forecasting* 24, 34-75.

Rogers, J. L., Van Buskirk, A., 2009. Shareholder litigation and changes in disclosure behavior. *Journal of Accounting and Economics* 47 (1), 136-156.

Schrand, C. M., & Walther, B. R., 2000. Strategic benchmarks in earnings announcements: The selective disclosure of prior-period earnings components. *The Accounting Review* 75 (2), 151-177.

Sheskin, D.J., 2011. *Handbook of Parametric and Nonparametric Statistical Procedures*, fifth ed. Chapman and Hall/CRC Press.

Shores, D., 1990. The association between interim information and security returns surrounding earnings announcements. *Journal of Accounting Research* 28, 164-181.

Singhal, A., 2001. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin* 24(4), 35-43.

Smith, J.E., Smith, N.P., 1971. Readability: a measure of the performance of the communication function of financial reporting. *The Accounting Review* 46, 552-561.

Soltes, E., 2014. Private interaction between firm management and sell-side analysts. *Journal of Accounting Research* 52 (1), 245-272.

Stickel, S. E., 1989. The timing of and incentives for annual earnings forecasts near interim earnings announcements. *The Accounting Review* 71, 289-315.

Suijs, J., 2007. Voluntary disclosure of information when firms are uncertain of investor response. *Journal of Accounting & Economics* 43 (2), 391-410.

Verrecchia, R.E., 1983. Discretionary disclosure. *Journal of Accounting & Economics* 5, 179-194.

Verrecchia, R.E., 2001. Essays on disclosure. *Journal of Accounting & Economics* 32, 97-180.

Appendix I

Additional Details on the Latent Dirichlet Allocation Model (LDA)

A. Intuition of LDA

We illustrate the intuition of LDA in Figure A1. Assume a collection of documents contains ten topics and each document has a different topic distribution. Further, each topic has a multinomial distribution over words. For example, the top four words in Topic 1 (Stores) in Figure A1 are: "new," "store," "open," and "square." Note that all the words in the vocabulary are associated with topics probabilistically. Top words are those with a high probability in a topic. A word can have high probabilities in multiple topics. For example, the word "new" has high probabilities in Topic 1 (Stores), 5 (Management) and 7 (Growth and Expansion), indicating that it is highly (but not equally) related to these three topics. Some words in the sample document have no topic labels because they are either stop words (e.g., "a, " "the, " "that") or words with low topic probability. LDA assumes each document is generated in two steps. First, a topic is randomly drawn based on the assumed topic distribution of the document; next, a word is randomly drawn based on the word distribution of the topic. Repeating this two-step word selection procedure for each word generates the complete document.

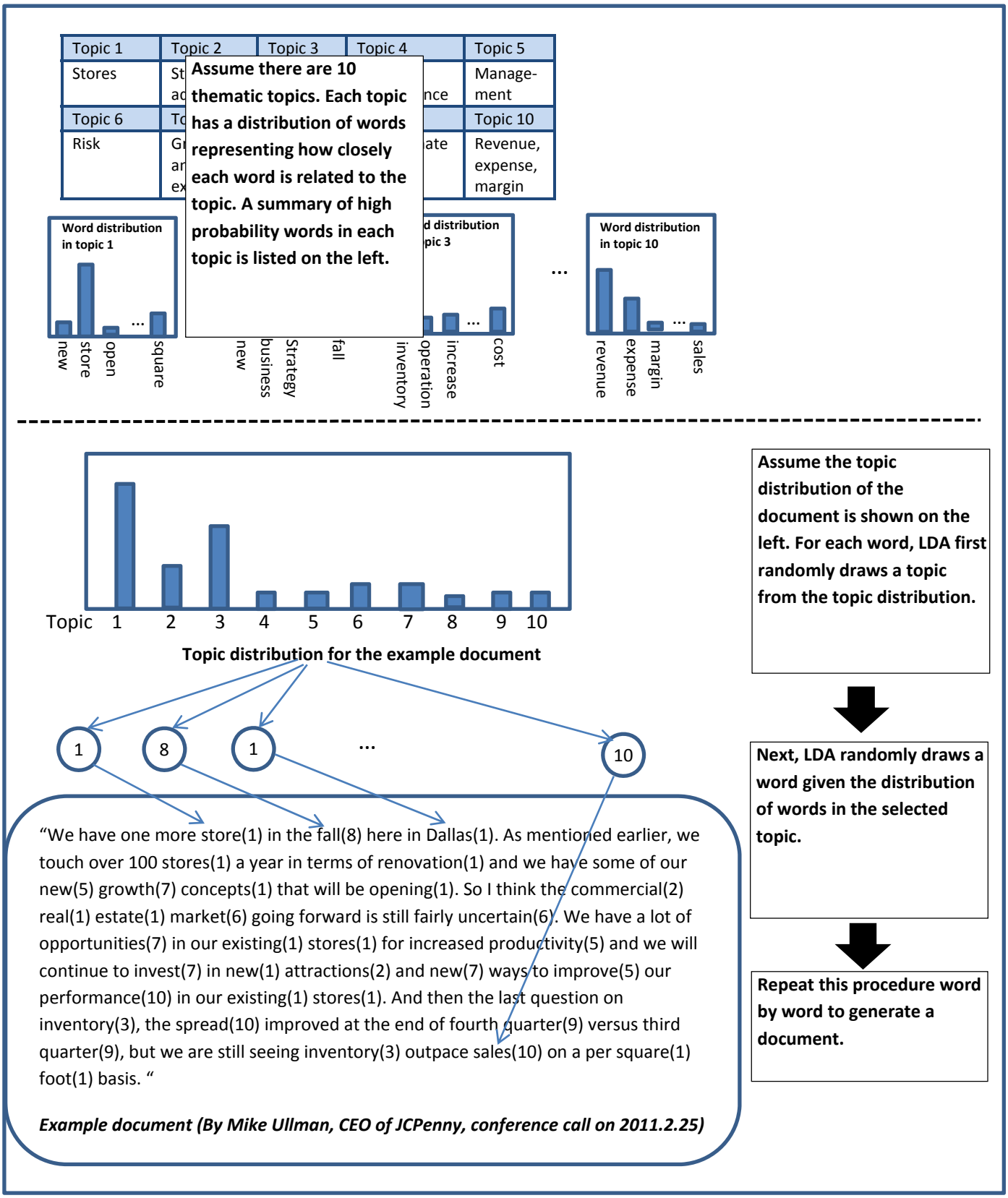


Figure A1. An illustration of how a document is generated

B. Technical Details of the Latent Dirichlet Allocation Model

Assume a corpus *D* consisting of a collection of documents contains a fixed number of latent topics. Each document, d, is characterized by a discrete probability distribution over topics (θ_d), and each topic, t, is characterized by a discrete probability distribution over words (ϕ_t). Given this framework, a document, *d*, can be generated by repeatedly sampling on the topic distribution θ_d to draw a topic, followed by a sampling on the word distribution ϕ_t for the given topic to draw a word. Formally, the LDA model generates the n^{th} word appearing in document d, w_{dn} , based on the following process:

1. Choose a topic z_{dn} ~ Multinomial (θ_d).
2. Choose a word w_{dn} ~ from $p(w_{dn}|z_{dn}, \phi_{z_{dn}})$,

where θ_d is the document d probability vector of topics, and $\phi_{z_{dn}}$is the word probability vector for topic z_{dn}. Topics $\{z_{dn}\}$ and words $\{w_{dn}\}$ are discrete random variables, and both follow a multinomial distribution. The objective of LDA is to estimate the parameters $\{\theta_d\}$ and $\{\phi_t\}$.

To simplify the computations and obtain the desired concentration of topics in a document, the model assumes that the multinomial topic and word posterior distributions are Dirichlet distributions with known parameters, i.e., $p(\theta_d)$~Dirichlet(α), $p(\phi_t)$~Dirichlet(β). We follow the literature (Steyvers and Griffiths, 2006) and use constant values of 0.1 and 0.01 for α and β, respectively.

Given this framework, the probabilistic generative process can be conveniently illustrated using a plate notation (Buntine, 1994). Figure A2 shows the graphical model of LDA used in Blei et al., 2003. Arrows indicate conditional dependencies between variables, while plates (the boxes in the figure) refer to repetitions of sampling steps with the variable in the lower right corner referring to the number of samples. For example, the inner plate over *z* and *w* illustrates the repeated sampling of topics and words until N_d words have been generated for document *d;* the plate surrounding θ_d illustrates the sampling of a distribution over topics for each document *d* for a total of *D* documents; the plate surrounding ϕ_t illustrates the repeated sampling of word distributions for each topic *z* until the word probabilities of *T* topics have been generated. LDA assumes that α and β are known parameters. The words (w_{dn}) are observed by LDA. The variables ϕ_t and θ_d, as well as z_{dn} (the assignment of word to topics) are the three sets of latent variables that the LDA intends to estimate.

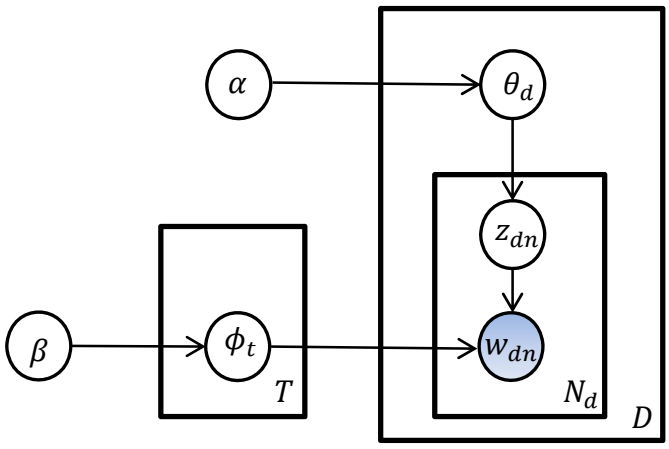


Figure A2. Plate notation depiction of LDA

The estimation problem of LDA is to compute the posterior distribution of the latent variables (i.e., ϕ_t, θ_d, and z_{dn}) given the observed documents and the assumed parameters (α, β, and T). However, these distributions are intractable to compute in general (Blei et al., 2003). The most commonly used estimation algorithm for LDA is collapsed Gibbs sampling, as proposed in Griffiths and Steyvers (2006). The collapsed Gibbs sampling procedure starts with sampling the value of variable z_{dn}. The probability of a topic assignment, z_{dn}, conditional on all other assignments z_{-dn} and other model parameters is equal to:

$$p(z_{dn} = t | w_{dn} = m, z_{-dn}, \alpha, \beta) \propto \frac{C^{WT}_{mt,-dn} + \beta}{\sum_{m'} C^{WT}_{m't,-dn} + W\beta} \times \frac{C^{T}_{t,-dn} + \alpha}{\sum_{t'} C^{T}_{t',-dn} + T\alpha}, \qquad \text{(A1)}$$

where z_{dn} is the topic assignment of the n^{th} word appearing in document d; z_{-dn} is the topic assignments of all words other than the n^{th} word appearing in document d; $C^{WT}_{mt,-dn}$and $C^{T}_{t,-dn}$ are the count matrices of the word-topic assignment of all words in document d other than the current word z_{dn}. The right hand side of (A1) is the posterior conditional probability of word *m* given the topic *t* multiplied by the probability of the topic *t*, i.e., $p(t|w) \propto p(w|t)p(t)$. See Blei et al. (2003) and Steyvers and Griffiths (2006) for more details.

Equation (A1) provides direct estimates of z_{dn}. However, many applications of topic modeling require the estimates of the word-topic distributions (ϕ_t) and topic-document distribution (θ_d). These distributions can be directly calculated from the count matrices as follows:

$$\phi_t = \frac{C^{WT}_{mt,-dn} + \beta}{\sum_{m'} C^{WT}_{m't,-dn} + W\beta}, \qquad \theta_d = \frac{C^{T}_{t,-dn} + \alpha}{\sum_{t'} C^{T}_{t',-dn} + T\alpha}.$$

C. Applying LDA to Conference Call Transcripts and Analyst Reports

Our corpus is composed of 18,607 earnings conference call transcripts and 476,633 analyst reports for S&P 500 firms from 2003-2012. We incorporate all available reports in the LDA to obtain the best representation of topics discussed in these reports. Earnings conference call transcripts are obtained from Thomson Reuter's Streetevent database and analyst reports are obtained from Thomson Reuter's Investext database. We conduct the LDA analysis by industry because many topics are likely industry-specific. We use the Global Industry Classification Standard (GICS) obtained from Compustat to identify industries. This classification is widely adopted by brokerages and analysts as their industry classification system and is superior to other industry classification schemes in identifying firms with their industry peers (Kadan, Madureira, Wang, and Zach, 2012; Boni and Womack, 2006; Bhojraj, Lee, and Oler, 2003)

Preprocessing of textual documents

We perform a set of standard preprocessing steps in information retrieval research on our dataset prior to the application of LDA. First, we convert all words into lower case and remove all non-English characters (e.g., punctuations and numbers). Second, we replace similar words that have the same root with a single representative word. This procedure is called "stemming" (Porter, 1980). For example, "increased" and "increases" are replaced by "increase." Last, we remove highly frequent functional words—also referred to as stop words. For example, "a," "of," and "the" are extremely frequent words, but convey relatively little meaning. These preprocessing steps help reduce the computational burden of the LDA model and enhance the interpretability of topics (Manning et al. 2008; Blei 2012). This process results in approximately 303 million words.

For analyst reports, we follow Huang, Zang and Zheng (2014) and remove the textual content in the tables, graphs, and "brokerage disclosures." Brokerage disclosures contain explanations of stock-rating system, disclosures regarding conflicts of interest, analyst certifications, disclosure required by regulations, disclaimers, glossaries, and descriptions of the brokerage or research firm. For conference calls, we exclude narratives from operators and standard greeting words used by speakers. For both analyst reports and conference calls, we remove companies' names and tickers to prevent the algorithm from identifying companies' names as topics.

Determining the number of topics

The LDA algorithm requires the researcher to input the number of topics in the documents. The choice of the number of topics can affect the interpretability of the results. For example, assuming too few topics can result in very broad topics and obscure specific topics. Conversely, assuming too many topics can introduce economically meaningless topics. To select the optimal number of topics, we follow the computational linguistic literature and calculate the *perplexity* of the LDA model based on different number of topics (Brown, Della Pietra, Mercer, and Della Pietra, 1992; Blei et al., 2003; Rosen-Zvi, Griffiths, Steyvers, and Smyth, 2004). Perplexity

measures the ability of an LDA model estimated on a subset of documents (training data) to predict word choices in the remaining documents (testing data). It is defined as the exponential of the negative normalized predictive likelihood under the model. Accordingly, the perplexity score is monotonically decreasing in the likelihood of observing the testing data given the model estimated from the training data. A lower perplexity score indicates better generalization performance of the model. Formally, for a testing data (D_{test}) with *M* documents, the perplexity is equal to:

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} logp(w_d)}{\sum_{d=1}^{M} N_d}\right\},$$

where N_d is the number of words in document d; w_d is a vector of all the words in document d; and $p(w_d)$ is the probability of observing the word vector w_d in document d given the LDA model estimated from the training data.

Following the literature (Blei et al., 2003; Rosen-Zvi et al., 2004), we compute and plot the perplexity of the LDA model for different numbers of topics ranging from 2 to 120. As can be seen in Figure A3, the perplexity score improves with the number of topics, but the improvement is marginally decreasing. The improvement diminishes significantly once the number of topics exceeds 60. Therefore, we choose 60 as the number of topics in our corpus.[1] This procedure is consistent with prior literature that uses LDA to analyze textual documents.[2]

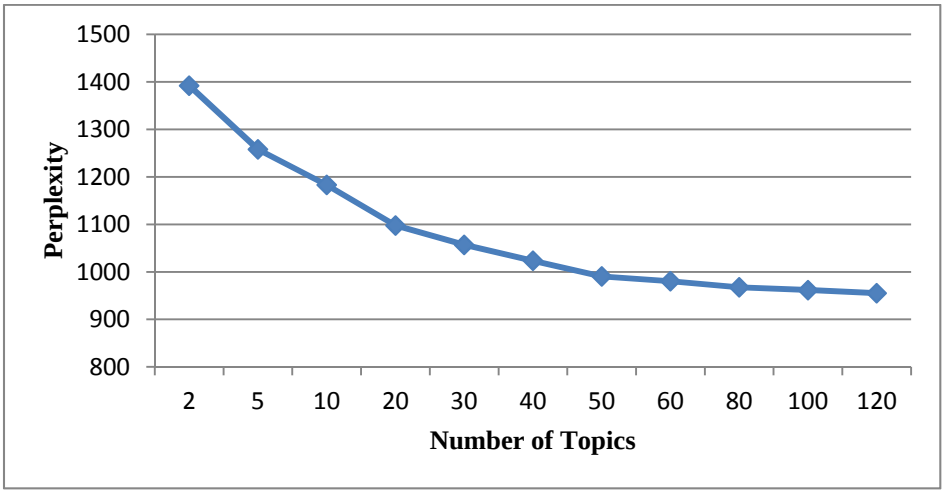


Figure A3. Perplexity of LDA model for different numbers of topics

Constructing a topic vector of a document

[1] We compare LDA results based on 30, 60, and 100 topics. Based on our comparison, we conclude that the LDA results with 60 topics outperform the other specifications in terms of its ability to identify intuitively important topics without generating many uninterpretable topics.

[2] For example, Ball et al. (2013) use 100 topics for MD&A text; Quinn et al. (2010) use 42 topics for political text; Atkins et al. (2012), use 100 topics for couples-therapy transcripts.

We use the following procedure to construct the topic vector (T_d) of a document d: first, we separate each sentence in a document into words; then, using the LDA output, which is a topic-word frequency matrix Φ, we construct a frequency vector for each word containing the number of times it appears in each of the K topics. With these vectors, we can construct a sentence-level matrix of word frequencies in each of the K topics (e.g., a sentence containing 10 words would have a $10 \times K$ frequency matrix). For each sentence, we then sum the frequencies of the words in each topic and assign the sentence to the topic with the highest combined frequency. Intuitively, we assume that a sentence containing words with the largest frequency in a given topic likely represents this topic.[3] The fraction of document d that is dedicated to a discussion of topic k (S_{dk}) equals the number of sentences that are assigned to the topic k divided by the total number of sentences in document d. Formally,

$$Topic\ vector\ of\ document\ d = T_d = (S_{d1}, S_{d2}, \ldots S_{d60}), \quad (1)$$

where S_{dk} represents the fraction of the discussion in a document devoted to topic k.

[3] In a sensitivity test, we assign each sentence into three topics based on the three highest combined frequencies. Our empirical results remain qualitatively similar.

Appendix II

Examples of analysts' information discovery and interpretation from excerpts of earnings conference calls and analyst reports

This appendix provides examples of topics that were classified by the LDA algorithm as information discovery and interpretation. We classify analyst discussions of exclusive topics that receive little or no attention from managers during the earnings conference calls as information discovery. We classify discussions in analyst reports of the top ten topics in the CC as interpretation.

Examples of Information Discovery

Example 1

On January 18th, 2006, Apple Inc. held a conference call to discuss the results of the period ending December 31, 2005. The LDA model identifies a topic labeled as "segment profit margin" discussed in several analyst reports issued immediately after the conference call. Because managers only briefly mention this topic in their call, analyst discussions of it is classified as information discovery. Below are excerpts from analyst discussions of this topic:

Christopher Kinney Whitmore (Analyst, Deutsche Bank):

"We believe Apple's PC margins are likely in the 28-30% range, above overall corporate gross margins, suggesting that additional uptake of Macs could drive EPS upside in coming quarters."

Bill Shope (Analyst, JPMorgan):

"… we believe that iPod gross margins are now trending above 26%."

Tsvetan Knitisheff (Analyst, Kintisheff Research):

"…since Intel-based Mac PC-s have the same pricing as previous models, we expect improved gross margins in the PC business of Apple as the cost base of Intel-based computers is expected to be lower compared to Power PC based PC-s. Instead, the fact that iPod revenues exceeded Mac PC sales for the first time, combined with still-low (albeit improving) margins in the iPod resulted in 27.3% corporate gross margin excluding stock compensation expenses (SCE)."

Peter Oppenheimer, Apple's CFO, provides a brief statement in regards to this topic:

"I don't want to be, for competitive reasons, specific, nor do we want to talk about specific iPod sales, but I will tell that you that the iPod gross margins in the December quarter were above 20%...As regards to the Intel-based Mac gross margins, we don't want to provide specific gross margins for any of our products.

As can be seen, the analysts provide specific, new information regarding the level and implications of Apple's Ipod and PC gross margin relative to the CFO discussion in the CC. We view this example as information discovery by the analysts.

Example 2

Another example of analyst information discovery role is related to a topic labeled as "Acquisitions," excerpted from Google's earnings conference call held on October 13, 2011. Management discussion includes very little in regards to its acquisitions. Analysts provide incremental information regarding the timing, motivation, and potential operating and financial implications of the Motorola acquisition. Accordingly, based on the classification of the LDA model and as can be seen from the excerpts below, we classify analyst discussion of this topic as information discovery.

Patrick Pichette, (CFO, Google Inc.):
"Additionally, acquisitions this quarter added a large number of people as well."

Larry Page, (CEO, Google Inc.):
"And as you know, Motorola Deals is under review and I think it will be premature for us to comment about anything we might do with regards to that."

Excerpts of analyst reports issued on the following day contain the additional information regarding this topic:

Jeetil Patel (Analyst, Deutsche Bank Research):

"The company still expects the Motorola deal to close early in 2012.

We suspect this may be part of the motivation behind the Motorola Mobility acquisition - the need to own more of the stack to control more of the search economics in addition to content and applications.

As such, we think that Google is strategically (and perhaps defensively) positioning itself similarly with the Motorola buy, whereas in the near term the core advertising business is executing exceptionally well.

As such, we see where Motorola fits in for Google, but we are hoping for quick deployment of Google-Mot handsets post deal-close, which should enable it to innovate from an application/functionality & ultimately ad standpoint.

Mayuresh Masurekar (Analyst, Collins Stewart):

Reiterate BUY, on global online advt growth, accelerating mobile revenues, incremental display, Android optionality and inexpensive valuation at 12x 2012 PF EPS ex cash even after Motorola acquisitions.

Nick Landell-Mills (Analyst, Indigo Equity Research):

In 2006, YouTube ($1.7 bn) & Postini were acquired.

Google pays the 3rd party websites fees for this; referred to as The majority of TAC (Total Acquisition Costs).

This acquisition places Google to compete with some of its partners, the handset makers who use Android.

Ben Schachter (Analyst, Macquarie Research):

Other than indicating that it plans to support and protect its Android ecosystem (presumably via patent acquisition and litigation), we expect GOOG will remain quiet on its broader MOT strategy until the deal closes.

Examples of Information Interpretation

Example 1

Recall that our empirical proxy for the amount of analyst information interpretation is equal to one minus the cosine similarity between the word usages by analysts and managers (bounded between [0,1]). We label this proxy as $Interpret$. Below we provide two examples for low and high values of $Interpret$. The first example is taken from the Applied Materials Inc. earnings conference call held on May 16, 2006. Management discusses a topic labeled as "product order." Analyst discussions of this topic are associated with a relatively low value of $Interpret$ at 0.141, suggesting a low level of interpretation by the analysts. As can be seen from the included excerpts, the analyst discussion resembles that of management and provides very little processing of the information provided by management:

Nancy Handel (Senior VP and CFO):

Orders by major geographic areas were Korea, 22%; Taiwan, 19%; North America, 18%; Japan, 17%; Southeast Asia and China, 14%; and Europe, 10%.

In the quarter, DRAM orders represented 27% of silicon systems orders, flash memory orders were 24% and foundry orders were 17%. Logic and other orders comprised the remaining 32%.

300 millimeter orders represented approximately 73% of total systems orders, and 74% of the system orders were for 100 nm and below process technology.

The following excerpts from analyst reports issued on the same date or the following date contain analysts' discussions of the same topic:

Shekhar Pramanick (Analyst, Moors and Cabot, Inc):

Orders by geography were as follows: Korea 22%, Taiwan 19%, North America 18%, Japan 17%, SE Asia/China 14%, and Europe 10%.

Orders by segment were as follows: DRAM 27%, Flash 24%, foundry 17% and logic/other 32%.

300mm orders represented 73% of total system orders and 74% of system orders was for the 100nm technology node and below.

Robert Maire (Analyst, Needham & Company):

Geographic Order Breakdown: The distribution of orders was as follows; Korea 22%, Taiwan 19%, North America 18%, Japan 17%, Southeast Asia and China 14%, and Europe 10%.

The Memory Monster - In the quarter, DRAM orders represented 27% of silicon systems orders, flash memory orders were 24% up from 18% in the first quarter and foundry orders were 17% down slightly from 19%.

300-mm orders represented about 73% of total systems orders and 74% of systems orders were for 100-nm and below process technologies.

Gavin X. Duffy (Analyst, A.G. Edwards & Sons, Inc):

Taiwan represented 19% of new orders in Q2, North America was 18%, Japan 17%, Europe 10%, Korea 22%, and Southeast Asia and China represented 14%.

DRAM orders represented 27% of total Q2 system orders (versus 28% in Q1), flash was 24%, (versus 18% sequentially), foundries accounted for 17% (versus 19% sequentially) and logic and other revenues accounted for the remaining 32% versus 35% in Q1.

300-millimeter tools represented approximately 73% of total system orders received in the quarter versus 84% sequentially.

Jay Deahna (Analyst, JPMorgan):

Logic was the greatest proportion of orders at 32%, with DRAM at 27%, flash at 24%, and foundry at 17%.

R. Kukreja (Analyst, W.R. Hambrecht & Co.):

On a more granular level, DRAM accounted for 27% of the total system orders (28% in FQ1), logic contributed 32% (35% in FQ1), foundries added 17% (19% in FQ1) while flash, which grew the most sequentially at 63% over FQ1, made up the remaining 24% (18% in FQ1).

Tim Summers (Analyst, Stanford Financial Group):

300mm systems accounted for 73% of total systems orders, lower than the 84% in 1Q06.

Example 2

Our second example is associated with a high value of $Interpret$ of 0.855 related to analyst interpretation of the *EZstore Initiative* discussion in the Dollar General Corporation's management earnings conference call from May 26, 2005. This discussion is part of a topic labeled "store operation" which was discussed in details in both the conference call and analyst reports. As can be seen, the analysts provide additional context, details, and opinion relative to management discussion of this topic.

David Perdue (Chairman and CEO, Dollar General):

Over 1200 stores served out of three distribution centers have been converted to the EZstore process. We are convinced that our EZstore effort will enhance our ability to manage our ever increasing number of small stores. While EZstore changes the way we replenish our stores, it also has a dramatic impact on management effectiveness at the store level. It is still our plan to have EZstore in about half of our stores by the end of fiscal '05. Improving our processes and execution of the stores remains our top priority.

Excerpts of analyst reports issued on the following day contain the following discussions of this topic:

Dan Wewer (Analyst, CIBC World Markets Inc.):

As a reminder, EZStore is a workflow initiative that simplifies DG's store operations by changing the way it pick, packs, and ships inventory in the distribution center.

Ralph Jean (Analyst, Wells Fargo Securities):

A key part of the EZstore initiative is the use of rolltainers that significantly reduces store labor costs associated with unloading delivery trucks.

Patrick McKeever (Analyst, Sun Trust Robinson Humphrey Capital Markets):

Before EZ Store, boxes were unloaded manually one by one and sorted in the back-room or elsewhere in the store. When the truck arrives at the store, the driver alone is responsible for rolling the container off the truck and into the back room. Employees then push the containers into designated areas of the store.

The EZ Store initiative, which has now been rolled out to more than 1,200 stores, or roughly 20% of the overall chain, is a process reengineering program that (in our opinion) revolutionizes the truck unloading process and has the potential to drive considerable efficiencies through what has been, until now, a labor intensive and generally inefficient process.

We believe EZ Store reduces the amount of time necessary to unload the truck from 12 hours to about an hour and a half.

Christine K. Augustine (Analyst, Bear, Stearns & Co., Inc.):

The benefits of EZ Store include lower turnover, lower costs to run a store, including lower workers' compensation costs, and fewer damages to merchandise.

Distribution center processes are also changing as a result of the EZ Store program. The EZ Store rollout has implications for hiring, training, scheduling, product presentation and product handling.

Mark Miller (Analyst, William Blair & Company):

The EZ Store initiative should facilitate improved better leverage of payroll going forward, although the timing and magnitude of that payback (relative to other cost pressures) is less clear.

John Zolidis (Analyst, Buckingham Research Group):

Finally, we expect the company's EZ Store initiative, which improves store operations and efficiency, should provide a benefit over the rest of the year.

Appendix III

Variable Definitions

Variable Name	Definition
Main variables	
$Discovery$	The number of sentences labelled by LDA as non-top-ten CC topics in AR scaled by the total number of sentences in AR. Top-ten CC topics are the ten topics with the most sentences labelled by LDA in CC;
$Interpret$	One minus the average within-topic cosine word similarity between CC and AR in the top-ten CC topics. The within-topic cosine word similarity between CC and AR for a given topic k is calculated as $\frac{\sum_{j=1}^{N}(w_{jk} \cdot v_{jk})}{\sqrt{\sum_{j=1}^{N}(w_{jk})^2} \cdot \sqrt{\sum_{j=1}^{N}(v_{jk})^2}}$, where, w_{1k} is word 1's frequency in the discussion of topic k in AR; v_{1k} is word 1's frequency in the discussion of topic k in CC; N is the total number of unique words in CC and AR. Top-ten CC topics are the ten topics with the most sentences labelled by LDA in CC;
Determinants of Discovery and Interpret	
$Competition$	Percentage of competition related words in CC in the industry during the 12 months prior to the conference call. Following Li et al. (2013), competition related words include "competition," "competitor," "competitive," "compete," and "competing." We include words with an "s" appended and remove phrases that contain negation, such as "less competitive," and "few competitors;"
$LitigRisk$	An indicator variable that equals one if a firm's SIC code belongs to the one of the four industries identified by Francis et al. (1994) to have a high incidence of litigation: Biotechnology (2833-2836, 8731-8734), Computers (3570-3577 and 7370-7374), Electronics (3600-3674), and Retailing (5200-5961), and zero otherwise;
$StockVol$	The standard deviation of the monthly return of the firm in the 12 months prior to the conference call, winsorized at the top and bottom 1%;
$Persistence$	Earnings persistence, defined as the slope coefficient from an autoregressive model of order one (AR1) for annual income before extraordinary items (*IB*), estimated over a ten-year rolling window prior to the conference call, winsorized at the top and bottom 1%;
$Miss$	An indicator variable that equals one if the actual EPS is less than the last consensus EPS forecast before the earnings announcement, both from I/B/E/S, and zero otherwise;
$Uncertain$	The number of words in CC that are in the Uncertainty word list created by Loughran and McDonald (2013), scaled by the total number of words in CC;

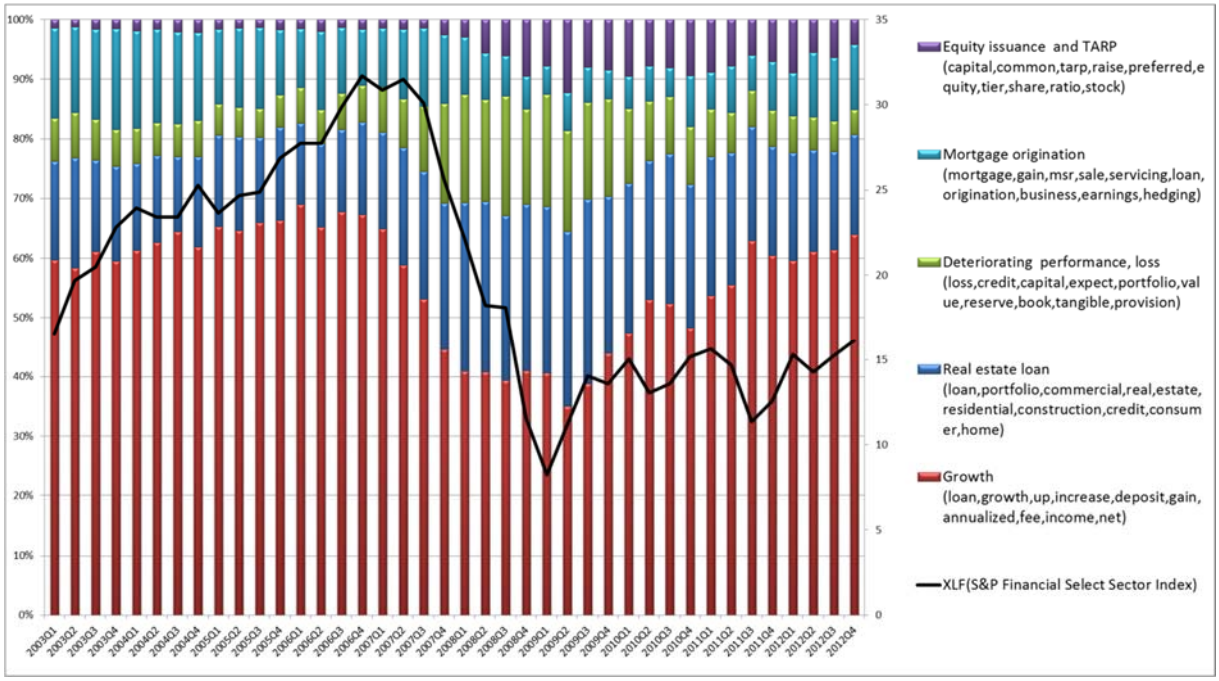
$Qualitative$	The number of sentences without a dollar sign or percent sign in CC scaled by the total number of sentence in CC;
$\#Segments$	The natural log of a firm's number of segments;
Control variables for Determinant Tests	
$\#Questions$	The natural log of one plus the number of questions raised by analysts in the conference call's Q&A session;
ABS_EPS_Surp	The absolute value of the earnings surprise, calculated as the absolute value of the difference between the actual EPS and the last consensus EPS forecast before the earnings announcement, both from I/B/E/S, scaled by the stock price 10 days prior to the earnings announcement, winsorized at the top 2%;
AR_Length	The number of sentences in analyst reports issued on the day of or the day following the conference call;
$Size$	The natural log of the market value of equity of the firm (*CSHOQ*×*PRCCQ*) at the end of the quarter prior to the conference call;
$BtoM$	The book value of equity (*CEQ*) scaled by the market value of equity (*CSHOQ*×*PRCCQ*) of the firm at the end of the quarter prior to the conference call, winsorized at the top and bottom 1%.
$\#Analyst$	The number of analyst reports issued on the day of or the day following the conference call;
Variables in the Market Reaction Test	
$CAR[0,1]$	The cumulative market-adjusted return over the [0, 1] window relative to the conference call date, winsorized at the top and bottom 1%, where the market-adjusted return is calculated as the raw return minus the buy-and-hold return on the NYSE/Amex/Nasdaq value-weighted market index;
$Tone\_Discovery$	The textual opinion of the sentences labelled by LDA as non-top-ten CC topics in AR. Top-ten CC topics are the ten topics with the most sentences labelled by LDA in CC. The textual opinion of the sentences is calculated as the percentage of positive sentences minus the percentage of negative sentences as classified by the naïve Bayes approach (Huang et al. 2014);
$Tone\_CC$	The textual opinion of the sentences labelled by LDA as top-ten CC topics in CC. Top-ten CC topics are the ten topics with the most sentences labelled by LDA in CC. The textual opinion of the sentences is calculated as the percentage of positive sentences minus the percentage of negative sentences as classified by the naïve Bayes approach (Huang et al. 2014);
EF_Rev	The consensus analyst earnings forecast for the next fiscal year immediately after the conference call minus that immediately before the conference call, scaled by the stock price of the firm 10 days prior to the conference call, winsorized at the top and bottom 1%;

Rec_Rev	The consensus analyst stock recommendation immediately after the conference call minus that immediately before the conference call. Analyst stock recommendations are coded as: 5 (Strong Buy), 4 (Buy), 3 (Hold), 2 (Underperform), and 1 (Sell);
TP_Rev	The consensus analyst target price immediately after the conference call minus that immediately before the conference call, scaled by the stock price of the firm 10 days prior to the conference call, winsorized at the top and bottom 1%;
EPS_Surp	Earnings surprise, calculated as the actual EPS minus the last consensus EPS forecast before the earnings announcement, both from I/B/E/S, scaled by the stock price 10 days prior to the earnings announcement, winsorized at the top and bottom 1%;
$Prior\_CAR$	The cumulative 10-day abnormal returns ending two days before the conference call winsorized at the top and bottom 1%, where abnormal return is calculated as the raw return minus the buy-and-hold return on the NYSE/Amex/Nasdaq value-weighted market index.

Figure 1
Temporal Variation in the Distribution of Key Topics

This figure presents the relative weights in the five topics with the highest variability in the banking and telecommunication industries, along with their respective sector indices (Financial Sector SPDR – XLF and iShares US Telecommunications – IYZ index respectively) in our sample period of 2003-2012.

Panel A: Banking industry (GICS 4010)



Panel B: Telecommunication industry (GICS 5010)

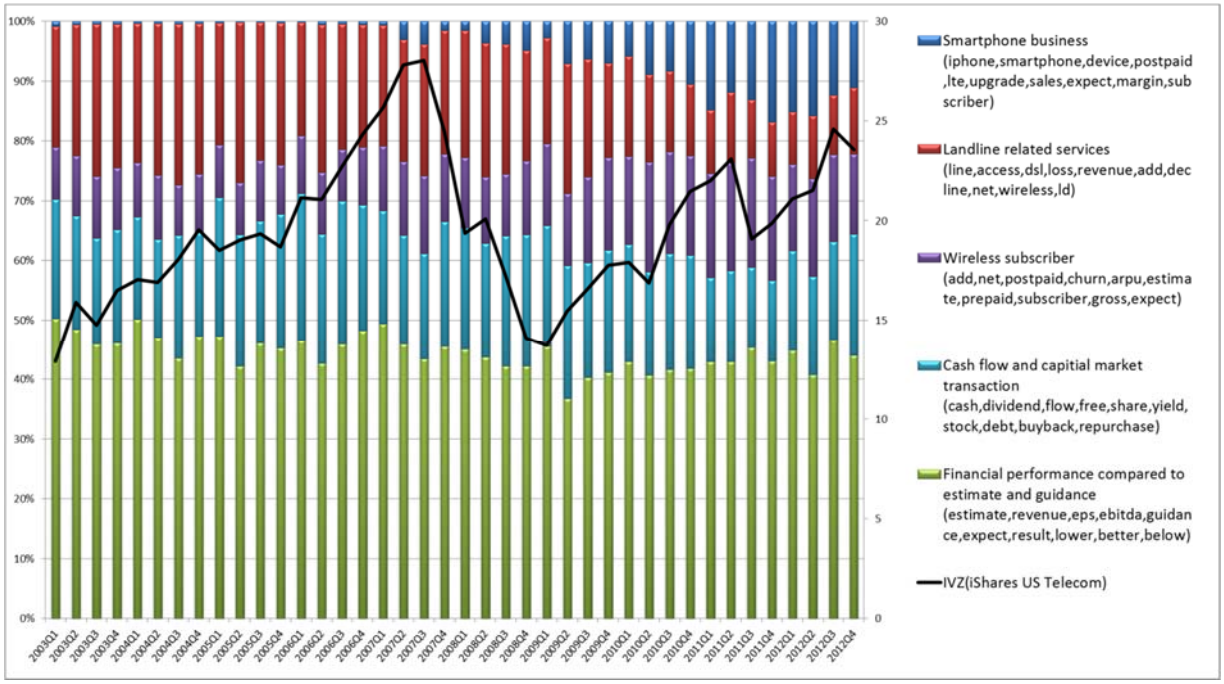


Table 1

Highest Probability Words in the Top Ten Topics of the Five Largest Industries

This table reports the top 20 words in each of the top ten topics and our inferred topic labels for the five largest industries in terms of the total number of conference calls in our sample.

Topic label	Top 20 Words
Capital Goods (GICS 2010)	
Comparing financial performance with expectation	margin, estimate, guidance, EPS, expect, consensus, operating, revenue, lower, bps, sales, expectation, below, higher, segment, management, forecast, up, beat, outlook
Sales	sales, increase, operating, margin, up, profit, higher, estimate, decline, volume, share, segment, result, improve, offset, rise, lower, cost, currency, earnings
Growth opportunities	growth, up, organic, strong, sales, digit, acquisition, business, rate, expect, down, strength, grow, line, margin, solid, core, single, segment, guidance
Business outlook	business, up, good, term, margin, down, looking, rate, big, better, customer, forward, guidance, market, off, area, line, opportunity, issue, new
Financial outlook	revenue, growth, operating, margin, segment, increase, business, expect, year-over-year, forecast, result, acquisition, higher, estimate, decline, compare, income, report, strong, EPS
Valuation	multiple, stock, earnings, price, target, valuation, estimate, cycle, risk, growth, market, EPS, current, PE, group, relative, view, investor, peak, upside
Defense contracts	system, program, defense, contract, space, service, budget, electronic, aircraft, information, ship, missile, government, technology, international, sales, air, support, navy, DOD (Department of Defense)
Cash flows and financing	cash, flow, free, share, capital, net, dividend, debt, balance, repurchase, increase, strong, sheet, margin, stock, working, earnings, growth, program, management
Raw materials and input price	cost, price, increase, material, pricing, margin, higher, raw, volume, expect, incremental, up, impact, commodity, product, issue, operating, steel, inventory, benefit
Geographic segments	market, growth, China, Europe, global, emerging, America, demand, region, Asia, India, investment, country, north, economy, expect, middle, economic, European, east
Energy (GICS 1010)	
Comparing financial performance with expectation	estimate, EPS, result, lower, higher, expect, earnings, expectation, share, report, consensus, cost, guidance, forecast, operating, below, management, above, tax, expense
Business outlook	up, term, cost, down, good, looking, market, price, rate, forward, opportunity, big, area, capital, project, issue, business, new, off, better
Cash flow and financing	share, cash, flow, dividend, increase, earnings, estimate, free, debt, repurchase, capital, stock, price, growth, expect, balance, acquisition, management, program, current
Oil and gas production	gas, price, production, natural, oil, MCF (thousand cubic feet), cost, BBL (barrel), higher, estimate, cash, flow, volume, commodity, increase, hedge, realize, crude, lower, share
New project opportunity	growth, capital, project, cost, return, expect, asset, management, opportunity, base, portfolio, cash, production, development, potential, strategy, position, key, significant, focus
Valuation	price, target, estimate, EPS, rating, multiple, base, buy, EBITDA, risk, history, EV/EBITDA, share, earnings, raising, report, expect, maintaining, consensus, increasing
Geographic segments	revenue, increase, activity, north, margin, operating, America, service, up, market, growth, pricing, international, drilling, Mexico, decline, strong, improvement, oilfield, Canada
Offshore drilling	contract, market, deepwater, fleet, drilling, offshore, jackup, rate, dayrate, expect, Mexico, utilization, gulf, new, cost, sea, newbuild, diamond, floater, demand

Income statement items	income, net, tax, expense, operating, interest, revenue, cash, share, asset, dilute, earnings, EBITDA, rate, cost, item, equity, margin, sales, EPS
Energy reserve	reserve, proved, cost, BOE (barrel of oil equivalent), production, asset, value, replacement, FD (finding and development), acquisition, MCFE (thousand cubic feet of gas equivalent), MMBOE (million barrels of oil equivalent), revision, gas, year-end, base, add, price, development, property

Software & Services (4510)

Growth	growth, revenue, margin, business, operating, expect, segment, acquisition, service, expansion, organic, grow, increase, digit, rate, investment, new, strong, improve, improvement
Comparing financial performance with expectation	revenue, estimate, EPS, growth, margin, increase, up, operating, expect, higher, guidance, result, management, lower, expectation, report, below, bps, share, grew
Valuation	price, target, estimate, multiple, share, EPS, rating, valuation, risk, PE, base, market, group, peer, stock, trade, earnings, current, trading, forward
Earnings guidance and expectations	estimate, guidance, EPS, revenue, consensus, expect, result, management, expectation, report, street, line, range, earnings, call, above, upside, growth, below, stock
Income statement items	income, revenue, operating, net, expense, margin, tax, EPS, cost, share, gross, interest, profit, GAAP, dilute, service, general, amortization, sales, pretax
Cash flow valuation model	cash, flow, share, free, value, growth, rate, capital, stock, valuation, terminal, equity, price, debt, DCF (discounted cash flow), estimate, forecast, earnings, base, analysis
Business outlook	business, up, term, good, new, down, product, looking, growth, rate, opportunity, market, customer, better, big, forward, guidance, deal, area, line
Competition	market, revenue, business, share, growth, industry, opportunity, acquisition, cost, product, position, large, margin, operating, significant, competitive, competitor, technology, advantage, management
Enterprise software and IT services	customer, product, sales, new, deal, application, service, license, enterprise, large, software, partner, market, base, vendor, solution, management, vertical, spending, system
Internet advertising	search, advertising, ad, display, revenue, advertiser, online, share, internet, user, paid, site, network, media, ads, EBITDA, growth, TAC (traffic acquisition cost), increase, market

Materials (1510)

Raw material pricing	volume, higher, increase, cost, price, sales, earnings, lower, offset, up, decline, material, segment, pricing, raw, result, expect, operating, improve, strong
Business outlook	business, up, good, down, term, price, cost, market, looking, pricing, customer, forward, better, rate, big, impact, volume, issue, area, start
Valuation	price, estimate, target, EPS, share, multiple, earnings, risk, forecast, expect, cost, increase, base, EBITDA, view, rating, current, reflect, valuation, result
Geographic segments	growth, America, north, Europe, volume, market, sales, asia, currency, strong, region, expect, new, demand, China, up, American, margin, Latin, global
Earnings guidance and expectations	estimate, EPS, guidance, expect, result, consensus, expectation, operating, report, lower, forecast, higher, below, volume, call, sales, segment, line, earnings, outlook
Cash flow and financing	cash, flow, debt, share, dividend, free, capital, balance, net, sheet, repurchase, credit, return, management, strong, expect, stock, earnings, shareholder, buyback
Growth	growth, business, market, new, opportunity, expect, product, management, cost, strategy, focus, key, customer, improvement, position, improve, return, investment, margin, plan
Income statement items	income, net, tax, operating, interest, share, expense, sales, margin, asset, cash, profit, EPS, dilute, equity, earnings, rate, debt, operation, liability
Steel prices and production	steel, ton, price, scrap, cost, market, shipment, product, mill, sheet, raw, increase, tubular, material, capacity, production, import, domestic, construction, flat-rolled

Agriculture	corn, roundup, seed, acre, product, traits, yield, gross, trait, share, market, profit, soybean, Smartstax, pipeline, technology, farmers, cotton, Brazil, biotech

Health Care Equipment & Services (3510)

Growth	growth, margin, revenue, expect, operating, business, rate, gross, digit, market, expansion, improvement, EPS, organic, mix, increase, drive, single, grow, new
Earnings guidance and expectations	estimate, EPS, guidance, share, expect, range, management, result, expectation, consensus, growth, earnings, impact, call, lower, new, below, revenue, report, stock
Geographic segments	sales, up, currency, constant, growth, report, expect, down, product, FX, gross, rate, Europe, business, impact, margin, international, foreign, tax, Japan
Income statement items	income, net, revenue, expense, operating, tax, EPS, margin, gross, interest, share, cost, profit, rate, SGA, dilute, pretax, amortization, item, adjust
Valuation	estimate, EPS, target, multiple, price, share, risk, growth, valuation, PE, stock, earnings, rating, base, trade, industry, group, forward, premium, peer
Medical cost	enrollment, MLR (medical loss rate), commercial, cost, trend, medical, earnings, share, Medicare, expect, ratio, membership, higher, prior, SGA, live, projection, increase, report, premium
Business outlook and opportunities	business, up, term, good, market, down, guidance, impact, looking, forward, rate, new, product, line, opportunity, better, call, cost, issue, start
Cash flow and financing	cash, debt, flow, share, net, asset, capital, current, liability, repurchase, balance, equity, note, investment, free, increase, stock, dividend, sheet, expense
Medicare and Medicaid	Medicare, plan, commercial, member, Medicaid, advantage, health, premium, care, benefit, cost, membership, group, enrollment, business, contract, government, risk, Tricare, individual
Drug trial	announce, disease, drug, product, category, treatment, trial, patient, update, system, new, agreement, Humira (a drug name), study, clinical, program, hub, pharmaceutical, administration, phase

Table 2

Sample Selection and Description

Panel A presents the sample selection procedures for the earnings conference calls. Panel B presents the sample selection procedures for the analyst reports. Revision reports consist of analyst reports issued on the day of or the day after a conference call that contain a revision in at least one of analyst quantitative measures (earnings forecast, stock recommendation, or price target). Panels C and D provide the distribution of reports by year and by industry, respectively.

Panel A: Sample selection – earnings conference call

Earnings conference calls of S&P 500 firms in 2003-2012	18,607
Less earnings conference calls not on days [0, +1] relative to the earnings announcement date	371
Less earnings conference calls without accompanying analyst reports	486
Earnings conference calls on days [0, +1] relative to the earnings announcement dates, with accompanying analyst reports	17,750

Panel B: Sample selection – analyst report sample

	All Reports	Revision Reports
Analyst reports issued for S&P 500 firms in 2003-2012	476,633	220,723
Less analyst reports not within [0, +1] relative to the earnings conference call dates	313,316	114,034
Less analyst reports issued before the start time of the earnings conference calls	4,107	4,107
Number of analyst reports issued on days [0, +1] after the earnings conference calls (denoted, ***AR***)	159,210	102,582
AR as a percentage of total analyst reports issued for S&P 500 firms	33.4%	46.5%

Panel C: Distribution of earnings conference calls and analyst reports (AR), by year

Year	# of conf. calls	# of ***AR***s	# of ***AR***s per call	# of Unique Firms
2003	1,605	11,793	7.35	445
2004	1,674	15,304	9.14	455
2005	1,723	15,570	9.04	469
2006	1,753	14,412	8.22	480
2007	1,767	14,283	8.08	488
2008	1,791	13,368	7.46	470
2009	1,819	14,880	8.18	497
2010	1,875	18,139	9.67	486
2011	1,857	19,118	10.30	487
2012	1,886	22,343	11.85	495
Total	17,750	159,210	8.97	686

Panel D: Distribution of earnings conference calls and prompt analyst reports, by industry

GICS	Industry Group	# of Conf. Calls	# of ARs	# of Unique Firms
2010	Capital Goods	1,395	12,795	48
1010	Energy	1,268	10,573	55
4510	Software & Services	1,207	14,190	49
1510	Materials	1,136	7,701	42
3510	Health Care Equipment & Services	1,107	12,086	42
5510	Utilities	1,037	4,698	41
2550	Retailing	983	10,806	41
4520	Technology Hardware & Equipment	983	10,527	40
4020	Diversified Financials	901	7,538	32
3020	Food, Beverage & Tobacco	883	6,693	32
3520	Pharmaceuticals, Biotechnology & Life Sciences	837	8,506	33
4030	Insurance	753	3,975	25
4010	Banks	731	6,772	31
4530	Semiconductors & Semiconductor Equipment	704	8,608	24
2520	Consumer Durables & Apparel	621	3,768	25
2540	Media	516	6,003	20
4040	Real Estate	447	2,687	21
2530	Consumer Services	442	4,492	16
2030	Transportation	347	2,951	12
2020	Commercial & Professional Services	345	1,896	14
3010	Food & Staples Retailing	335	3,357	11
5010	Telecommunication Services	322	4,477	16
3030	Household & Personal Products	243	2,341	8
2510	Automobiles & Components	207	1,770	8
Total		17,750	159,210	686

Table 3

Summary Statistics on the Conference Calls and Analyst Reports Issued on the Day of or the Day after the Conference Call

Panel A: All individual topics

This panel presents the summary statistics for the number of topics in earnings conference calls and prompt analyst reports. These statistics are presented for the managers' comments during the presentation and the Q&A part (CC), the presentation part of the conference call (CCP), the manager answers in the Q&A part of the conference call (CCA), and the set of analyst reports issued promptly after the conference call (AR).

Document Type	# of documents	Number of such topics in the document					Avg. combined length of these topics
		Mean	Median	Std	Min	Max	
CC	17,750	28.57	29	4.77	3	51	100.00%
CCP	17,748	21.99	22	4.93	2	51	100.00%
CCA	17,328	23.13	23	4.97	1	43	100.00%
AR	17,750	25.94	26	6.75	2	53	100.00%

Panel B: Individual topics with discussion length exceeding 2.5% of the entire discussion

This panel presents the summary statistics for the number of topics for which the discussion length exceeds 2.5% of the entire document in earnings conference calls and prompt analyst reports. These statistics are presented for the managers' comments during the presentation and the Q&A part (CC), the presentation part of the conference call (CCP), the manager answers in the Q&A part of the conference call (CCA), and the set of analyst reports issued promptly after the conference call (AR).

Document Type	# of documents	Number of such topics in the document					Avg. combined length of these topics
		Mean	Median	Std	Min	Max	
CC	17,750	10.51	10	1.88	3	18	83.41%
CCP	17,748	10.26	10	2.00	2	19	86.76%
CCA	17,328	10.04	10	2.18	1	20	83.98%
AR	17,750	9.56	9	2.00	2	18	86.25%

Panel C: Difference in the topic distributions of conference calls and analyst reports

This panel presents the statistics from the Pearson's chi-square tests for the homogeneity between AR and CC with respect to the proportion of sentences in each of the 60 topics (i.e., the null that $T_{AR} = T_{CC}$, where T_{AR} and T_{CC} are topic vectors of AR and CC, respectively, as defined in Section 4.3), and that between AR and CCP, AR and the analyst questions in the Q&A part of the conference call (CCQ), AR and CCA, CCQ and CCA, CCA and CCP. If the two documents are homogeneous, the proportion of sentences in topic i will be equal, i.e., the observed number of sentences in each topic will be equal to the expected number of sentences for the two documents (see Sheskin 2011, P. 644, Eq. 16.2). The chi-square test statistic is calculated as: $\chi^2 = \sum_{j=1}^{60} \frac{[n_{AR}\cdot(S_{AR,j}-p_j)]^2}{n_{AR}\cdot p_j} + \sum_{j=1}^{60} \frac{[n_{CC}\cdot(S_{CC,j}-p_j)]^2}{n_{CC}\cdot p_j}$, where n_{AR} (n_{CC}) is the total number of sentences in the AR (CC); $S_{AR,j}$ ($S_{CC,j}$) is the fraction of sentences in topic j in AR (CC); $p_j = (n_{AR}\cdot S_{AR,j} + n_{CC}\cdot S_{CC,j})/(n_{AR} + n_{CC})$ is the overall proportion of sentences in the two documents that belong to topic j. The degree of freedom of the chi-square test between the two documents is the vector length minus one (i.e., 60 – 1 = 59).

	Pearson's chi-square tests for the homogeneity of the topic distribution in pairs of analyst reports and conference calls						
	# of doc pairs	χ^2			Degrees of freedom	% of the sample document pairs for which the homogeneity is rejected	
		Mean	Std	Median		Significant at 10%	Significant at 5%
AR vs. CC	17,750	145.89	61.32	137.42	59	90.80%	88.58%
Benchmarks:							
AR vs. CCP	17,748	102.60	46.81	93.78	59	71.53%	66.31%
AR vs. CCQ	17,320	86.66	38.09	81.25	59	58.96%	53.73%
AR vs. CCA	17,328	145.24	63.88	137.42	59	88.45%	86.37%
CCQ vs. CCA	17,302	29.10	9.59	28.16	59	0.06%	0.02%
CCA vs. CCP	17,326	63.93	20.24	61.58	59	28.04%	21.63%

Table 4

Descriptive Statistics

This table reports the summary statistics for the variables. Variable definitions are provided in Appendix III.

Variables:	# of obs.	Mean	Median	Std	Q1	Q3
Discovery	17,750	0.273	0.265	0.097	0.205	0.333
Interpret	17,749	0.509	0.503	0.087	0.446	0.565
Determinants of Discovery and Interpret:						
Competition (%)	17,750	0.068	0.065	0.027	0.048	0.084
LitigRisk	17,750	0.280	0.000	0.449	0.000	1.000
StockVol	17,725	0.086	0.074	0.048	0.053	0.104
Persistence	17,743	0.530	0.550	0.482	0.205	0.876
Miss	17,633	0.222	0.000	0.416	0.000	0.000
Uncertain (%)	17,750	0.836	0.811	0.265	0.651	0.986
Qualitative (%)	17,750	80.699	80.919	7.362	76.018	85.663
#Segments	17,750	0.751	0.693	0.747	0.000	1.386
Control Variables for Determinant Tests:						
#Questions	17,750	3.202	3.296	0.659	2.996	3.555
ABS_EPS_Surp	17,623	0.002	0.001	0.004	0.000	0.003
AR_Length	17,750	410.782	366.000	259.774	213.000	558.000
Size	17,724	9.339	9.233	1.083	8.594	9.952
BtoM	17,746	0.468	0.393	0.326	0.248	0.609
#Analysts	17,750	8.970	8.000	4.982	5.000	12.000
Market Reaction Test:						
CAR[0,1]	17,734	0.000	0.000	0.057	-0.030	0.030
Tone_Discovery	17,737	0.139	0.141	0.137	0.057	0.224
Tone_CC	17,750	0.280	0.281	0.112	0.200	0.359
EF_Rev	17,750	-0.001	0.000	0.032	0.000	0.001
Rec_Rev	17,750	0.000	0.000	0.123	-0.050	0.060
TP_Rev	17,750	0.001	0.011	0.158	-0.020	0.046
EPS_Surp	17,623	0.001	0.001	0.005	0.000	0.002
Prior_CAR	17,700	0.003	0.002	0.049	-0.024	0.028

Table 5

Tests of the Determinants of the Analyst Information Discovery and Interpretation Roles

This table reports the coefficient estimates and the *t*-statistics from OLS regressions of $Discovery$ and $Interpret$ on their determinants and control variables. Specifically, we report the results from the following: $Discovery = \alpha + \beta_1 Competition + \beta_2 LitigRisk + \beta_3 StockVol + \beta_4 Persistence + \beta_5 Miss + \beta_6 \#Questions + \beta_7 ABS\_EPS\_Surp + \beta_8 AR\_Length + \beta_9 Size + \beta_{10} BtoM + \beta_{11} \#Analysts + \sum_i \gamma_i I_i + \sum_t \delta_t I_t + \varepsilon$ in Column (1) and $Interpret = \alpha + \beta_1 Uncertain + \beta_2 Qualitative + \beta_3 \#Segment + \beta_4 Miss + \beta_5 \#Questions + \beta_6 ABS\_EPS\_Surp + \beta_7 AR\_Length + \beta_8 Size + \beta_9 BtoM + \beta_{10} \#Analysts + \sum_i \gamma_i I_i + \sum_t \delta_t I_t + \varepsilon$ in Column (2). We further supplement the $Discovery$ and $Interpret$ regressions with the determinants of $Interpret$ ($Uncertain$, $Qualitative$, and $\#Segment$) and $Discovery$ ($Competition$, $LitigRisk$, $StockVol$, and $Persistence$) respectively and report the results in Columns (3) and (4). Variable definitions are provided in Appendix III. Coefficient estimates are shown in bold and their *t*-stats based on standard errors clustered at the firm and year level are displayed in parentheses below. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively using two-tailed tests.

	Dependent Variables			
	$Discovery$	$Interpret$	$Discovery$	$Interpret$
	(1)	(2)	(3)	(4)
$Competition$	**0.220****		**0.195****	**0.001**
	(2.6)		(2.3)	(0.0)
$LitigRisk$	**0.026*****		**0.024*****	**-0.002**
	(3.7)		(3.5)	(-0.3)
$StockVol$	**0.070****		**0.062****	**0.004**
	(2.2)		(2.0)	(0.2)
$Persistence$	**0.001**		**0.002**	**0.001**
	(0.4)		(0.6)	(0.5)
$Miss$	**0.004****	**0.003****	**0.004****	**0.003****
	(2.4)	(2.2)	(2.6)	(2.2)
$Uncertain$		**0.016*****	**-0.022*****	**0.016*****
		(3.3)	(-3.4)	(3.2)
$Qualitative$		**0.002*****	**0.001*****	**0.002*****
		(7.2)	(3.9)	(7.3)
$\#Segments$		**0.005*****	**-0.002**	**0.005*****
		(2.7)	(-0.8)	(2.7)
$\#Questions$	**-0.001**	**-0.012*****	**-0.004****	**-0.012*****
	(-0.6)	(-7.4)	(-2.2)	(-7.5)
ABS_EPS_Surp	**1.068*****	**0.316**	**1.089*****	**0.318**
	(3.3)	(1.4)	(3.5)	(1.4)
AR_Length	**0.004****	**-0.000**	**0.003***	**-0.000**
	(2.1)	(-0.3)	(1.9)	(-0.3)
$Size$	**0.017****	**-0.000**	**0.019*****	**0.000**
	(2.3)	(-0.0)	(2.6)	(0.0)
$BtoM$	**-0.000**	**-0.000*****	**-0.000**	**-0.000*****
	(-1.4)	(-18.0)	(-1.2)	(-17.6)

#Analysts	**0.000**	**-0.001****	**-0.000**	**-0.001****
	(0.1)	(-2.5)	(-0.1)	(-2.6)
Intercept	**0.223*****	**0.507*****	**0.158*****	**0.506*****
	(11.1)	(24.2)	(5.4)	(22.5)
Fixed Effect	Industry, Year	Industry, Year	Industry, Year	Industry, Year
Observations	17,609	17,620	17,609	17,608
Adjusted R^2	0.237	0.512	0.247	0.512

Table 6

Investor Reaction to Analyst Information Discovery and Information Interpretation

This table reports the coefficient estimates and the *t*-statistics from the following OLS regression: $CAR[0,1] = \alpha_1 Tone\_Discovery + \beta_1 Tone\_CC + \gamma_1 EF\_Rev + \gamma_2 Rec\_Rev + \gamma_3 TP\_Rev + \gamma_4 EPS\_Surp + \gamma_5 Miss + \gamma_6 Prior\_CAR + \gamma_7 Size + \gamma_8 BtoM + \gamma_9 \#Analysts + \sum_t \delta_t I_t + \varepsilon$ in Column (1). We then augment the model by including $Tone\_Discovery * Discovery$ and $Discovery$, and $Tone\_CC * Interpret$ and $Interpret$ separately in Columns (2) and (3), respectively, and all of them in Column (4). All variables are defined in Appendix III. Coefficient estimates are shown in bold and their *t*-stats based on standard errors clustered at the firm and year level are displayed in parentheses below. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively using two-tailed tests.

	Dependent Variable $CAR[0,1]$			
	(1)	(2)	(3)	(4)
$Tone\_Discovery$	**0.053*** **	**0.027*** **	**0.053*** **	**0.027*** **
	(12.5)	(3.5)	(12.4)	(3.5)
$Tone\_Discovery * Discovery$		**0.100*** **		**0.101*** **
		(3.5)		(3.5)
$Discovery$		**-0.014****		**-0.014****
		(-2.1)		(-2.1)
$Tone\_CC$	**0.015*** **	**0.014*** **	**-0.004**	**-0.006**
	(3.3)	(3.2)	(-0.4)	(-0.5)
$Tone\_CC * Interpret$			**0.038***	**0.039***
			(1.8)	(1.9)
$Interpret$			**0.004**	**0.004**
			(0.6)	(0.7)
EF_Rev	**-0.054***	**-0.053***	**-0.054***	**-0.053***
	(-1.8)	(-1.7)	(-1.8)	(-1.7)
Rec_Rev	**0.073*** **	**0.073*** **	**0.073*** **	**0.073*** **
	(12.7)	(12.7)	(12.7)	(12.7)
TP_Rev	**0.062*** **	**0.062*** **	**0.062*** **	**0.062*** **
	(3.1)	(3.1)	(3.1)	(3.1)
EPS_Surp	**1.050*** **	**1.034*** **	**1.047*** **	**1.031*** **
	(5.5)	(5.4)	(5.5)	(5.4)
$Miss$	**-0.018*** **	**-0.017*** **	**-0.018*** **	**-0.017*** **
	(-12.5)	(-12.5)	(-12.6)	(-12.5)
$Prior\_CAR$	**-0.079*** **	**-0.079*** **	**-0.078*** **	**-0.079*** **
	(-4.8)	(-4.8)	(-4.8)	(-4.8)
$Size$	**-0.000**	**-0.000**	**-0.000**	**-0.000**
	(-0.7)	(-0.9)	(-0.6)	(-0.7)
$BtoM$	**0.016*** **	**0.017*** **	**0.016*** **	**0.017*** **
	(7.4)	(7.5)	(7.4)	(7.5)
$\#Analysts$	**-0.000*** **	**-0.000*** **	**-0.000****	**-0.000****
	(-3.2)	(-3.2)	(-2.3)	(-2.3)
$Intercept$	**-0.008**	**-0.004**	**-0.012***	**-0.008**
	(-1.6)	(-0.7)	(-1.8)	(-1.1)

Fixed Effect	Year	Year	Year	Year
Observations	17,531	17,531	17,530	17,530
Pseudo R^2	0.152	0.153	0.152	0.153