Working Paper

A Thematic Analysis of Analyst Information Discovery and Information Interpretation Roles

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
March 2014

A Thematic Analysis of Analyst Information Discovery and Information Interpretation Roles

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

March 2014

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.

A Thematic Analysis of Analyst Information Discovery and Information Interpretation Roles

Abstract

Evidence in extant literature on the information interpretation and information discovery roles of sell side analysts is inconclusive. While studies in this literature employ different research designs and sample periods, they uniformly rely on equity market reaction to capture the analyst information role. Because equity market reaction may be incomplete or confounded by the simultaneous release of other information (e.g., earnings release and conference calls), this design choice may have hindered researchers' ability to identify the precise information role analyst play in the capital market. In this study, we introduce novel measures of analyst information discovery and information interpretation that are based on the thematic content of a large sample of analyst reports. These measures allow us to explicitly identify and empirically quantify the amount of information analysts discover and interpret in their reports, without referencing to the equity market reaction. Consistent with information discovery, we document that analyst reports issued promptly after conference calls contain a significant amount of discussion on exclusive topics that were not referred to in the conference calls. Moreover, when analysts do discuss the topics covered in the conference call, they frequently use a different vocabulary from that used by managers, consistent with their information interpretation role. Cross-sectionally, we document evidence that analysts not only respond to investor demand for their services and play a greater information discovery role when firms' proprietary cost is high, but also provide more interpretation when the processing cost of the information in conference calls is high. Finally, we show that investors value both the information interpretation as well as the information discovery role played by analysts.

1. Introduction

Financial analysts serve an important information intermediary role in capital markets (Beaver 1998, p. 10; Frankel, Kothari, and Weber 2006; Ramnath, Rock, and Shane 2008). Through their research, they process and interpret public corporate disclosures, corporate events, and news for investors, and also provide investors new information they discover from their private research efforts.

Evidence in the extant academic literature on the relative importance of the analyst information interpretation and information discovery roles is mixed. For example, Francis, Schipper, and Vincent (2002) and Frankel et al. (2006) conclude that the informativeness of analyst research and earnings announcements complement each other, consistent with the analyst information interpretation role. In contrast, Ivkovic and Jegadeesh (2004) and Chen, Cheng, and Lo (2010) interpret their evidence to suggest that the discovery of new information by analysts is more valuable to investors than their interpretation of the information in earnings announcements. Livnat and Zhang (2012) extend the analyses of Chen et al. (2010) by including other corporate disclosures; they reach the opposite conclusion – the analysts' information interpretation role after firms announce their earnings dominates their information discovery role. Recognizing the conflicting nature of these findings, a literature review by Ramnath et al. (2008) calls for more research on the distinction between the analyst interpretation and discovery roles.

While these studies employ different samples and research methodologies, they uniformly infer analyst information roles from the immediate equity market reaction to the issuance of analyst reports. This research design suffers from two potential limitations. First, it assumes that investors fully understand and instantaneously incorporate the information provided by analysts into stock prices. However, evidence in prior literature suggests that investors have limited attention and processing power and their initial response to information events such as earnings announcements or analyst reports is incomplete (e.g., Hirshleifer, Lim, and Teoh 2011; Zhang 2006). Second, because a vast number of analyst reports are issued immediately after corporate

disclosures (46.5% of analyst revision reports in our sample are issued on the same day or one day after earnings conference calls), researchers are unable to disentangle the immediate market reaction to the information content of analyst reports and that of the earnings release or the conference call. This limitation confounds the ability of the equity market-based measure to identify investor reaction to the information content in analyst reports. Researchers overcome this limitation either by excluding from their samples analyst reports issued in close proximity to the corporate disclosure, potentially resulting in a biased sample (e.g., Chen et al. 2010) or by assuming that such reports serve the information interpretation role (e.g., Livnat and Zhang 2012; Ivkovic and Jegadeesh 2004). These two limitations likely contribute to the appearance of conflicting evidence in the literature on the information role of analysts, in general, and, specifically, the relative importance of their information interpretation and information discovery roles.

In this study, we compare and contrast the information content of sell side analyst reports issued on the day of or the day after conference calls (AR) to that in earnings conference calls (CC).[1] Specifically, we ask (1) do the vast number of analyst reports issued promptly after one of the most important regular corporate disclosure events – earnings release and the adjoining conference call – provide incremental information to the discussions in these calls, and (2) when do analysts play an information discovery role and when do they play an information interpretation role? In stark contrast to prior work, we do not rely on the immediate stock price reaction around the earnings release as a measure of information content; rather, we introduce several novel measures of the thematic content of analyst reports and conference calls and use these measures to test the relative information content of analyst reports as well as their information discovery and interpretation roles.

[1] We limit our analysis to prompt analyst reports, i.e., reports issued on the same day and the day following the conference call, to avoid the confounding effects of other information released between the time of the conference call and the issuance of the analyst report and because these reports constitute a disproportionally large proportion of the total number of reports issued by analysts throughout the year. For simplicity, we treat all analyst reports issued inside this two-day window as a single report (denoted AR). As explained below, because our information content measures do not depend on the immediate equity market reaction, we are able to test the relative information content of AR and CC despite their close proximity.

To construct these measures, we rely on an advanced technique in information retrieval research called *Latent Dirichlet Allocation* (LDA). Developed in a seminal paper by Blei, Ng, and Jordan (2003), LDA is a robust method that relies on statistical correlations among words in a large set of documents to discover and quantify their latent thematic structure (or, topics). LDA can be thought of as a dimensionality-reduction technique, similar to cluster analysis or principle components analysis, but designed specifically for use with text. The algorithmic procedure of LDA is able to handle a massive collection of documents impossible for human coders to process. Moreover, as an unsupervised statistical learning method, LDA does not require training data, annotation, or any prespecified topic labels from the researchers. These desirable features lead to a widespread application of LDA in a variety of fields (e.g., political science, psychology, and economics).

LDA's most desirable feature in the context of this study, is its ability to discover economically interpretable topics within analyst research reports and conference call transcripts. As described in detail in Section 4, the output from LDA can be used to construct precise quantitative measures of the exclusive topics discussed in AR relative to those in CC, as well as measures of the difference in the composition of words used to describe topics that both analysts and managers discuss in their respective AR and CC. We use these measures in tests of the relative information content of analyst reports and the role they play in discovering and interpreting information.

We begin our analysis with a detailed description of the classification of topics discussed in analyst reports and conference calls. Based on a sample of about 18,000 conference call transcripts and over 470,000 analyst reports, we document that the presentation (question and answer) part of the conference call comprises an average of 22 (26) meaningful topics and that analyst reports issued promptly after these calls consist of an average of 26 topics. Over 80% of the content of the CC (AR), however, is devoted to discussions of 9 (11) key topics, suggesting that over ten distinct topics are mentioned but not discussed in great detail in these documents. Following standard validation procedures in the literature, we use the LDA output of word-to-

topic assignment to label the economic intuition of key topics in each industry. As expected, we find that managers and analysts routinely discuss topics related to growth, financial performance (current and outlook), and valuation. In addition, we label many industry specific topics discussed by managers and analysts, for example, drilling in the energy industry, internet advertising in the software industry, and drug trials in the health care industry.

As an additional validation test, we plot the temporal variation in the proportion of discussions devoted to key topics in the banking and telecommunication industry groups during the period 2003 to 2012 and visually examine whether trends in these topics correspond to key economic developments in these industry groups (similar analysis is conducted by Quinn, Monroe, Colaresi, Crespin, and Radev 2010 for U.S. Senate Congressional Record). These graphs visually confirm that manager and analyst discussions closely track economic developments in these industries. For example, during the financial crisis, managers and analysts in the banking sector devoted more of their discussion to mortgage-related issues and deteriorating financial performance and less discussion to mortgage origination and loan growth. The topic discussion in the telecommunications sector gradually shifted from landline services in the early period to the impact of the smartphone business on financial performance in the later period. We interpret these validation tests as support of the ability of LDA to meaningfully discover the thematic content in conference call transcripts and analyst research reports.

To conduct our empirical analyses, we construct measures of the amount of information discovery and information interpretation in AR and test whether, and under what circumstances, they discover new information and interpret existing information relative to the information in quarterly CC. We document that the distribution of topics in prompt analyst research reports is statistically different from that in the presentation part of the CC in about 70% of the cases; this finding suggests that analysts frequently provide new information by discussing exclusive topics that were not referred to in the CC, consistent with the analyst information discovery role. We test the analyst information interpretation role by comparing the vocabularies used by analysts and managers to discuss the top ten topics of the CC (these top ten topics account for about 87%

of the CC presentation). We document that the words used by analysts to describe the top ten CC topics are statistically different from those used by managers to describe the same topics 49% of the time. Further, 98% of the analyst reports contain two or more key CC topics with a statistically different word usage. We interpret this evidence as supporting the information interpretation role played by analysts.

This evidence motivates several hypotheses on the determinants of the cross-sectional variation in analyst information discovery and interpretation roles. Following Verrecchia (1983), we predict that, when proprietary cost is high, managers are more likely to withhold information from their CC discussions. In these cases, we hypothesize that analysts will respond to investors' demand for additional, new information by increasing the amount of private information disclosed in their reports. In a similar vein, when the processing cost of the information in CC is high, we predict that analysts will respond to investor demand to clarify this information by increasing their efforts to interpret the hard-to-understand information in CC.

Evidence from our empirical tests supports these predictions. We document that the amount of new (potentially private) information in AR is increasing in the level of competition (our primary measure of proprietary cost) and that the analyst information interpretation role is positively and significantly related to empirical proxies for the processing cost of the information in CC. Furthermore, we provide evidence that investors value analyst overall information interpretation role and their information discovery role for firms operating in highly competitive environments.

Our study provides several contributions to related literatures. First, by explicitly quantifying the semantic content of analyst research reports and contrasting it with managers' discussions in earnings conference calls, our study advances the understanding of the information role of analysts, as well as the determinants of their information discovery and information interpretation roles. This is an important contribution given the inconclusive evidence in the existing literature on the information role of analyst reports. Second, based on a long-established empirical methodology in information retrieval research, we construct novel measures of the

information content in textual disclosures. Notably, these measures are based on the semantic discussions of economically meaningful topics in these disclosures, rather than the common practice of relying on the immediate market reaction to the release of these disclosures. Because reliance on market reaction to measure information content has the potential to obscure and even bias inferences on the information content of these reports, we encourage other researchers to consider the measures introduced in this study as a viable alternative. Finally, recent research in accounting and finance primarily focuses on the textural characteristics (e.g., readability and tone) of corporate financial disclosures (e.g., MD&A in 10-K and S-1) to examine how texts are being said. With the increased popularity of this type of research, we believe that the topic-modeling methodology introduced in this study will enable interested researchers to significantly expand their analysis of the textual content of corporate financial disclosures beyond the basic understanding of "*how* texts are being said" to a broader understanding of "*what* is being said" in these corporate financial disclosures.

Section 2 reviews related literature. We develop our hypotheses in section 3. Section 4 provides a detailed explanation of our empirical methodology and variable measurement. Sections 5 and 6 describe our empirical tests and results. We conclude the study in Section 7.

2. Review of related literature

2.1. Research on the information role of sell side analysts

Extant research examines the relative importance of the information interpretation and information discovery roles played by analysts (see Ramnath et al. 2008 for a review). Early research (e.g., Dempsey 1989 and Shores 1990) interprets evidence that market reaction to earnings announcements decreases with analyst coverage as consistent with the analyst information discovery role. Using the sum of the absolute stock price reactions to individual analyst reports issued during the year to measure the information content of analyst reports, subsequent studies by Francis et al. (2002) and Frankel et al. (2006) document that the information content of analyst reports and that of earnings announcements are complements,

consistent with analysts predominantly engaging in information interpretation. Ivkovic and Jegadeesh (2004) document a weaker market reaction to analyst revisions during the week after earnings announcements compared to other periods and conclude that the value of analysts' discovering information dominates the value they provide in interpreting information. Chen et al. (2010) document that the association between the market reaction to earnings announcements and that of analyst reports is overall negative before and after earnings announcements, except during the first week after earnings announcements. They suggest the overall negative association implies that analysts' dominant role is information discovery.[2] Livnat and Zhang (2012) extend Chen et al. (2010) by including other types of public corporate disclosure and analyst reports issued within three trading days after the public disclosure, which they refer to as "prompt reports." Livnat and Zhang (2012) find that the majority of analyst reports are prompt reports and that they trigger greater market reaction (measured as the three-day abnormal returns centered on the report date) than non-prompt reports. Livnat and Zhang (2012) assume that prompt reports serve an information interpretation role and conclude that analysts are more valuable interpreting information than discovering information. This conclusion is in contrast to Chen et al. (2012) and Ivkovic and Jegadeesh (2004), but consistent with Francis et al. (2002) and Frankel et al. (2006).

These studies uniformly infer the information content of analyst reports and determine the distinction between their information interpretation and information discovery roles based on the immediate market reaction to the issuance of these reports. Inferences based on this market reaction are influenced by whether the immediate investor reaction is complete and by the confounding effect of the simultaneity of the earnings release, the conference call, and the issuance of analyst reports on the immediate stock price reaction. Perhaps more troublesome, however, is that using market reaction surrounding analyst reports to infer their information content imposes the restrictive assumption that analysts substitute their discovery and

[2] Chen et al. (2010) measure information content as the absolute value of the abnormal stock return on the announcement date, and exclude a large amount of analyst reports issued on days [-1, +1] relative to the earnings announcement dates to mitigate the problem that the two information events are confounded.

interpretation roles. This assumption prevents researchers from investigating whether the analyst information role, even immediately after corporate disclosures, is a combination of both their information interpretation and information discovery roles. For example, Ivkovic and Jegadeesh (2004), Chen et al. (2010), and Livnat and Zhang (2012) assume that prompt reports mostly interpret the information in earnings release while other reports primarily discover information. This assumption may be challenged because a revision issued promptly after public disclosures likely contains both the analyst own private information along with an interpretation of the firm's public disclosures, just as a nonprompt revision may reflect some private information the analyst has discovered as well as a belated interpretation of previously-disclosed public information. Further, Chen et al. (2010) and Francis et al. (2002) make the assumption that a positive (negative) relation between the market reaction to earnings announcements and that of analyst reports implies an information interpretation role (information discovery role). This assumption suffers from a similar problem that a positive correlation between the market reaction to earnings announcements and that of analyst reports can be driven both by information interpretation and by information discovery.

Several other studies examine the textual content of analyst reports either manually or using a computational linguistic classification approach. Asquith, Mikhail, and Au (2005) manually categorize the content of 1,126 reports issued by 56 *Institutional Investor* All-American "First Team" analysts, into 14 justification variables with either positive or negative tone. They find that the market reacts to the tone of some of these variables conditional on earnings forecasts, stock recommendations, and target prices. Livnat and Zhang (2012) read 200 analyst reports and find that 78.5% of them refer to some recent public news. They argue that the evidence is consistent with the analyst information interpretation role. Kothari, Li, and Short (2009) use a dictionary-based method based on General Inquirer to classify positive and negative content in analyst report text. They find an insignificant relation between analyst report content and cost of capital, suggesting that analysts might respond to the market events after the events have taken place. Huang, Zang, and Zheng (2014) classify the textual opinions in over 360,000 analyst

reports using a naïve Bayes machine learning approach and find that the incremental information content of analyst report text is economically significant and its cross-sectional variation can be explained by the characteristics of the reports. Overall, these studies underscore the importance of examining the textual content of analyst reports to understand their information role.

2.2. Information retrieval research that applies the LDA model

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 include Quinn et al. (2010), who use LDA to analyze legislative speech and measure political attention, and Griffiths and Steyvers (2004), who use it to analyze the abstracts in *Proceedings of the National Academy of Sciences* and identify "hot topics." These, and many other studies, establish the legitimacy and validity of the LDA model as an effective method of discovering and summarizing the thematic content in linguistic data.

The application of the LDA model in accounting and finance research is fairly limited. Ball, Hoberg, and Maksimovic (2014) use LDA to extract topics from the MD&A part of corporate 10-K filings and to measure corporate disclosure quality based on the topics extracted. 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.

3. Hypotheses development

3.1. Analyst information role immediately after earnings conference calls

Quarterly earnings announcements and the adjoining earnings conference calls are arguably the most important corporate disclosure events that occur during the year. Immediately following these events, an overwhelming number of sell-side analysts issue revised research reports in which they review the earnings call and provide their clients with an interpretation of

management discussion in the conference call and potentially discover new information to supplement these discussions.[3]

In their information discovery role, analysts provide value to investors by collecting, processing, and providing information that is not as readily available to investors as public disclosure (Ivkovic and Jegadeesh 2004). The information can be collected from public and private channels by visiting stores to collect information on traffic, surveying customers to evaluate customer satisfaction or product quality, investigating suppliers to assess potential input shocks, and conducting research on major competitors to understand the company's competitive advantage. For example, in an Morgan Stanley report issued on August 12, 2011, immediately after J.C. Penney's conference call, the analyst notes: *"The top reason consumers say they shop JCP is due to 'low prices, great discounts' (as per our most recent consumer survey)."* The consumer survey and its conclusions are a result of the analyst's search for private information, and similar information was not included in the conference call. This example anecdotally illustrates the analyst information discovery role immediately after an earnings call.

The information interpretation role suggests that analysts provide value by rephrasing public information in a clearer manner, offering their independent opinions on issues discussed in the conference call, providing comparisons to an objective benchmark, and proposing quantitative assessments of management's subjective statements.[4] The demand for information interpretation arises because analysts possess the financial expertise, in-depth industry and institutional knowledge, and intimate knowledge of the firms they follow, and they are able to dedicate time and effort to process the public information. The analyst information interpretation role could be

[3] About 33% of all analyst reports and 47% of all revision reports in our sample are issued on the day of or the day following the conference calls. After reading 200 analyst reports, Livnat and Zhang (2012) find that the percentage of analyst reports that refer to some recent public news can be as high as 74%.

[4] As an anecdotal example, a Google executive provides the following comments in an October 13, 2011, earnings conference call, "we are pleased with the paid click growth, which increased this quarter, while keeping ad quality high." A report issued on the following day by a Deutsche Bank analyst noted, "Google's core search business continues to maintain healthy growth rates, indicated by the 28% YoY growth in paid clicks." In this example, the analyst modified Google CEO's subjective evaluation on paid click growth with his own opinion on Google's core business's growth and provided quantitative justification using the year-to-year growth rate.

particularly valuable when the amount of public information is vast or when its content is difficult for investors to comprehend (Livnat and Zhang 2012).[5]

In contrast to prior research that assigns analysts either an information interpretation or an information discovery role, a more realistic depiction is that analysts engage in a combination of these roles.[6] First, theoretical models in Kim and Verrecchia (1994, 1997) and Barron, Byard, and Kim (2002) suggest that public information disclosure triggers analysts to produce idiosyncratic information, because analysts combine their private research effort with the information disclosed during public corporate disclosure to produce "uniquely privately inferred information" about firms' future prospects (Mayew et al. 2013; Mayew 2008). In other words, analysts wait until after the earnings conference calls to decide whether the information they discover through their private research efforts would provide investors incremental value beyond the information contained in earnings conference calls. Second, because of the importance of quarterly earnings announcements combined with the adjoining conference calls, analysts compete to issue reports quickly after this significant information event (e.g., Stickel 1989; Mozes 2003). Analysts likely take this opportunity to signal their superior ability to interpret the large amount of qualitative and quantitative information in the conference calls, as well as provide new, private information heretofore unknown to their clients and investors. Accordingly, we hypothesize that:

***H1a**: Analysts serve the information discovery role in reports issued immediately after earnings conference calls;*

***H1b**: Analysts serve the information interpretation role in reports issued immediately after earnings conference calls.*

[5] It is possible that, at times, analysts neither interpret nor discover information, but rather repeat information provided by managers. As explained below, our empirical measures would not consider such discussions as information interpretation or information discovery.

[6] While some prior studies recognize that analyst reports are likely to reflect a combination of both roles, their empirical tests are unable to distinguish between these two roles. Accordingly, they conclude that analysts either interpret or discover information.

3.2. Cross-sectional determinants of analyst information discovery role

We predict that the importance of the analyst information discovery role is increasing in firms' proprietary costs. An extensive theoretical literature on proprietary cost (see reviews in Verrecchia 2001; Dye 2001; Healy and Palepu 2001) demonstrates that proprietary costs represent a significant consequential disclosure cost that prevents managers from being forthcoming because disclosing proprietary information can damage the company's competitive position in the product market. When proprietary cost is high, the impact on the capital market due to proprietary information being withheld tends to be greater, implying a greater value of the withheld information to investors, as well as a higher investor demand for additional sources of information.

When managers withhold value-relevant information, analysts are likely to supplement managers' limited disclosure with information obtained through their private research efforts. For example, managers may withhold early information on research and development of an innovative product or a drug; at the same time, recognizing the project's potential value implication to investors, analysts are expected to use their private research efforts, 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 and provide value-relevant information.[7] Therefore, we hypothesize that:

***H2a**: The importance of analyst information discovery role immediately after earnings conference calls increases with firms' proprietary cost.*

Implicit in our arguments motivating *H2a* is the assumption that analysts engage in greater information discovery for firms with high proprietary cost as a response to investors' demand for more information. We empirically test this assumption by examining whether investors value the

[7] We assume that analysts engage in private information acquisition throughout the quarter. Based on whether the firm they follow is operating in an industry associated with high proprietary costs or when managers withhold information during the conference calls due to proprietary cost concerns, analysts decide whether to include their private information in their reports. Our empirical tests measure proprietary cost at both the industry and the conference call level.

analyst information discovery role, and, specifically, whether investors place a greater weight on analyst information discovery for firms with high proprietary cost. Thus, we hypothesize:

***H2b**: Investors value the analyst information discovery role, and view this role as more valuable for firms with high proprietary cost.*

3.3. Cross-sectional determinants of analyst information interpretation role

Evidence in prior research suggests that the information processing cost of earnings conference calls is generally high because some of managers' spoken disclosure tends to be informal and unstructured, involves ambiguous language, subjective evaluation, and a significant amount of non-financial information (Frankel, Johnson, and Skinner 1999; Brochet, Naranjo, and Yu 2013). Prior research also documents that corporate disclosures that involve high processing costs result in an increasing demand for analyst service and a greater collective effort by these analysts (Lehavy, Li, and Merkley 2011). Accordingly, we predict that analysts are more likely to serve in their interpretation role when the information disclosed during the conference call is harder to process. Formally, we predict that:

H3a*: *The importance of the analyst information interpretation role immediately after earnings conference calls increases with the costs of processing the information in these calls.

Similar to our discussion above, *H3a* relies on the assumption that analysts increase their efforts to interpret the discussions in conference calls when these discussions are hard to interpret, in response to investors' demand for clearer information. Accordingly, our final prediction is that investors value the analyst information interpretation role particularly when the processing costs of this information is high. Formally, we predict:

***H3b**: Investors value the analyst information interpretation role, and view this role as more valuable when the information processing cost of the discussions in earnings conference calls is high.*

4. Empirical methodology

4.1. Topic Modeling and Latent Dirichlet Allocation

A large body of research in computational linguistics investigates the ability of unsupervised machine learning algorithms to analyze the semantic content in large collections of linguistic data and to uncover the thematic structure of this data (see Blei 2012 for a review). These topic modeling algorithms are capable of handling a massive amount of textual data, assigning individual words to specific themes (or topics), and providing a concise probabilistic overview of the themes in the data. 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 simultaneously estimates the topics in large collections of texts and sorts documents into the estimated topics. The researcher is able to use the LDA output to categorize texts according to topics or identify portions of texts that are highly related to specific topics. Topic modeling has several desirable features. First, it generates topics automatically from the texts based on the statistical correlation among words, hence it is capable of handling a massive collection of documents impossible for human coders to process. Second, it is unsupervised, in the sense that it does not require data training or a prespecification of the topics in the data. Therefore, the entire procedure is consistent and replicable (as described below, LDA does require the researcher to input the total number of topics). Finally, the resulting topics are typically interpretable. That is, the distribution of words within topics allows the researcher to discern the content of the topic (Blei 2012; Quinn et al. 2010). Taken together, topic modeling allows analysis and comparison of textual data at the theme level, and it produces measures for testing hypotheses of substantive and theoretical interest.

Latent Dirichlet Allocation (LDA)

Introduced in an influential paper by Blei et al. (2003), Latent Dirichlet Allocation (LDA) has become the most widely used topic modeling algorithm. LDA has been applied and validated in many research areas including political science, psychology, biomedical, economics, and

science.[8] LDA uses a statistical process to imitate the process of generating a document. To do so, the algorithm assumes that all documents share the same set of topics, but the proportion of topics in each document is different. Accordingly, each document is modeled as a probability distribution over these topics, and each of these topics is modeled as a probability distribution over the words in the documents. To generate the entire document, the algorithm assumes that each word in a document is generated in two steps: first, the author selects a topic from the distribution of all available topics; second, given the topic, the author selects a word from the distribution of words representing this topic. Repeating this process word-by-word will probabilistically generate a document by sampling words based on these two (Dirichlet) distributions (See Appendix I for a detailed technical description of the LDA estimation process).

We illustrate the document generation process used by LDA in Figure 1. Assume a collection of documents D contains ten topics and each document has different probabilities over these topics (described by a multinomial distribution with ten parameters). Further, each topic has a multinomial distribution over words. For example, the top four words in Topic 1 (Stores) are: "new," "store," "open," and "square."[9] To generate a document, LDA starts by randomly drawing a topic based on the assumed topic distribution and then randomly drawing a word based on the word distribution of the topic. For example, assume LDA draws Topic 1 and then draws the word "store" from the word distribution of Topic 1. The complete document is generated by repeating this two-step sampling procedure for each word.

Given the assumptions described above, LDA implements a Bayesian procedure to find the model parameters that best fit the textual data. This Bayesian procedure relies on the co-occurrences of words to determine the model parameters. If two words appear frequently in the same document, there is a higher likelihood that LDA will assign them into the same topic. The

[8] For example, Quinn et al. 2010; Grimmer 2010; Atkins, Rubin, Steyvers, Doeden, Baucom, and Christensen 2012; Bao and Dutta 2012; Girffiths and Steyvers 2004; Kaplan and Vakili 2013)

[9] 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.

output from the LDA algorithm comprises a matrix of word frequencies in each topic (**Φ**). Based on the elements of this matrix, we can calculate the probability of a word appearing in a given topic, which is its frequency in the topic divided by the total frequency of all words in the topic.

As described in Appendix II, we applied several preprocessing steps to the conference call transcripts and analyst reports prior to applying LDA. We also set the number of topics to 60 based on the documents' *Perplexity Score* (also described in Appendix II). We perform the LDA analysis on the combined set of all available conference call transcripts and analyst reports, by industry and use the resulting matrix of word frequencies in each topic to construct measures of the overall information content of the analyst reports and the conference calls, as well as the amount of information discovery and information interpretation in these reports (explained in Section 4.3).

4.2. Validation tests of the LDA output

A standard validation technique in studies employing LDA is to manually read high-probability words in key topics and the sentences assigned to these topics in an attempt to discern the underlying content of the topic (e.g., Quinn et al., 2010; Atkins et al., 2012; Bao and Dutta, 2012). This technique provides the researcher with the ability to label the various topics and provides some assurance that the LDA output represents meaningful contextual topics.

Table 1 presents the results of applying this validation technique in our sample. The table reports the top 20 words in each of the top ten topics as well as our inferred topic labels. Results are presented for the five largest industries in our sample (ranked by the total number of conference calls). Overall, the LDA algorithm appears effective in identifying distinct, economically meaningful topics in conference calls and analyst reports. First, words assigned to a specific topic appear semantically related. For example, the frequent appearance of the words "multiple," "target," "price," "valuation," "eps," and "PE," in a specific topic in the Capital Goods industry, suggests that this topic of discussion is related to valuation models and target price. Similarly, the frequent appearance of the words "drug," "trial," "announce," "clinical," and "phase," in a specific topic in the Health Care Equipment & Services industry, suggests that the

discussion in this topic primarily relates to drug trials. LDA also appears effective in uncovering common topics related to a firm's financial performance as well as many industry-specific topics. For example, among the top ten topics, all industries contain discussions of growth- and performance-related topics. In addition, LDA 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 materials. Finally, the LDA algorithm is capable of assigning the same word to multiple topics recognizing the polysemy or the contextual nature of words. For example, LDA classifies the word "price" in both "Valuation" and "Raw Materials and Input Price" in the Capital Goods industry, to reflect the notion that, when used in combination with other words, "price" has different economic meanings. Overall, the evidence in Table 1 suggests that the output from the LDA model provides a reliable delineation of economically meaningful topics in analyst reports and conference call transcripts.

Another commonly used validation test is to examine the correspondence between the temporal variation in the amount of discussion dedicated to key topics and important contextual events. For example, Quinn et al. (2010) examine the Congressional Record of the U.S. Senate and demonstrate that the proportion of key political topics tracks exogenous events such as the 9/11 attack and the Iraq War. Similarly, we visually examine whether the temporal variation in key topics is related to changes in industry and economy-wide conditions. Figure 2 depicts 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). The banking industry experienced significant turmoil over the past decade while the telecommunication industry underwent a significant technology evolution during that time. Therefore, we expect their key topics (and their associated key words) to track the economic developments in these two industries.

Panel A of Figure 2 presents visual evidence of a reliable relation between the temporal variation in the distribution in key topics and the 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." With the advent of the financial crisis in 2007, the proportion of discussion of these two topics declines substantially, while that of "Real Estate Loans" and "Performance and Losses" increases. It also can be seen that the proportion of discussion on the topic labeled "Equity Issuance and TARP" gradually evolves starting in the third quarter of 2008 (the Troubled Asset Relief Program, or TARP was approved on October 3, 2008).

A similarly compelling correspondence between the topics discussed in earnings conference calls and analyst reports and economic conditions is observed for the telecommunication industry. Panel B of Figure 2 depicts a close relation between technological development and topic discussion. For example, discussion of landline related business (e.g., DSL technology), has shrunk steadily, while that on wireless services and smartphone business has grown steadily over time (see topics labeled "Landline Related Services," "Smartphone Business," and "Wireless Subscribers"). Taken together, we interpret the evidence of the validation tests presented in Table 1 and Figure 1 (as well as similar validation tests in related papers) as supporting the effectiveness of LDA to qualitatively identify and quantitatively measure economically meaningful contextual topics.

4.3. Measurement of information discovery and information interpretation

A key contribution of this study stems from our ability to construct explicit empirical measures of the information content of analyst reports issued immediately after earnings conference calls, measures that do not rely on equity market reaction to the issuance of these reports and are not confounded by adjacent public disclosure events. In the next subsections, we describe the process used to construct four related measures for the information discovery and information interpretation roles played by analysts.

4.3.1. *Measuring information discovery based on differences in the distribution of topics between analyst reports and conference calls*

Our empirical tests of the analyst discovery role are based on statistical comparisons of the distribution of topics discussed by management in the presentation part of the conference calls and those discussed in analyst reports issued on the day of or the day following the call.[10] Evidence of a statistically significant difference between the distribution of topics in CC and the adjoining AR is consistent with the information discovery role played by analysts.

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 topic-word frequency matrix $\mathbf{\Phi}$, we construct a frequency vector for each word containing the number of times it appears in each of the K topics. This step results in 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.[11] 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.

4.3.2. *Measuring information interpretation based on differences in word usage between analyst reports and conference calls*

Our empirical tests of the analyst interpretation role are based on a statistical comparison of the distribution of words used by analysts and managers to discuss the top ten topics in the presentation part of the conference call. Statistical evidence that analysts have used different

[10] We focus our analyses on comparisons of the information content of analyst reports with the presentation section of the conference calls. This empirical choice is motivated by the evidence in Matsumoto, Pronk and Roelofsen (2011) that the information content in the Q&A section is primarily driven by analysts' active involvement and by the evidence in Lee (2014) that managers often answer questions from analysts by repeating remarks they had already made in their earlier presentation.

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

words to describe the most important CC topics than those used by managers supports the information interpretation role played by analysts.

To conduct this test, we extract the amount of discussion dedicated of each of the top ten CC topics in the AR and the CC and construct a vector of the word usage for each topic:

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

where each element of these vectors (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).

4.3.3 Measuring the amount of information discovery in analyst reports

Our empirical tests of the determinants of the analyst information discovery role require a summary measure of the amount of information discovery contained in analyst reports issued promptly after earnings conference calls. We define this measure as one minus the cosine similarity between the distribution of topics in a conference call and that in the adjoining analyst reports (i.e., one minus the cosine similarity between the topic vector of CC and AR in Eq. 1).

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.[12] Cosine similarity is neatly bounded in [0, 1], easy to evaluate and calculate, 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).

Formally, we measure analyst information discovery as:

[12] For example, assume there are two topics and two documents. One document has 30% sentences in topic 1 and 70% in topic 2 and the other document has 60% sentences in topic 1 and 40% in topic 2. The cosine similarity of their topic distributions is: $(0.3 \times 0.6 + 0.7 \times 0.4)/\sqrt{(0.3^2 + 0.7^2) \times (0.6^2 + 0.4^2)} = 0.8376$.

$$
\begin{aligned}
&Discovery \\
&= 1 - cosine\ similarity\ between\ T_{AR}\ and\ T_{CC} \\
&= 1 - \frac{\sum_{k=1}^{K}\left(S_{AR,k} \cdot S_{CC,k}\right)}{\sqrt{\sum_{k=1}^{K}\left(S_{AR,k}\right)^2} \cdot \sqrt{\sum_{k=1}^{K}\left(S_{CC,k}\right)^2}}.
\end{aligned} \tag{3}
$$

where $S_{AR,k}$ ($S_{CC,k}$) is the fraction of the discussion in AR (CC) devoted to topic *k*. Intuitively, information discovery occurs when analysts introduce topics that are not included in CC, or are less emphasized by managers in their CC.

4.3.4 Measuring the extent of information interpretation in analyst reports

Our empirical tests of the determinants of the analyst information interpretation role require a summary measure of the extent to which analyst reports provide interpretation of the information contained in CC. We define this measure as the average, over the top ten CC topics, of one minus the cosine similarity between the words used by the analysts to describe each of these topics and the words used by managers to discuss the same topics in their CC (i.e., the difference between W_{AR} and W_{CC} for each of the top ten CC topics). We focus on analyst interpretation of the top ten CC topics to avoid the noise introduced by potentially less important topics, which constitute a small fraction of the CC (on average, the top ten CC topics account for 86% of the CC discussion).

If analyst discussion of a key CC topic contains similar words as those used by managers, our interpretation measure will be associated with low values. If analysts employ a vastly different set of words than managers to describe a top ten CC topic, our interpretation measure will be associated with high values, suggesting that analysts supplement and facilitate investor understanding of the CC discussion of this topic. Formally, our interpretation measure is defined as:

$$
\begin{aligned}
Interpret\ &= \frac{1}{10}\sum_{k=1}^{10}\left(1 - cosine\ similarity\ between\ W_{AR,k}\ and\ W_{CC,k}\right) \\
&= \frac{1}{10}\sum_{k=1}^{10}\left(1 - \frac{\sum_{j=1}^{N}\left(w_{jk} \cdot v_{jk}\right)}{\sqrt{\sum_{j=1}^{N}\left(w_{jk}\right)^2} \cdot \sqrt{\sum_{j=1}^{N}\left(v_{jk}\right)^2}}\right).
\end{aligned} \tag{4}
$$

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; k is one of the top ten topics discussed in the CC.

5. Sample selection and tests of analyst information discovery and interpretation roles

5.1. Sample selection

Our initial sample comprises quarterly earnings conference calls transcripts and all analyst reports issued on the same day or the day following these conference calls for the S&P 500 firms during 2003 through 2012. Table 2 describes our sample selection criteria. As shown in Panel A, from Thomson Reuter's Streetevent Database 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 and keep 18,236 conference calls during days [0, +1] relative to the earnings announcement dates (this is the sample used to perform the LDA model). Excluding 486 conference calls unaccompanied by any analyst reports results in 17,750 earnings conference calls with matched analyst reports in the final sample.

[Insert Table 2 here]

As described in Panel B of Table 2, our initial sample of sell-side analyst reports contains all reports available in the Investext Database for the S&P 500 firms during 2003-2012 (476,633 reports; these are the reports used to perform the LDA analysis). We exclude reports not issued on the day of or the day following earnings conference calls and reports issued on the day of but prior to the start time of the call. We impose this criteria to avoid the potential confounding effects on our analysis of new information issued between the end of the *CC* and the issuance of the *AR*. Our final sample comprises 159,210 analyst reports. The proportion of analyst reports issued for S&P 500 firms on the day of or the day after a conference call constitutes 33% of the entire population of analyst reports (or 47% if we only consider AR containing revisions); it is consistent with the importance of this corporate disclosure event and supports our decision to

focus our analysis on these reports only. The percentage is remarkable considering that these reports are issued in only eight days of the entire year.[13]

Panel C of Table 2 indicates that the number of conference calls increases steadily from 1,605 in 2003 to 1,886 in 2012. The number of prompt analyst reports and the number of reports per call dipped in 2008 to 13,368 and 7.46, respectively, but reached 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.

Panel D presents the GICS industry composition of our sample. The industries with the largest number of earnings conference calls in our sample include capital goods, energy, software and services, materials, and health care equipment and services.

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

Table 3 reports summary statistics of the number of topics in earnings conference calls and analyst reports. Panel A indicates that an average earnings conference call contains discussions on 30 distinct topics; the presentation and Q&A sections contain 22 and 26 topics, respectively. An average set of analyst reports issued promptly after the call contains 26 topics. The relatively high standard deviation in the number of topics in AR suggests a greater variation in the thematic content of these reports.

It is not always the case that all 60 topics generated by LDA for each four-digit GICS industry are discussed in each individual document. Therefore, we report in Panel B of Table 3 summary statistics on the number of topics whose weight in a given document exceeds 2.5% of the entire length of the document. Panel B indicates that, on average, CC (AR) contain discussion of 11 (9) key topics with a standard deviation of around 2; the combined length of these key topics accounts for around 86% of the entire discussion in the presentation part of the CC and the

[13] In untabulated results we find that 83.7% of AR issued during the ten calendar days following conference calls are issued on days [0, +1] relative to the calls, suggesting that our two-day window covers the overwhelming majority of analyst reports prompted by earnings announcements and the adjoining conference calls.

AR. Overall, the summary statistics reported in Table 3 indicate that managers and analysts discuss a large array of topics, but devote the lion's share of the discussion to fewer key topics.

[Insert Table 3 here]

5.3. Tests of Analyst Information Discovery and Information Interpretation Roles

Our first hypothesis (H1a) is that analysts serve the information discovery role in reports issued immediately after earnings conference calls. To test this hypothesis, we compute the Pearson's chi-square statistics and 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). The Pearson's chi-square test is a standard statistical test for the homogeneity of the frequency distribution of certain events (i.e., the frequency of the sentences in each of the 60 topics in our setting) observed in two or more samples (Sheskin 2011, P.638; see a definition of the chi-square test in Table 4).

Table 4 presents the results of these tests. The mean (median) value of the chi-square statistic across all 17,750 pairs of AR and CC is 103.1 (94.17), indicating that the homogeneity between the topic distribution in these documents is rejected 71.7% of the time (at the 10% level). That is, in 71.7% of prompt analyst reports, the set of topics discussed is statistically different than those discussed in the immediately preceding CC. This evidence supports the information discovery role in analyst reports.

For completeness, we also present the chi-square statistics between the topic distributions of AR and the Q&A part of the conference calls, and between the presentation and the Q&A parts of the calls. Because analyst reports tend to include an overview of the discussion in the presentation part but not of the discussion in the Q&A part of the conference call, we expect a greater difference between the topic distribution of the Q&A part of the conference call and AR. In contrast, we expect that the topics discussed in the presentation to be fairly similar to those discussed in the Q&A, because managers often answer questions from analysts by repeating their scripted remarks from the earlier presentation (Lee 2014). Consistent with this expectation, the

difference between the topic distribution of AR and Q&A is significant (at the 10% level) for 91.4% of the sample, suggesting that most prompt analyst reports provide new information beyond the Q&A section of the earnings conference calls. The difference between the topic distribution of the presentation and Q&A part of conference calls is significant (at the 10% level) for only 39.3% of the sample.

[Insert Table 4 here]

Next, we empirically test the analyst information interpretation role (H1b). Our tests attempt to statistically distinguish between analyst reports that describe the key topics in CC using words that are similar to those used by managers to describe the same topics, from those reports that use a different set of words to describe these topics. The latter set of reports likely transformed and paraphrased the information contained in the public disclosure or provided a new perspective, consistent with a meaningful interpretation role.

Empirically, for each CC-AR pair, we use the Pearson's chi-square to test whether the words used by managers and analysts for a given CC topic are significantly different (i.e., we test the null that $W_{AR,k} = W_{CC,k}$ in equation 2 for each of the top ten CC topics). We focus on the top ten CC topics to avoid the noise introduced by economically less important topics.

Table 5 reports the results of these tests. Out of a total of 167,544 top ten CC topics in our sample, the homogeneity between the distribution of words used to describe these topics in CC and AR is rejected (at the 10% level) for 49.4% of the sample.[14] That is, in each set of reports issued promptly after the CC, analysts provide statistically significant interpretation for an average of five of the top ten CC topics. Untabulated results indicate that analyst reports provide statistically meaningful interpretation of at least one (four) of the top ten CC topics in 99.7% (80.7%) of all CC-AR pairs. Taken together, these findings support the information interpretation role analysts play immediately after earnings conference calls.

[Insert Table 5 here]

[14] The number of topic pairs (167,544) is less than 177,500 (17,750 conference calls ×10) because some of the top ten CC topics are not discussed in the associated AR.

6. Tests of the determinants of analyst information discovery and information interpretation roles

6.1. Variable Definitions

Measures of proprietary cost

A recent study by Li, Lundholm, and Minnis (2013) introduces a measure of proprietary cost based on the relative occurrence of words related to competition in the MD&A section of 10-K filings. Li et al. (2013) argue that their measure is superior to existing measures because it directly captures managers' perception of their firms' competitive environment, does not rely on a definition of industry boundary, captures competition from many sources that are hard to identify (e.g., competition from private firms, foreign firms, and potential new entrants), and captures the variation in competition among firms in the same industry, as well as variation in competition across industry.[15] Using this measure they document that the number of references to competition in corporate disclosure captures firms' diminishing ability to earn profits due to new and existing rivals.

Following Li et al. (2013), we construct two measures for firm's proprietary costs. The first relies on the perception of managers of their firms' competitive environment and is measured as the number of references made to competition in the presentation part of the conference call ($Competition\_Firm$). In an alternative specification, we measure competition at the industry level ($Competition\_Industry$) using the percentage of competition references in CC transcripts of all firms in a given industry-year.[16] H1a predicts a positive relation between the competition measures and analyst information discovery.

[15] 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 or non-U.S. companies, or potential entrants and because they rely on certain industry definitions and fail to capture within industry variation in competition. They argue that these limitations likely result in unreliable measures of product market competition.

[16] Following Li et al. (2013), our firm-level competition measure is based on the number of occurrences of "competition," "competitor," "competitive," "compete," and "competing" in the CC. We include words with an "s" appended, and remove phrases that contain negation, such as "less competitive," and "few competitors," and 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 the presentation part of the *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).

Measures of processing costs

We use three measures to capture the cost of processing the information in earnings conference calls. The first measure is the percentage of uncertain words contained in a CC ($Uncertain$). This measure, as well as the list of uncertain words, is developed by Loughran and McDonald (2013).[17] They argue that when managers use words like "may," "assume," "possibly," and "approximately," it is difficult to judge the quality of the information and investors treat such disclosure as ambiguous information signals (see also Epstein and Schneider 2008). Consistent with their argument, they find that more uncertain text in Form S-1 filings makes it more difficult for investors to assimilate value-relevant information, resulting in more volatility in the valuation of an IPO.

Our second measure of processing costs follows Huang et al. (2014) who demonstrate that it is harder to process information in corporate financial disclosures when the information is described using more qualitative and subjective language, as opposed to quantitative analyses.[18] Similarly, we measure the extent qualitative vocabulary is used to discuss firm performance in the CC ($Qualitative$) as one minus the percentage of sentences that contain "$" or "%."

Our last measure of information processing cost is the firms' number of segments ($NSegment$). This measure captures the complexity of the firm's operations and thus correlates with the complexity of the information disclosed by the firm during the CC (Frankel et al. 2006). H3a predicts a positive relation between $Uncertain$, $Qualitative$ and $NSegment$ and $Interpret$.

Control Variables

In our tests of H2 and H3 we control for several firm and report characteristics. We control for the length of the combined prompt analyst reports (AR_Length) following the evidence in

[17] The complete list of uncertain words is available at http://www.nd.edu/mcdonald/word-lists.html.
[18] For example, compare the following two statements from a Google conference call on October 13, 2011: (1) "Turning to our geography performance, the U.S. and rest of worlds are growing at a very healthy pace." (2) "Our revenue from the U.S. was up 26% year-over-year to $4.4 billion; our non-U.S. revenue accounted for 55% of our total revenue or $5.3 billion, up 41% year-over-year." The former statement uses more qualitative and subjective language, thus it is harder for investors to process compared to the latter.

Brown and Tucker (2011) that measures based on cosine similarity are positively correlated with document length. To control for the possibility that the analyst information role is related to the magnitude or the sign of the earnings news, we include the absolute value of the earnings surprise (ABS_EPS_Surp) and an indicator variable for negative earnings surprise (Neg_EPS_Surp). We also control for firm size ($Size$), growth opportunities (book-to-market ratio, $BtoM$) and firm profitability (return on assets, ROA) because they characterize a firm's information environment. In our tests of the equity market reaction to the information discovery and interpretation roles played by analysts we also control for the percentage of CC topics with positive weight (CC_Topic) and the percentage of analyst revision reports (REV_Pct). Finally, we include year and industry fixed effects to control for the common effect across all firms in a year and in an industry, respectively.

6.2. Descriptive statistics

Table 6 reports descriptive statistics on the variables used in our cross-sectional analyses. The mean values of $Discovery$ and $Intepret$ are 0.23 and 0.54, respectively. These values provide initial indication that the discussion in promptly issued AR likely reflects *both* information discovery and information interpretation. The mean of $Competition\_Firm$ ($Competition\_Industry$) is 0.054 (0.056) words per hundred words in a CC, which is comparable to the sample mean of 0.058 in Li et al. (2013). Further, a CC in our sample contains a median of four competition-related words. The mean value of $Uncertain$ is 0.856 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 of $Uncertain$ reported in Loughran and McDonald (2013) for the S-1 filing is 1.41. The mean value of $Qualitative$ indicates that, on average, 71.5% of the sentences in CC are qualitative, higher than the mean of 64.7% reported in Huang et al. (2014) for analyst reports. This finding suggests that managers tend to use more qualitative comments in their earnings conference calls than do analysts in their research reports. On average, a CC in our sample covers 37% of the topics (CC_Topic) in the industry. The median length of the combined

prompt analyst reports (AR_Length) is 5.771 or 366.1 sentences. Of the prompt analyst reports, 63.3% of them contain a revision in earnings forecast, stock recommendation, or target price (REV_Pct). The median number of business segments ($NSegment$) is two (log value of 0.751) and its standard deviation is 2.1. The average earnings surprise (ABS_EPS_Surp) in our sample is 7.5 cents per share and 22% of our sample observations have negative earnings surprise (as indicated by the mean of Neg_EPS_Surp).

6.3. Empirical results of tests of H2 and H3

Table 7 reports the regression results of tests of the cross-sectional determinants of the analyst information discovery role. The dependent variable in columns (1)-(3) is the raw values of $Discovery$ while that in columns (4)-(6) is the decile rankings of $Discovery$. As can be seen in this table, the coefficient estimates on the proprietary cost measures ($Competition\_Firm$ or $Competition\_Industry$) are positive and significant in all specifications. When $Competition\_Firm$ ($Competition\_Industry$) changes from the 25th percentile to the 75th percentile, the raw value of $Discovery$ increases by 0.008 (0.010). To put it into perspective, simulation results show that increasing $Discovery$ by 0.08 (0.010) is equivalent to analysts increasing the length of a discussion of an exclusive topic (i.e., a topic not referred to in the CC) by an average of 4.2% (4.7%) of the total length of the AR. This evidence supports H2a that the amount of information analysts discover in reports issued promptly after conference calls is increasing in firms' proprietary cost. The coefficient estimates on the control variables indicate that the amount of information discovery in prompt analyst reports is also increasing in the sign and magnitude of the earnings news, but is decreasing in the length of the analyst reports.

[Insert Table 7 here]

We tabulate the regression results of tests of the cross-sectional variation in the analyst information interpretation role in Table 8. The dependent variable in columns (1)-(2) is the raw values of $Interpret$ while that in columns (3)-(4) is the decile rankings of $Interpret$. In all specifications, the coefficient estimates on all three measures of the processing cost of the

information in the presentation part of earnings conference calls ($Uncertain$, $Qualitative$ and $NSegment$) are positive and significant. This evidence suggests that, consistent with H3a, the amount of information interpretation analysts provide in reports issued promptly after earnings conference calls increases when the information contained in these calls is more difficult for investors to process. We focus on the coefficient on $Uncertain$ reported in Column (2) to interpret the economic magnitude. Based on simulation results, we find that when $Uncertain$ moves from the 25th percentile to the 75th percentile (i.e., from 0.61% to 1.05%), $Interpret$ increases by 0.011, which is equivalent to analysts using 10% more different words (i.e., words not used by managers for this topic) in discussing a top ten CC topic. The coefficient estimates on the control variables indicate that analyst information interpretation role is greater for firms reporting negative earnings surprise, smaller firms, and firms with lower ratios of book-to-market (i.e., growth firms).

[Insert Table 8 here]

The evidence in tables 7 and 8 suggests that analysts respond to investor demand and simultaneously discover a greater amount of information when firms' proprietary costs are high and provide more interpretation of management CC discussions when the information processing costs are high. In our final analysis, we formally test whether investors appear to value analyst information discovery and interpretation roles and whether they find these roles more valuable when firms' proprietary and information processing costs are high (H2b and H3b).

To test these two hypotheses, we regress investor reaction to the issuance of AR on our measures of analyst information discovery and information interpretation, as well as interaction terms of these measures with $Competition$ and $Uncertain$. We measure investor reaction as the absolute value of the cumulative market adjusted returns on days [-1, +2] around the conference call date (ABS_CAR).[19] Because $Discovery$ and $Interpret$ measure the quantity of the information content in prompt analyst reports but not the favorableness of this information, these

[19] This return window encompasses earnings announcement, conference calls, and all analyst reports in our sample. The four-day window attempts to capture a more complete investor reaction to these information events.

variables might relate to market reaction in a non-linear manner. Therefore, we use an indicator variable for cases where the market reaction is greater than our sample median ($ABS\_CAR\_Indicator$) and perform a logistic regression to test whether the likelihood of observing a large market reaction is related to analyst information roles.

[Insert Table 9 here]

Table 9 reports the results of these regressions. Column (1) presents the results of a baseline regression of market reaction on the thematic-based measures of the information content of analyst reports (i.e., $Discovery$ and $Interpret$) and the control variables. The coefficients on $Discovery$ and $Interpret$ are positive and significant, indicating that the information discovery and interpretation provided by analysts trigger incremental market reactions. This result provides further validation that our information content measures based on the thematic topics generated by LDA likely capture the informativeness of these documents, as perceived by investors.

Column (2) presents the results for a regression that includes interaction terms between measures of analyst information roles and their economic determinants. The coefficient estimates on the interaction term between $Discovery$ and competition is positive and significant (at the 10% level), suggesting that investors value the information discovery role played by analysts when firms appear to withhold information due to proprietary cost concerns, consistent with H2b. That the coefficient on $Discovery$ is statistically insignificant in Column (2), indicates that the informativeness of the information discovery role is likely driven by firms with high proprietary costs. The coefficient on the interaction term between $Interpret$ and $Uncertain$ is statistically insignificant, inconsistent with H3b. Perhaps when the processing cost for the underlying information signal is high, investors' initial reaction to the information event is incomplete. The control variables indicate that market reaction is increasing in the number of topics included in CCs, in the length and the proportion of AR containing a revision, but is decreasing in size. Overall, the evidence in Table 9 is consistent with our hypotheses that investors appear to value both the information discovery and the information interpretation roles

in analyst reports issued promptly after the earnings announcement and the adjoining earnings conference calls.

7. Conclusions

We analyze the information content of analyst research reports and the role they play in discovering and interpreting corporate financial disclosures to capital market participants. To do so, we introduce novel measures of the information content of textual data that rely on algorithmic analyses of the themes (or topics) discussed in this data.

Consistent with information discovery, we find that analyst reports issued promptly after earnings conference calls contain substantial amounts of discussion on exclusive topics that were not referred to in the conference calls. Moreover, when analysts do discuss topics covered in conference calls, they frequently use a different vocabulary than that used by managers, consistent with their information interpretation role. Cross-sectionally, we document evidence that analysts respond to investor demand for their services and play a greater information discovery role when firms' proprietary cost is high, and they provide greater interpretation when the processing cost of the information in conference calls is high. Finally, we show that investors value both the information interpretation as well as the information discovery role played by analysts.

Our study advances the understanding of the information role of analysts as well as the determinants of their information discovery and information interpretation roles, by explicitly quantifying the semantic content of analyst research reports and contrasting it with managers' discussions in earnings conference calls. Additionally, we introduce measures of the information content of textual disclosures that do not rely on equity market reactions to the release of these disclosures. Because reliance on market reaction has the potential to obscure and even bias inferences of the information content of these disclosures, we encourage future research to consider the measures introduced in this study as a viable alternative. Finally, with the increased popularity of textual research in accounting and finance, we believe that the topic-modeling

methodology introduced in this study will enable interested researchers to significantly expand their analysis of the textual content of corporate financial disclosures beyond the basic understanding of "*how* texts are being said" to a broader understanding of "*what* is being said" in these texts.

References

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.

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.

Beaver, W., 1998. Financial reporting: An accounting revolution, third ed. Prentice-Hall, Upper Saddle River, NJ.

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.

Bhorjraj, S., Lee, C.M.C., Oler, D.K., 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41 (5), 745-774.

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.

Boni, L., Womack, K.L., 2006. Analysts, industries, and price momentum. *Journal of Financial & Quantitative Analysis* 41 (1), 85-109.

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

Brown, P., Della Pietra, V.J., Mercer, R.L., Della Pietra, S.A., Lai, J.C., 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics* 18 (1), 31-40.

Buntine, W.L., 1994. Operations for learning with graphical models. *Journal of Artificial Intelligence Research* 2, 159-225.

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.

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

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.

Fang, F., Dutta, K., Datta, A., 2013. LDA-based industry classification. Proceeding of International Conference on Information Systems.

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.

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.

Harris, M.S., 1998. The association between competition and managers' business segment reporting decisions. *Journal of Accounting Research* 36, 111-128.

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.

Hirshleifer, D., Lim, S.S., Teoh, S.H., 2011. Limited investor attention and stock market misreactions to accounting information. *Review of Asset Pricing Studies* 1, 35-73.

Huang, A., Zang, A., Zheng, R., 2014. Evidence on the information content of text in analyst reports. Working paper, Hong Kong University of Science and Technology

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

Kadan, O., Madureira, L.M., Wang, R., Zach, T., 2012. Analysts' industry expertise. *Journal of Accounting & Economics* 54, 95-120.

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.

Lee, J., 2014. Inadvertent disclosure and scripted earnings conference calls. Working paper, Washington University in St. Louis.

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., 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.

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.

Manning, C., Raghavan, P., Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

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.

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

Porter, M., 1980. An algorithm for suffix stripping. *Program* 14(3), 130-137.

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.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The Author-Topic model for authors and documents. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence.

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.

Steyvers, M., Griffiths, T., 2006. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis: A road to meaning.* (427-448). Mahwah, NJ:Erlbaum.

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

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.

Zhang, X.F., 2006. Information uncertainty and stock returns. *The Journal of Finance* 61, 105-137.

Appendix I

Technical Details of the Latent Dirichlet Allocation Model

Assume a corpus consisting of a collection of *D* 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 A1 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.

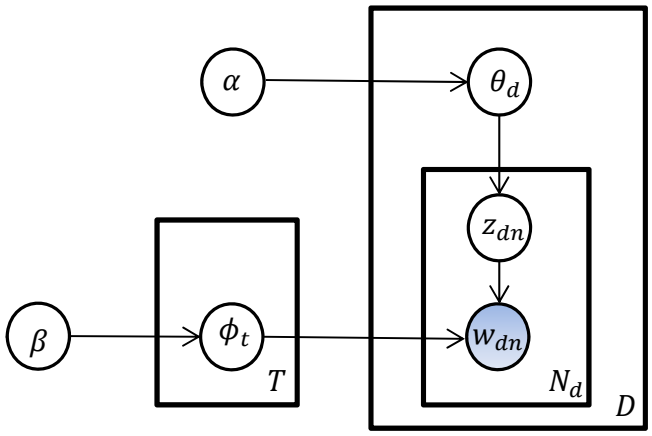


Figure A1. Plate notation depiction of LDA

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 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:

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

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 the 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:

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

Appendix II

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 the S&P 500 firms during 2003-2012. We incorporate all available reports in the LDA to obtain the best representation of topics discussed in these reports. Earnings conference calls are from Thomson Reuter's Streetevent database and analyst reports are 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.

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 number of topics ranging from 2 to 120. As can be seen in Figure A2, 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 document.[2]

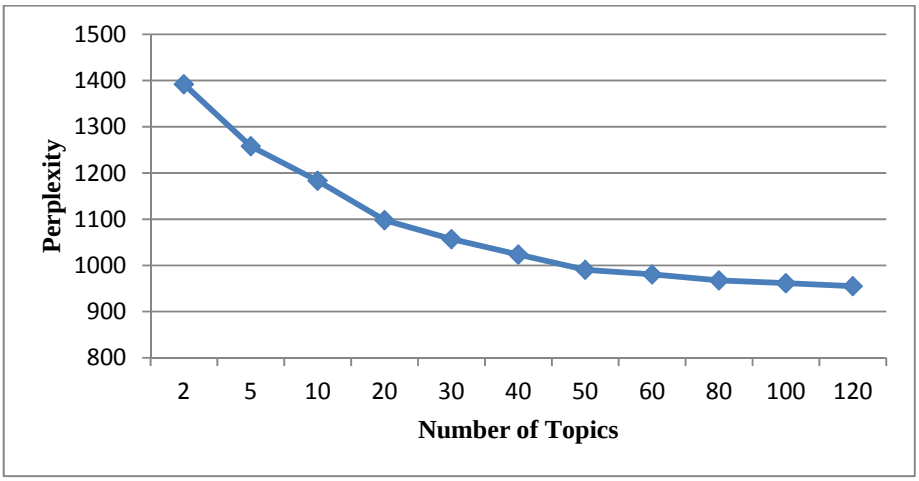


Figure A2. Perplexity of LDA model for different number of topics

[1] We compared LDA results based on 30, 60, and 100 topics for banking and retailing industries. 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 couple-therapy transcripts.

Appendix III

Variable Definition

Variable Name	Definition
$Discovery$	One minus the topic cosine similarity between the CC and the AR.
$Interpret$	One minus the average within topic word cosine similarity in the top 10 topics in CC and the corresponding topics in AR.
$Competition\_Industry$	Percentage of competition related words in CC in the same industry during the last year.
$Competition\_Firm$	Percentage of competition related words in the last CC of the same firm.
$Uncertain$	Percentage of uncertain words in the CC.
$Qualitative$	Percentage of sentences without dollar sign or percent sign in the CC.
CC_Topic	Percentage of non-empty topics in the CC.
AR_Length	Natural log of the total number of sentences in the AR.
$NSegment$	Natural log of a firm's number of segments.
ABS_EPS_Surp	Absolute value of earnings surprise, calculated as the actual EPS minus the last consensus EPS forecast before the earnings announcement, both from I/B/E/S, winsorized at the 98% level.
Neg_EPS_Surp	An indicator dummy variable that equals one if an earnings surprise is negative, and zero otherwise.
$Size$	Natural log of the market value of equity of the firm at the end of the quarter.
$BtoM$	Book value of equity scaled by the market value of equity of the firm at the end of the quarter, winsorized at the top and bottom 1%.
ROA	Net income scaled by the average total assets of the firm in the last quarter, winsorized at the top and bottom 1%.
$ABS\_CAR\_Indicator$	An indicator variable that equals one if the absolute value of the cumulative market-adjusted return over the [-1, 2] window around *CC* is larger than the sample median, and zero otherwise.
REV_Pct	Percentage of AR issued on the day of or the day after a CC that contain a revision in analyst quantitative measures (earnings forecast, stock recommendation, or target price).

Figure 1: An Illustration of the Document Generation Process

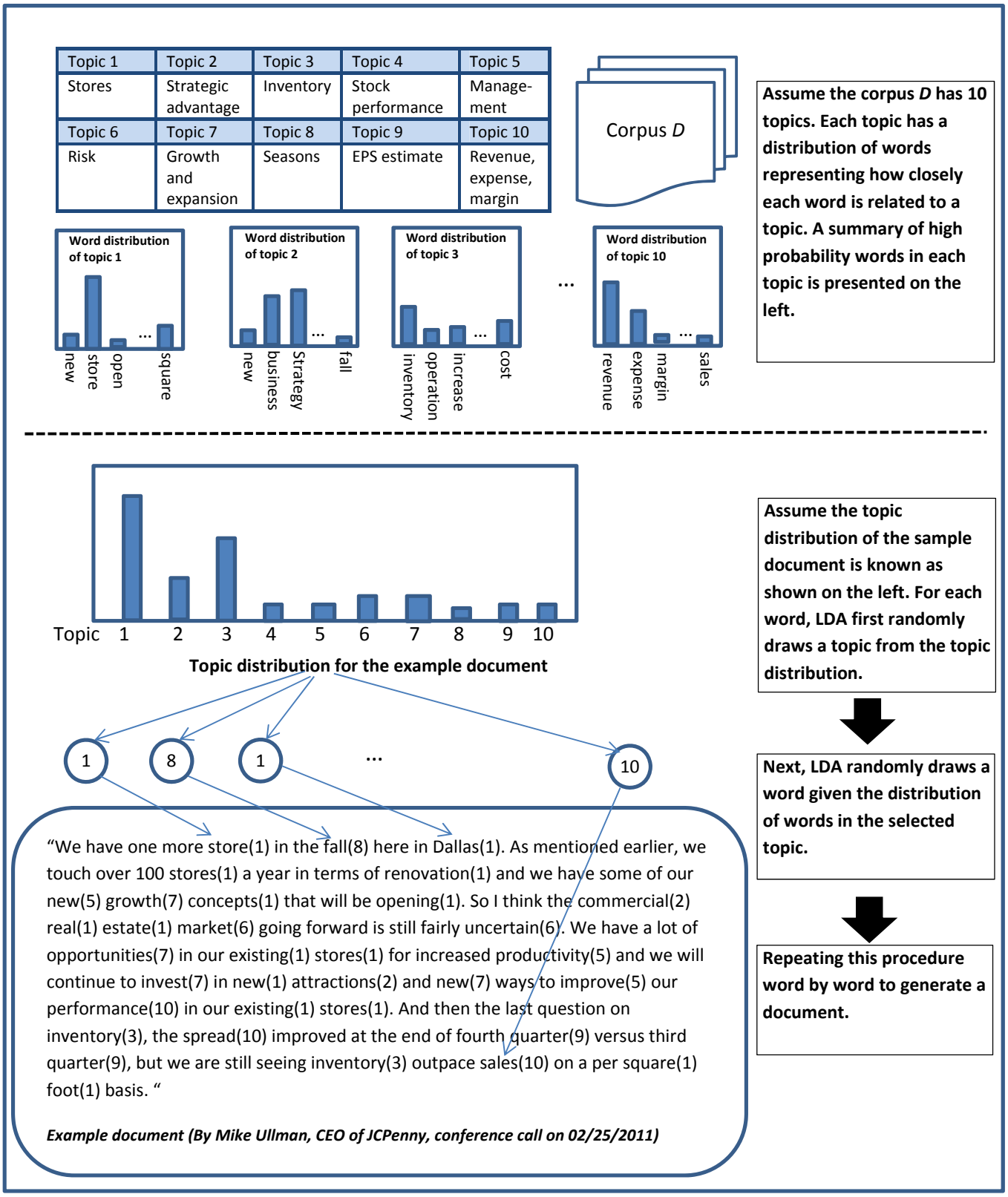
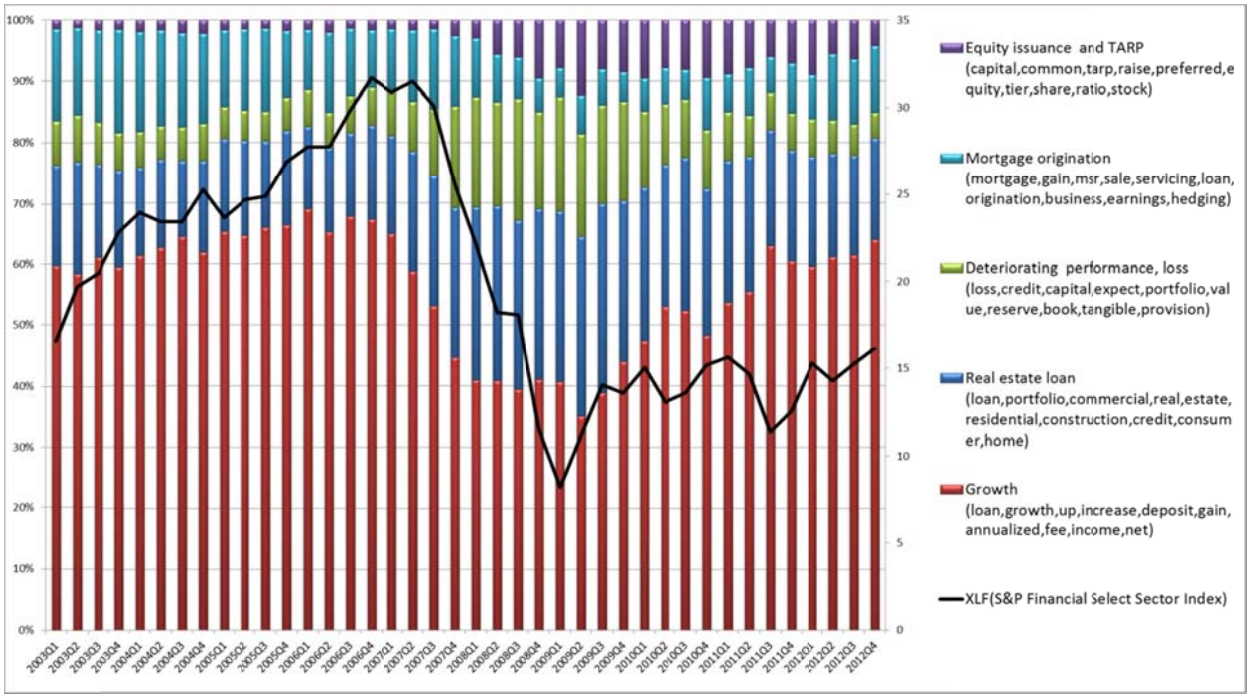


Figure 2
Temporal Variation in the Distribution of Key Topics

Panel A: Banking industry (GICS 4010)



Panel B: Telecommunication industry (GICS 5010)

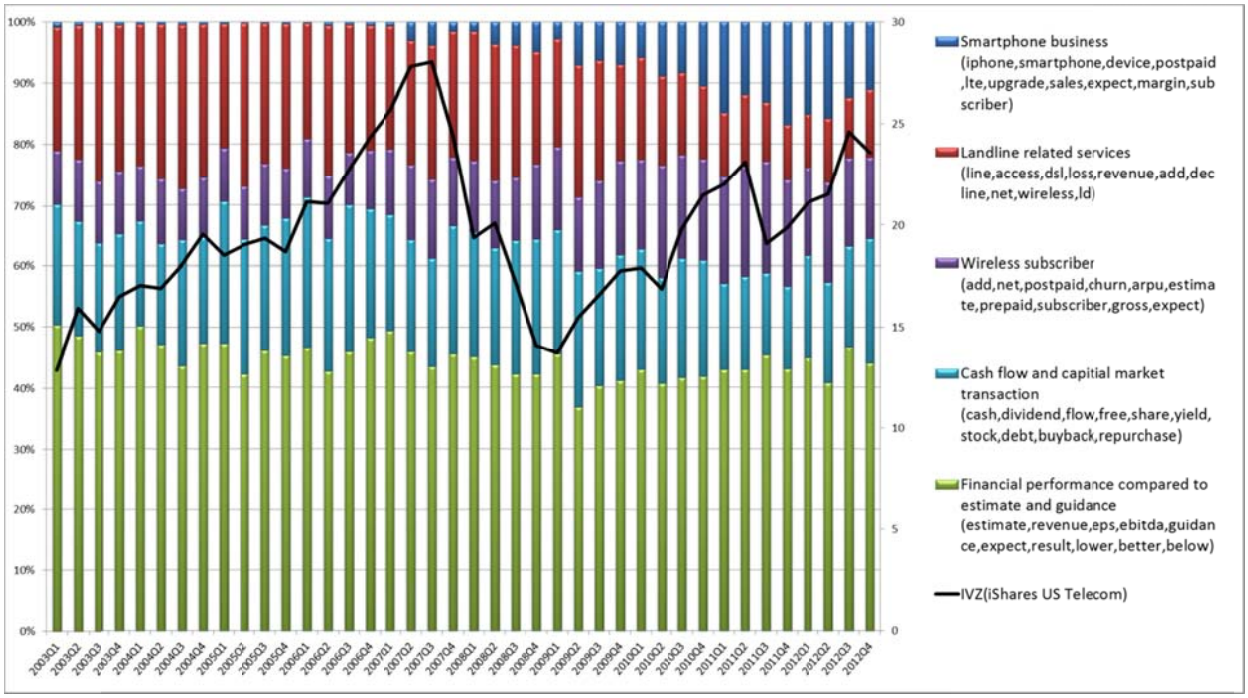


Table 1

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 largest five industries in our sample. Industries are ranked by the total number of conference calls.

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 of the earnings conference call. Panel B presents the sample selection procedures of the analyst report. Revision reports consist of analyst reports issued on the day of or the day after a CC that contain a revision in analyst quantitative measures (earnings forecast, stock recommendation, or price target). Panel 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 ***CC***s	# of ***AR***s	# of ***AR***s per ***CC***	# 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 CC	# 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 Distributions of the Number of Topics in Conference Calls and Analyst Reports Issued on the Day of or the Day after the Conference Call

This table presents the summary statistics of the distribution of topics in earnings conference calls and prompt analyst reports. These statistics are presented for the entire CC, the presentation part of the CC (CC), the Q&A part of the CC ($CCQA$), and the set of AR issued promptly after the CC. Panel A presents statistics of all topics in a given document. Panel B presents the statistics for topics for which the discussion length exceeds 2.5% of the entire document in the document.

Panel A: All individual topics

Document Type	# of documents	Number of such topics in a document					Avg combined length of these topics
		Mean	Median	Std	Min	Max	
Entire conference call	17,750	30.18	30	4.72	9	51	100%
CC	17,750	22.06	22	5.00	2	51	100%
$CCQA$	17,346	25.94	26	4.92	1	45	100%
AR	17,750	25.94	26	6.75	2	53	100%

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

Document Type	# of documents	Number of such topics in the document					Avg combined length of these topics
		Mean	Median	Std	Min	Max	
Entire conference call	17,750	10.53	11	1.85	4	17	82.95%
CC	17,750	10.26	10	2.00	2	19	86.73%
$CCQA$	17,346	10.00	10	2.04	1	18	82.87%
AR	17,750	9.56	9	2.00	2	18	86.25%

Table 4

Tests of Analyst Information Discovery Role

This table presents statistics of 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 $T_{CC} = T_{AR}$, where T_{CC} and T_{AR} are topic vectors of CC and AR, respectively, as defined in Section 4.3). 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%
Information discovery role:							
AR vs. CC	17,750	103.10	47.36	94.17	59	71.7%	66.6%
Benchmarks:							
AR vs. $CCQA$	17,346	152.97	63.64	146.20	59	91.4%	89.6%
CC vs. $CCQA$	17,346	70.33	21.80	67.81	59	39.3%	31.5%

Table 5

Tests of Analyst Information Interpretation Role

This table presents statistics on Pearson's chi-square tests for the homogeneity between AR and CC with respect to their word distributions in the same topic (i.e., test the null that $W_{CC,k} = W_{AR,k}$, where $W_{CC,k}$ and $W_{AR,k}$ are word vectors of CC and AR in topic k, respectively, as defined in Section 4.3). If the discussions of the same topic in CC and AR are homogeneous, the observed frequency of each word will be equal to the expected frequency of the word for both documents (see Sheskin, 2011, P. 644, Eq. 16.2). The chi-square test statistic is calculated as: $\chi^2 = \sum_{i=1}^{N} \frac{(w_{AR,i} - m_{AR} \cdot p_i)^2}{m_{AR} \cdot p_i} + \sum_{i=1}^{N} \frac{(w_{CC,i} - m_{CC} \cdot p_i)^2}{m_{CC} \cdot p_i}$, where $w_{AR,i}$ ($w_{CC,i}$) is the frequency of word i in this topic in AR (CC); $m_{AR} = \sum_{i=1}^{N} w_{AR,i}$ ($m_{CC} = \sum_{i=1}^{N} w_{CC,i}$) is the total number of words in this topic in the AR (CC); $p_i = (w_{AR,i} + w_{CC,i})/(m_{AR} + m_{CC})$ is the overall proportion of word i in this topic in the two documents; N is the total number of unique words in this topic in the two documents. The degree of freedom of the chi-square test between the two documents in the same topic is the number of unique words in the topic minus one (i.e., $N - 1$). These tests are conducted for the top ten topics discussed in the conference call.

	Pearson's chi-square tests for the homogeneity of the distributions of word used to describe a given topic in the CC and AR						
	# of topic pairs	χ^2			Average Degrees of freedom	% of the sample topic pairs for which homogeneity is rejected	
		Mean	Std	Median		Significant at 10%	Significant at 5%
Information interpretation role:							
AR vs. CC	167,544	240.24	209.34	173.97	191.72	49.4%	41.4%
Benchmarks:							
CC vs. $CCQA$	164,497	143.64	108.30	110.85	124.93	30.9%	21.1%

Table 6

Descriptive Statistics

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

Variables:	**# of obs.**	**Mean**	**Median**	**Std**	**Q1**	**Q3**
Discovery	17,750	0.230	0.214	0.116	0.144	0.295
Interpret	17,749	0.543	0.537	0.089	0.480	0.600
CC and AR characteristics:						
Competition_Firm (%)	17,123	0.054	0.034	0.071	0.000	0.080
Competition_Industry (%)	17,750	0.056	0.054	0.024	0.037	0.070
Uncertain (%)	17,750	0.856	0.811	0.357	0.610	1.045
Qualitative (%)	17,750	71.510	71.809	10.916	64.626	78.761
CC_Topic (%)	17,750	36.768	36.667	8.326	31.667	41.667
AR_Length	17,750	5.771	5.903	0.798	5.361	6.324
REV_Pct (%)	17,750	63.305	66.667	26.938	50.000	83.333
Firm characteristics:						
NSegment	17,750	0.751	0.693	0.747	0.000	1.386
ABS_EPS_Surp	17,633	0.075	0.030	0.116	0.010	0.080
Neg_EPS_Surp	17,633	0.220	0.000	0.414	0.000	0.000
Size	17,746	9.351	9.252	1.084	8.605	9.967
BtoM	17,746	0.468	0.393	0.326	0.248	0.609
ROA	17,750	0.016	0.014	0.019	0.005	0.025
ABS_CAR	17,734	0.046	0.032	0.048	0.015	0.060

Table 7

Tests of the Determinants of the Analyst Information Discovery Role

This panel reports the coefficient estimates and the *t*-statistics from OLS regressions: $Discovery = \alpha_1 Competition[Firm\ or\ Industry\ or\ both] + \beta_1 ABS\_EPS\_Surp + \beta_2 Neg\_EPS\_Surp + \beta_3 Size + \beta_4 BtoM + \beta_5 ROA + \beta_6 AR\_Length + \sum_t \gamma_t I_t + \varepsilon$. Variable definitions are provided in Appendix III. Coefficient estimates are shown in bold and their *t*-statistics are displayed in parentheses below. ***, **, and * indicate significance at the 1%, 5%, and 10% levels using two-tailed tests.

	Dependent Variable					
	Discovery			Decile of *Discovery*		
	(1)	(2)	(3)	(4)	(5)	(6)
Competition_Firm	**0.095*** **		**0.096*** **	**2.457*** **		**2.440*** **
	(3.6)		(2.8)	(4.4)		(3.3)
Competition_Industry		**0.350*** **	**0.303** **		**9.510*** **	**8.246** **
		(2.7)	(2.2)		(2.8)	(2.4)
ABS_EPS_Surp	**0.037***	**0.042***	**0.044***	**0.824***	**1.016***	**1.086***
	(1.9)	(1.7)	(1.9)	(1.8)	(1.7)	(1.8)
Neg_EPS_Surp	**0.007** **	**0.004**	**0.004**	**0.162** **	**0.084**	**0.085**
	(2.3)	(1.0)	(1.0)	(2.2)	(0.9)	(0.9)
Size	**0.003**	**-0.002**	**-0.002**	**0.023**	**-0.071**	**-0.072**
	(1.2)	(-0.6)	(-0.6)	(0.4)	(-1.0)	(-1.0)
BtoM	**0.007**	**-0.016**	**-0.017***	**0.211**	**-0.367**	**-0.379**
	(0.9)	(-1.6)	(-1.7)	(1.2)	(-1.5)	(-1.5)
ROA	**0.029**	**0.078**	**0.113**	**0.925**	**1.921**	**2.832**
	(0.3)	(0.5)	(0.8)	(0.4)	(0.5)	(0.8)
AR_Length	**-0.046*** **	**-0.039*** **	**-0.040*** **	**-0.939*** **	**-0.802*** **	**-0.813*** **
	(-15.7)	(-11.5)	(-12.4)	(-14.6)	(-10.5)	(-11.0)
Intercept	**0.463*** **	**0.438*** **	**0.439*** **	**9.580*** **	**8.879*** **	**8.877*** **
	(19.5)	(14.5)	(14.1)	(16.3)	(11.9)	(11.5)
Fixed Effect	Industry, Year	Year	Year	Industry, Year	Year	Year
Observations	17,015	17,629	17,015	17,015	17,629	17,015
Adjusted R^2	27.0%	8.1%	8.4%	28.7%	6.2%	6.5%

Table 8

Tests of the Determinants of the Analyst Information Interpretation Role

This panel reports the coefficient estimates from OLS regressions: $Interpret = \alpha_1 Uncertain + \alpha_2 Qualitative + \alpha_3 NSegment + \beta_1 ABS\_EPS\_Surp + \beta_2 Neg\_EPS\_Surp + \beta_3 Size + \beta_4 BtoM + \beta_5 ROA + \beta_6 AR\_Length + \sum_t \gamma_t I_t + \varepsilon$. Variable definitions are provided in Appendix III. Coefficient estimates are shown in bold and their *t*-statistics are displayed in parentheses below. ***, **, and * indicate significance at the 1%, 5%, and 10% levels using two-tailed tests.

	Dependent Variable			
	Interpret		Decile of *Interpret*	
	(1)	(2)	(3)	(4)
Uncertain	**0.025*** **	**0.025*** **	**0.724*** **	**0.741*** **
	(5.4)	(5.5)	(6.0)	(6.2)
Qualitative	**0.001*** **	**0.001*** **	**0.017*** **	**0.017*** **
	(3.9)	(3.9)	(3.7)	(3.7)
NSegment		**0.005** **		**0.159** **
		(2.4)		(2.4)
ABS_EPS_Surp	**0.023** **	**0.023** **	**0.642***	**0.653***
	(2.0)	(2.0)	(1.7)	(1.7)
Neg_EPS_Surp	**0.004*** **	**0.004*** **	**0.107*** **	**0.111*** **
	(3.1)	(3.2)	(2.9)	(3.0)
Size	**-0.002**	**-0.002**	**-0.102***	**-0.112** **
	(-1.2)	(-1.3)	(-1.9)	(-2.1)
BtoM	**-0.008***	**-0.009** **	**-0.269***	**-0.314** **
	(-1.8)	(-2.1)	(-1.7)	(-2.0)
ROA	**0.060**	**0.065**	**2.043**	**2.220**
	(1.0)	(1.0)	(1.0)	(1.1)
AR_Length	**-0.067*** **	**-0.067*** **	**-2.044*** **	**-2.036*** **
	(-37.3)	(-37.1)	(-34.9)	(-34.5)
Intercept	**0.900*** **	**0.898*** **	**15.963*** **	**15.902*** **
	(38.9)	(38.6)	(23.0)	(22.8)
Fixed Effect	Industry, Year	Industry, Year	Industry, Year	Industry, Year
Observations	17,628	17,628	17,628	17,628
Adjusted R^2	46.3%	46.4%	44.1%	44.3%

Table 9

Investor Reaction to Analyst Information Discovery and Information Interpretation

This panel reports the coefficient estimates and the z-statistics from the following logistic regressions: $ABS\_CAR\_Indicator = \beta_1 Discovery + \beta_2 Interpret + \beta_3 CC\_Topic + \beta_4 AR\_Length + \beta_5 ABS\_EPS\_Surp + \beta_6 REV\_Pct + \beta_7 Size + \beta_8 BtoM + \sum_t \gamma_t I_t + \varepsilon$ in Column (1). We then augment the model by including $Discovery * Competition\_Firm\ Indicator$, $Competition\_Firm\ Indicator$, $Interpret * Uncertain\ Indicator$, and $Uncertain\ Indicator$ in Column (2). $Competition\_Firm\ Indicator$ ($Uncertain\ Indicator$) is a dummy variable that equals one if $Competition\_Industry$ ($Uncertain$) is larger than the sample median, and zero otherwise; all other variable definitions are defined in Appendix III. Coefficient estimates are shown in bold and their z-stats are displayed in parentheses below. ***, **, and * indicate significance at the 1%, 5%, and 10% levels using two-tailed tests.

	Dependent Variable *ABS_CAR_Indicator*	
	(1)	(2)
Discovery	**0.314****	**0.037**
	(2.1)	(0.2)
Interpret	**0.402***	**0.670****
	(1.7)	(2.2)
*Discovery * Competition_Firm Indicator*		**0.484***
		(1.7)
Competition_Firm Indicator		**-0.074**
		(-1.0)
*Interpret * Uncertain Indicator*		**-0.524**
		(-1.5)
Uncertain Indicator		**0.382****
		(2.0)
CC_Topic	**1.417*****	**1.443*****
	(6.9)	(6.9)
AR_Length	**0.379*****	**0.366*****
	(13.2)	(12.5)
ABS_EPS_Surp	**0.976*****	**0.959*****
	(6.7)	(6.5)
REV_Pct	**0.002*****	**0.002*****
	(3.0)	(3.4)
Size	**-0.329*****	**-0.322*****
	(-19.5)	(-18.7)
BtoM	**-0.054**	**-0.060**
	(-1.0)	(-1.1)
Intercept	**-0.363**	**-0.543***
	(-1.2)	(-1.8)
Fixed Effect	Year	Year
Observations	17,593	16,995
Pseudo R^2	3.83%	3.96%