"Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach"

Allen Huang, Reuven Lehavy, Amy Zang, Rong Zheng

Internet Appendix

Table of Content:

Internet Appendix I

Technical Details of Latent Dirichlet Allocation (LDA)

Assume that a corpus *D* consisting of a collection of documents contains a fixed number of latent topics. Each document, d, is characterized by a discrete probability distribution over topics (θ_d). Each topic, t, is characterized by a discrete probability distribution over words (ϕ_t). Given this assumption, a document, *d*, can be generated by repeating the two-step sampling: 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}, in the following two steps:

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

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 multinomial distributions. The objective of LDA is to estimate the parameters $\{\theta_d\}$ and $\{\phi_t\}$ that maximize the probability of observing the actual documents.

To simplify the computation, the model assumes that the multinomial topic and word distributions follow Dirichlet priors with known parameters of α and β, i.e., $p(\theta_d)\sim\text{Dirichlet}(\alpha)$ and $p(\phi_t)\sim\text{Dirichlet}(\beta)$.

Given this framework, the probabilistic process can be illustrated using a plate notation (Buntine, 1994). Figure IA1 shows the graphical model of LDA used in Blei et al. (2003). Arrows indicate the 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 sampling. 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. As discussed above, the LDA assumes that α and β are known parameters. The words (w_{dn}) are observed by LDA. The parameters 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.

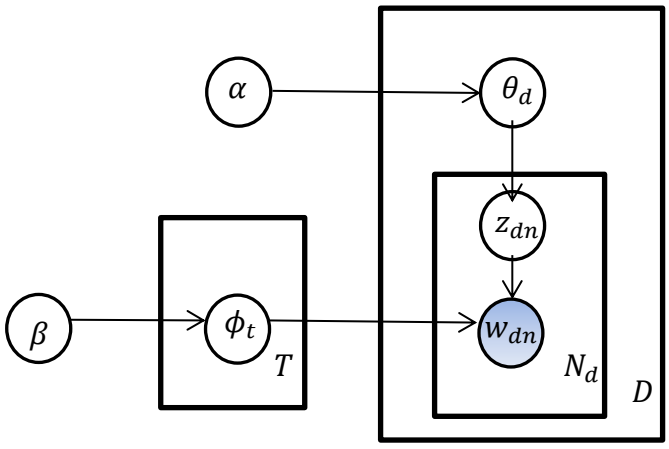

Figure IA1. Plate notation depiction of LDA

The estimation problem of LDA is to find the most likely distribution of the latent variables (i.e., ϕ_t, θ_d, and z_{dn}), given the observed documents and the assumed parameters (α, β). These distributions, however,

are intractable to estimate with closed-form solutions in general (Blei et al., 2003). The most commonly used estimation algorithm for LDA is the collapsed Gibbs sampling, as described in detail 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:

$$p(z_{dn} = t|w_{dn} = m, z_{-dn}, \alpha, \beta) \propto \frac{C^{WT}_{mt,-dn} + \beta}{\sum_{m'} C^{WT}_{m't,-dn} + W\beta} \times \frac{C^{T}_{t,-dn} + \alpha}{\sum_{t'} C^{T}_{t',-dn} + T\alpha} \quad \text{(IA1)}$$

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 Eq. IA1 is the posterior conditional probability of word *m*, given topic *t*, multiplied by the probability of topic *t*, i.e., $p(t|w) \propto p(w|t)p(t)$. Please see Blei et al. (2003) and Steyvers and Griffiths (2006) for more details.

Many applications of the LDA algorithm, including the application in this study, require estimates of the word-topic distribution (ϕ_t) and the topic-document distribution (θ_d). These distributions can be directly calculated from the count matrices as follows:

$$\phi_t = \frac{C^{WT}_{mt,-dn} + \beta}{\sum_{m'} C^{WT}_{m't,-dn} + W\beta}, \quad \text{(IA2)}$$

$$\theta_d = \frac{C^{T}_{t,-dn} + \alpha}{\sum_{t'} C^{T}_{t',-dn} + T\alpha}. \quad \text{(IA3)}$$

Internet Appendix II

Examples of Analyst Information Discovery and Interpretation from Excerpts of Earnings Conference Calls and Analyst Reports

This appendix provides examples of analysts' information discovery and interpretation from excerpts of earnings conference calls and analyst reports. The analyst information discovery role is analysts' discussion in the prompt reports on topics that receive little or no attention from managers during the CC, and the information interpretation role is their discussion of CC topics in the prompt reports.

A. Examples of information discovery

Example 1

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

Christopher Kinney Whitmore (Analyst, Deutsche Bank):
"We believe Apple's PC margins are likely in the 28-30% range, above overall corporate gross margins, suggesting that additional uptake of Macs could drive EPS upside in coming quarters."
Bill Shope (Analyst, JPMorgan):
"We believe that iPod gross margins are now trending above 26%."
Tsvetan Knitisheff (Analyst, Kintisheff Research):
"Since Intel-based Mac PC-s have the same pricing as previous models, we expect improved gross margins in the PC business of Apple as the cost base of Intel-based computers is expected to be lower compared to Power PC based PC-s. Instead, the fact that iPod revenues exceeded Mac PC sales for the first time, combined with still-low (albeit improving) margins in the iPod resulted in 27.3% corporate gross margin excluding stock compensation expenses (SCE)."

Peter Oppenheimer, Apple's CFO, provides a brief statement in regards to this topic:
"I don't want to be, for competitive reasons, specific, nor do we want to talk about specific iPod sales, but I will tell that you that the iPod gross margins in the December quarter were above 20%. . . . As regards to the Intel-based Mac gross margins, we don't want to provide specific gross margins for any of our products."

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

Example 2

Another example of the analyst information discovery role is related to a topic labeled as "Acquisitions," excerpted from Google's earnings conference call held on October 13, 2011. Management discussion

includes very little in regards to its acquisitions. Analysts provide incremental information regarding the timing, motivation, and potential operating and financial implications of the Motorola acquisition. Accordingly, based on the classification of the LDA model, and as can be seen from the excerpts below, we classify analyst discussion of this topic as information discovery.

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

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

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

Jeetil Patel (Analyst, Deutsche Bank Research):
"The company still expects the Motorola deal to close early in 2012.
We suspect this may be part of the motivation behind the Motorola Mobility acquisition--the need to own more of the stock to control more of the search economics in addition to content and applications.
As such, we think that Google is strategically (and perhaps defensively) positioning itself similarly with the Motorola buy, whereas in the near term the core advertising business is executing exceptionally well.
As such, we see where Motorola fits in for Google, but we are hoping for quick deployment of Google-Mot handsets post deal-close, which should enable it to innovate from an application/functionality & ultimately ad standpoint."

Mayuresh Masurekar (Analyst, Collins Stewart):
"Reiterate BUY, on global online advt growth, accelerating mobile revenues, incremental display, Android optionality and inexpensive valuation at 12x 2012 PF EPS ex cash even after Motorola acquisitions."

Nick Landell-Mills (Analyst, Indigo Equity Research):
"In 2006, YouTube ($1.7 bn) & Postini were acquired.
Google pays the 3rd party websites fees for this; referred to as the majority of TAC (Total Acquisition Costs).
This acquisition places Google to compete with some of its partners, the handset makers who use Android."

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

B. Examples of Information Interpretation

Example 1

Below we provide two examples of analyst information interpretation with low and high values of $NewLanguage$. Recall that to measure the extent to which analysts use their own language to discuss CC topics, we use one minus the cosine similarity between the word usages by analysts and managers (bounded between [0,1]). The first example is taken from the Applied Materials Inc. earnings conference

call held on May 16, 2006. Management discusses a topic labeled as "product order." Analyst discussions of this topic are associated with a relatively low value of $NewLanguage$ at 0.141, suggesting that the analyst discussion resembles that of management and provides little clarification for the original statement:

Nancy Handel (Senior VP and CFO):

> *"Orders by major geographic areas were Korea, 22%; Taiwan, 19%; North America, 18%; Japan, 17%; Southeast Asia and China, 14%; and Europe, 10%.*
> *In the quarter, DRAM orders represented 27% of silicon systems orders, flash memory orders were 24% and foundry orders were 17%. Logic and other orders comprised the remaining 32%.*
> *300 millimeter orders represented approximately 73% of total systems orders, and 74% of the system orders were for 100 nm and below process technology."*

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

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

> *"Orders by geography were as follows: Korea 22%, Taiwan 19%, North America 18%, Japan 17%, SE Asia/China 14%, and Europe 10%.*
> *Orders by segment were as follows: DRAM 27%, Flash 24%, foundry 17% and logic/other 32%.*
> *300mm orders represented 73% of total system orders and 74% of system orders was for the 100nm technology node and below."*

Robert Maire (Analyst, Needham & Company):

> *"Geographic Order Breakdown: The distribution of orders was as follows; Korea 22%, Taiwan 19%, North America 18%, Japan 17%, Southeast Asia and China 14%, and Europe 10%.*
> *The Memory Monster - In the quarter, DRAM orders represented 27% of silicon systems orders, flash memory orders were 24% up from 18% in the first quarter and foundry orders were 17% down slightly from 19%.*
> *300-mm orders represented about 73% of total systems orders and 74% of systems orders were for 100-nm and below process technologies."*

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

> *"Taiwan represented 19% of new orders in Q2, North America was 18%, Japan 17%, Europe 10%, Korea 22%, and Southeast Asia and China represented 14%.*
> *DRAM orders represented 27% of total Q2 system orders (versus 28% in Q1), flash was 24%, (versus 18% sequentially), foundries accounted for 17% (versus 19% sequentially) and logic and other revenues accounted for the remaining 32% versus 35% in Q1.*
> *300-millimeter tools represented approximately 73% of total system orders received in the quarter versus 84% sequentially."*

Jay Deahna (Analyst, JPMorgan):

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

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

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

Tim Summers (Analyst, Stanford Financial Group):

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

Example 2

Our second example is associated with a high value of $NewLanguage$ of 0.855 related to analyst interpretation of the *EZstore Initiative* discussion in the Dollar General Corporation's management

earnings conference call from May 26, 2005. This discussion is part of a topic labeled "store operation," which was discussed in detail in both the conference call and analyst reports. As can be seen, the analysts provide additional context, details, and opinions relative to management discussion of this topic.

David Perdue (Chairman and CEO, Dollar General):

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

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

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

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

Ralph Jean (Analyst, Wells Fargo Securities):

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

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

> *"Before EZ Store, boxes were unloaded manually one by one and sorted in the back-room or elsewhere in the store. When the truck arrives at the store, the driver alone is responsible for rolling the container off the truck and into the back room. Employees then push the containers into designated areas of the store.*
>
> *The EZ Store initiative, which has now been rolled out to more than 1,200 stores, or roughly 20% of the overall chain, is a process reengineering program that (in our opinion) revolutionizes the truck unloading process and has the potential to drive considerable efficiencies through what has been, until now, a labor intensive and generally inefficient process.*
>
> *We believe EZ Store reduces the amount of time necessary to unload the truck from 12 hours to about an hour and a half."*

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

> *"The benefits of EZ Store include lower turnover, lower costs to run a store, including lower workers' compensation costs, and fewer damages to merchandise.*
>
> *Distribution center processes are also changing as a result of the EZ Store program.*
>
> *The EZ Store rollout has implications for hiring, training, scheduling, product presentation and product handling."*

Mark Miller (Analyst, William Blair & Company):

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

John Zolidis (Analyst, Buckingham Research Group):

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

Table IA1

Highest Probability Words in the Top Ten Topics of Three Large Industries

This table reports the top 20 words in each of the top ten topics and our inferred topic labels for three of the five largest industries in terms of the total number of conference calls in our sample (another two are reported in Table 1 of the paper).

Topic Label	Top 20 Words
	Energy (GICS 1010)
Comparing financial performance with expectation	estimate, EPS, result, lower, higher, expect, earnings, expectation, share, street, consensus, cost, guidance, forecast, operating, below, management, per-share, tax, in-line
Business outlook	margin, term, cost, pressure, good, looking, market, price, rate, forward, opportunity, big, area, capital, project, issue, business, new, earnings, better
Cash flow and financing	share, cash, flow, dividend, cash-flow, earnings, estimate, free, debt, repurchase, capital, stock, price, growth, expect, balance, acquisition, management, free-cash-flow, 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 model	target-price, estimate, EPS, rating, multiple, base, buy, EBITDA (, risk, history, EV/EBITDA (enterprise value/EBITDA), share, method, earnings, raising, report, price-to-earnings, maintaining, consensus, increasing
Geographic segments	revenue, increase, activity, north, margin, operating, America, service, grow, 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, tax-rate, earnings, EBITDA, rate, cost, item, after-tax, 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, revenue-growth, strong, improve, improvement
Comparing financial performance with expectation	revenue, estimate, EPS, growth, margin, increase, beat, operating, expect, consensus, guidance, result, management, lower, expectation, report, below, in-line, share, grew

Valuation model	price, target, target-price, multiple, share, EPS, rating, valuation, risk, price-to-earnings, 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, in-line, range, earnings, call, EPS-estimate, upside, growth, below, per-share
Income statement items	income, revenue, operating, net, expense, margin, tax, EPS, cost, share, gross, interest, profit, GAAP, dilute, tax-rate, general, amortization, sales, pre-tax
Operating cash flow	cash, flow, share, cash-flow, value, growth, rate, capital, stock, valuation, terminal, equity, free-cash-flow, debt, DCF (discounted cash flow), estimate, forecast, earnings, base, analysis
Business outlook	business, growth, term, good, new, positive, product, looking, growth, rate, opportunity, market, customer, better, big, forward, guidance, deal, area, future
Competition	market, revenue, business, share, growth, industry, opportunity, acquisition, cost, product, position, large, margin, operating, significant, competitive, competitor, technology, advantage, competition
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, off-set, inflation, decline, material, segment, pricing, raw, result, expect, operating, improve, strong
Business outlook	business, growth, good, improving, term, price, cost, market, looking, pricing, customer, forward, better, rate, big, impact, volume, issue, area, positive
Valuation model	price, estimate, target, EPS, share, multiple, earnings, risk, forecast, expect, cost, target-price, base, EBITDA, price-to-earnings, rating, current, reflect, valuation, result
Geographic segments	growth, America, north, Europe, volume, market, sales, Asia, currency, strong, region, expect, new, demand, China, segment, American, margin, Latin, global
Earnings guidance and expectations	estimate, EPS, guidance, expect, result, consensus, expectation, operating, EPS-estimate, lower, forecast, higher, per-share, volume, call, sales, segment, in-line, earnings, outlook
Cash flow and financing	cash, flow, debt, share, dividend, free-cash-flow, capital, balance, net, sheet, repurchase, credit, return, cash-flow, 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, operating-income, 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

Table IA2

Difference in the Topic Distributions of Prompt Analyst Reports and Conference Calls

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

	Pearson's chi-square tests for the homogeneity of the topic distribution in pairs of analyst reports and conference calls						
	# of doc pairs	χ^2			Degrees of freedom	% of the sample document pairs for which the homogeneity is rejected	
		Mean	Std	Median		Significant at 10%	Significant at 5%
AR vs. CC	17,749	136.49	52.36	130.47	59	91.01%	88.97%
Benchmarks:							
AR vs. CCP	17,749	97.74	40.93	93.78	59	70.52%	65.61%
AR vs. CCQ	16,947	86.63	37.48	81.25	59	60.00%	54.36%
AR vs. CCA	17,491	137.57	57.53	137.42	59	88.59%	86.19%
CCQ vs. CCA	16,939	34.73	10.44	28.16	59	0.17%	0.09%
CCA vs. CCP	17,491	61.14	17.11	61.58	59	21.65%	15.31%

Table IA3

Investor Reaction to Analyst Information Discovery and Information Interpretation, Controlling for Other Analyst Outputs

This table investigates whether investors react to analyst information discovery and information interpretation roles, controlling for other analyst outputs. It reports the coefficient estimates and *t*-statistics from the following OLS regression: $CAR[0,1] = \alpha_1 Tone\_Discovery + \beta_1 Tone\_Interpret + \gamma_1 Tone\_CC + \gamma_2 EF\_Rev + \gamma_3 REC\_Rev + \gamma_4 TP\_Rev + \gamma_5 EPS\_Surp + \gamma_6 Miss + \gamma_7 EPS\_Surp * Miss + \gamma_8 Prior\_CAR + \gamma_9 Size + \gamma_{10} BtoM + \gamma_{11} \#Analysts + \sum_t \delta_t I_t + \varepsilon$. All variables are defined in Appendix III of the paper and Table IA7 of this Internet Appendix. *t*-stats based on standard errors clustered at the firm and year levels are displayed in parentheses below the coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively, using two-tailed tests.

	Dependent Variable $CAR[0,1]$
$Tone\_Discovery$	0.034***
	(8.7)
$Tone\_Interpret$	0.051***
	(13.7)
$Tone\_CC$	0.031***
	(6.3)
EF_Rev	-0.033
	(-1.3)
REC_Rev	0.068***
	(12.5)
TP_Rev	0.053***
	(3.0)
EPS_Surp	2.553***
	(10.2)
$Miss$	-0.012***
	(-8.5)
$EPS\_Surp * Miss$	-2.425***
	(-7.3)
$Prior\_CAR$	-0.093***
	(-6.0)
$Size$	-0.000
	(-0.1)
$BtoM$	0.016***
	(7.7)
$\#Analysts$	-0.000***
	(-3.3)
$Intercept$	-0.021***
	(-3.9)
Year Fixed Effect	Yes
Observations	16,923
Adjusted R^2	0.182

Table IA4

Investor Reaction to Analyst Information Discovery and Information Interpretation, Conditional on the Consistency between the Tone of Analyst Discovery and the Tone of Interpretation

This table investigates whether investors' reactions to analyst information discovery and information interpretation vary when analyst information discovery and interpretation tones are more consistent. It reports the coefficient estimates and *t*-statistics from the following OLS regression: $CAR[0,1] = \alpha_1 Tone\_Discovery + \alpha_2 Tone\_Discovery * Diff\_D\_I + \beta_1 Tone\_Interpret + \beta_2 Tone\_Interpret * Diff\_D\_I + \gamma_1 Tone\_CC + \gamma_2 EPS\_Surp + \gamma_3 Miss + \gamma_4 EPS\_Surp * Miss + \gamma_5 Prior\_CAR + \gamma_6 Size + \gamma_7 BtoM + \gamma_8 \#Analysts + \sum_t \delta_t I_t + \varepsilon$. All variables are defined in Appendix III of the main paper and Table IA7 of this Internet Appendix. *t*-stats based on standard errors clustered at the firm and year levels are displayed in parentheses below the coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively, using two-tailed tests.

	Dependent Variable $CAR[0,1]$
$Tone\_Discovery$	0.049***
	(10.1)
$Tone\_Discovery * Diff\_D\_I$	-0.026***
	(-2.7)
$Tone\_Interpret$	0.057***
	(11.0)
$Tone\_Interpret * Diff\_D\_I$	0.004
	(0.4)
$Tone\_CC$	0.039***
	(7.9)
EPS_Surp	2.820***
	(10.5)
$Miss$	-0.013***
	(-8.8)
$EPS\_Surp * Miss$	-2.496***
	(-7.8)
$Prior\_CAR$	-0.060***
	(-4.8)
$Size$	-0.000
	(-0.5)
$BtoM$	0.016***
	(8.3)
$\#Analysts$	-0.000***
	(-3.6)
$Intercept$	-0.021***
	(-3.9)
Year Fixed Effect	Yes
Observations	16,923
Adjusted R^2	0.139

Table IA5

Placebo Test of the Determinants of the Analyst Information Interpretation Role

This table reports the findings of a placebo test, in which we randomly divide analyst reports into two groups and use the word distribution differences between them ($NewLanguage\_Analyst$) as the dependent variable. It reports the coefficient estimates and *t*-statistics from OLS regressions of $NewLanguage\_Analyst$ on the determinants of $NewLanguage$ and control variables. Variable definitions are provided in Appendix III of the paper and Table IA7 of this Internet Appendix. *t*-stats based on standard errors clustered at the firm and year levels are displayed in parentheses below the coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively, using two-tailed tests.

	Dependent Variables $NewLanguage\_Analyst$
$Uncertain$	-0.000
	(-1.5)
$Qualitative$	0.000
	(1.9)
$\#Segments$	0.003
	(1.2)
$Miss$	0.003
	(1.2)
ABS_EPS_Surp	0.529**
	(2.1)
$Expr$	-0.001
	(1.4)
$Star$	0.003
	(0.5)
$\#Questions$	-0.007***
	(-3.5)
$Size$	0.010***
	(5.1)
$BtoM$	0.011*
	(1.7)
AR_Length	-0.000***
	(-21.3)
$\#Analysts$	-0.014***
	(-10.5)
$Intercept$	0.706***
	(27.2)
Industry and Year Fixed Effects	Yes
Observations	17,291
Adjusted R^2	0.694

Table IA6

Determinants of Analyst Information Discovery and Interpretation Roles for Individual Reports

This table examines the determinants of analyst information discovery and interpretation roles at the individual report level. It reports the coefficient estimates and *t*-statistics from OLS regressions of $Discovery$ and $NewLanguage$ on their determinants and control variables at the individual report level. Variable definitions are provided in Appendix III of the main paper and Table IA7 of this Internet Appendix. *t*-stats based on standard errors clustered at the firm and year levels are displayed in parentheses below the coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively, using two-tailed tests.

	Dependent Variables	
	$Discovery$	$NewLanguage$
	(1)	(2)
$Competition$	0.019**	
	(2.2)	
$LitigRisk$	0.148***	
	(4.0)	
$Uncertain$		0.020***
		(4.5)
$Qualitative$		0.001***
		(6.0)
$\#Segments$		-0.002
		(-1.0)
$Miss$	0.006***	-0.001*
	(2.8)	(-1.8)
ABS_EPS_Surp	0.040	0.030
	(0.1)	(0.1)
$AnalystExpr$	-0.000	0.000
	(-0.6)	(0.6)
$AnalystStar$	-0.002	-0.003
	(-0.9)	(-1.4)
$\#Questions$	0.105	-0.303
	(1.0)	(-1.4)
$Size$	0.005***	-0.000
	(2.6)	(-0.0)
$BtoM$	-0.006	0.003
	(-1.0)	(0.8)
$Report\_Length$	-0.000***	0.001***
	(-5.7)	(19.1)
$Intercept$	0.255***	0.643***
	(10.8)	(37.1)
Industry and Year Fixed Effects	Yes	Yes
Observations	117,979	121,365
Adjusted R^2	0.095	0.169

Table IA7

Definitions of Variables that Appear in the Internet Appendix

Variable Name	Definition
EF_Rev	The consensus analyst earnings forecast for the next fiscal year immediately after the conference call minus that immediately before the conference call, scaled by the stock price of the firm 10 days prior to the conference call, winsorized at the top and bottom 1%.
Rec_Rev	The consensus analyst stock recommendation immediately after the conference call minus that immediately before the conference call. Analyst stock recommendations are coded as: 5 (Strong Buy), 4 (Buy), 3 (Hold), 2 (Underperform), and 1 (Sell).
TP_Rev	The consensus analyst target price immediately after the conference call minus that immediately before the conference call, scaled by the stock price of the firm 10 days prior to the conference call, winsorized at the top and bottom 1%.
$Diff\_D\_I$	The absolute value of the difference between $Tone\_Discovery$ and $Tone\_Interpret$.
$AnalystExpr$	The experience of the analyst issuing the report, measured as the number of years since the analyst first issued a forecast in I/B/E/S.
$AnalystStar$	An indicator variable that equals one if the analyst is ranked as Institutional Investor All-Star analyst during the year, and zero otherwise.
$Report\_Length$	The number of sentences in the analyst report.
$NewLanguage\_Analyst$	One minus the average within-topic cosine word similarity between $AR1$ and $AR2$, two random sets of analyst reports, in the AR topics. The within-topic cosine word similarity between $AR1$ and $AR2$ for a given topic k is calculated as $\frac{\sum_{j=1}^{N}(w_{jk} \cdot v_{jk})}{\sqrt{\sum_{j=1}^{N}(w_{jk})^2} \cdot \sqrt{\sum_{j=1}^{N}(v_{jk})^2}}$, where, w_{jk} is word j's frequency in the discussion of topic k in $AR1$; v_{jk} is word j's frequency in the discussion of topic k in $AR2$; and N is the total number of unique words in $AR1$ and $AR2$. AR topics are the topics in which the discussion length exceeds 2% of the AR.

References

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J. Machine Learn. Res*. 3(Jan):993–1022.

Buntine, WL (1994) Operations for learning with graphical models. *J. of Artificial Intelligence Res*. 2:159-225.

Sheskin DJ (2011) Handbook of Parametric and Nonparametric Statistical Procedures. Fifth ed. Chapman and Hall/CRC Press.

Steyvers M, Griffiths T (2006) Probabilistic topic models. *Handbook of Latent Semantic Analysis.*