Business News and Business Cycles*

Leland Bybee
Yale University

Bryan Kelly
Yale University,
AQR Capital,
and NBER

Asaf Manela
Washington University
in St. Louis and IDC

Dacheng Xiu
Booth School of Business
University of Chicago

First Draft: December 2017

This Draft: July 2023

Abstract

We propose an approach to measuring the state of the economy via textual analysis of business news. From the full text of 800,000 *Wall Street Journal* articles for 1984–2017, we estimate a topic model that summarizes business news into interpretable topical themes and quantifies the proportion of news attention allocated to each theme over time. News attention closely tracks a wide range of economic activities and can forecast aggregate stock market returns. A text-augmented VAR demonstrates the large incremental role of news text in forecasting macroeconomic dynamics. We retrieve the narratives that underlie these improvements in market and business cycle forecasts.

Key words: Textual analysis, stock market timing, macroeconomic news, attention, Wall Street Journal, volatility, VAR, machine learning, narrative retrieval

*Previously titled "The Structure of Economic News." We benefited from discussions and comments from George-Marios Angeletos, Scott Baker, Nick Bloom, Steve Davis, Rob Engle, Xavier Gabaix, Will Goetzmann, Toby Moskowitz, Lasse Pedersen, Jesse Shapiro, Bob Shiller, Alp Simsek, and participants at many seminars and conferences. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. The authors have nothing further to disclose.

1 Introduction

A central function of the economics profession is measuring the state of the economy and developing models that link these measurements to distributions of future outcomes. This arms consumers, investors, and policymakers with the information and structural context necessary to allocate resources efficiently. But the economy is a complex system whose current state defies simple measurement. Vast resources, both public and private, are devoted to measuring the many facets of economic activity.[1] Researchers routinely analyze this overlapping ecosystem of numerical business cycle indicators. Yet ascertaining the nature and evolution of the state of the economy from these data is a notoriously difficult task.

In this paper, we offer an approach to measuring the state of the economy via textual analysis of business news. The media sector, as a central information intermediary in society, continually transforms perceptions of economic events into a verbal description that we call "news." This transformation involves describing events, interpreting their meaning, forecasting their impacts, and inferring their causes. The information disseminated by news media is an equilibrium outcome determined by the confluence of consumer preferences and news production technologies (Mullainathan and Shleifer, 2005). As such, news text is a mirror of the prevailing economic issues that are important to both news consumers and producers. News is a one-stop shop for simultaneously understanding many facets of the state of the economy and how they interact with each other.

While media reflects information that consumers rely on to make allocation decisions in an evolving environment, little work has directly studied the structure of news. We focus our analysis on the full text of the *The Wall Street Journal* (*WSJ*), consisting of approximately 800,000 articles from 1984–2017. We summarize this dense verbal description of the state of the economy via a topic model. Topic models are a popular dimension reduction technique from the fields of machine learning and natural language processing. They have two essential elements. Just as principal component analysis condenses large data matrices into a comparatively low number of common factors, a topic model's first element reduces an inherently ultra-high dimension representation of a text corpus into a relatively low-dimensional set of common "topics." The formation of topics is unsupervised—they are estimated as clusters of terms (words or phrases) that tend to co-occur in articles. Those clusters are optimized so that relatively few clusters (many fewer than the number of distinct terms in the data set)

[1]For example, the BEA maintains detailed national accounts for income and spending at the aggregate and industry level, the BLS specializes in measuring employment, wages, and general price levels, and the Federal Reserve disseminates large data sets of macroeconomic indicators (McCracken and Ng, 2016). In addition, equity and debt markets provide a barometer for economic conditions facing the corporate sector. And academic researchers maintain and distribute indices of economic volatility (Engle, 2019) and recession probabilities (Chauvet and Piger, 2008), to name a few.

preserve as much of the meaning in the original corpus as possible by best explaining the variation in term usage across articles.

A topic model's second main element estimates the proportion of text dedicated to each topic in an article. These proportions are a valuable map from the common topics to the individual, article-level narratives that invoke those topics. But more importantly, they quantify the amount of news *attention* allocated to each topic. This makes it possible, for example, to analyze the interaction between news and economic activity. In short, our topic model accomplishes two important tasks of 1) summarizing dominant themes throughout the history of the *WSJ* and 2) tracking how media attention to news topics evolves. We study these as a new quantitative description of the state of the economy.

Our main empirical findings are the following. First, we characterize the topical structure in business news. *WSJ* news decomposes into easily interpretable topics with intuitive time series patterns. A model with 180 topics is a statistically optimal specification according to a Bayes factor criterion. Models with fewer topics tend to mix news themes into overly broad clusters, while allowing for more topics adds parameters without reliably improving model fit.

Almost all topics exhibit strong time series persistence. A long open question in financial research is why do asset returns exhibit such strong volatility clustering. While price volatility could partly drive news coverage, it is unclear why volatility would persist. A leading hypothesis for this fact is that volatility is driven by news arrivals, and that news itself arrives in clusters (e.g., Engle et al., 1990). Newspaper coverage is naturally persistent because journalists pay a fixed cost to learn a new topic and then write cheaper follow-up stories (Boydstun, 2013). The persistence in news topic attention supports this hypothesis. For example, *WSJ* coverage of news associated with the "oil drilling" topic {key terms: *exxon mobil, cubic feet, drill rig, offshore oil*} and "oil market" topic {key terms: *opec, nonopec, oil minister, oil demand*} has similar time series dynamics as the volatility of crude oil prices.

We find that news is a combination of recurrent, seasonal, and emergent topics. Recurrent topics are those garnering media attention consistently throughout the sample. For example, the "Federal Reserve" topic {key terms: *greenspan, yellen, federalfunds rate, fomc*} and "health insurance" topic {key terms: *hmo, health plan, health coverage, blue cross*} are regularly active throughout the sample. Seasonal topics include "presidential elections" {key terms: *obama, romney, dukakis, campaign finance*} and "earnings forecasts" {key terms: *analyst poll, earn forecast, earningspershare, earn expectation*}[2], which draw attention with clockwork regularity. Emergent topics are dormant for much of the sample, then are triggered by

[2]The phrasing of key terms reflects text processing choices described in Appendix A. For example, "earn forecast" and "earn expectation" reflect stemming and lemmatization, which reduce "earnings" to "earn" and "expectations" to "expectation." Likewise, "earningspershare" is an example of the treatment of punctuation (dropping hyphens in this case).

particular events and often remain elevated after. Examples are the "terrorism" topic {key terms: *taliban, queda, suicide bomber, osama*} and the "natural disasters" topic {key terms: *katrina, quake, tsunami, hurricane*}. These examples also illustrate the fact that many topics describe subjects of the news (elections, the Fed, earnings), and do not assign an obvious value of assessment of good news versus bad news. For other topics, the subject itself carries a value assignment (terrorism, recession, natural disasters). And there are a small number of sentimental topics, such as the "concerns" topic {key terms: *raise concern, major concern, express concern, increase concern*}, that provide directional color to the subject topics which co-occur in a given article.

We report a variety of validation checks showing that economic topics identified from the *WSJ* text coincide with conceptually related measures of specific economic activities. For many data series, ranging from output and employment to financing activity, asset prices, and uncertainty, we find that a small subset of thematically related news topics provide a close match to the paths of numerical macro series. Remarkably, we find that news attention explains 25% of the variation in aggregate stock market fluctuations.

More detailed examples of economic activity further demonstrate the explanatory power of news attention. We find that the volume of leveraged buyout (LBO) transactions is most associated with the "takeovers" news topic {key terms: *poison pill, hostile takeover, share tender, higher offer*} and "control stakes" news topic {key terms: *control stake, majority stake, minority shareholder, acquire stake*}, which help explain 58% of the variation in LBO activity. Similarly, the time series behavior of IPO volume is closely tracked by news attention to the "IPO" topic {key terms: *ipo market, ipo price, roadshow, lockup*} and "venture capital" topic {key terms: *joint venture, venture capitalist, venture fund*}. The key term lists of statistically related topics offer an immediate narrative for each numerical time series that we study. These are not causal narratives. In some cases they may reflect proximate causes, and in others they may represent anticipated impacts.

Our second contribution is to show that news text contains distinct or incremental information versus standard numerical indicators. We consider two important applications of our topic model: macroeconomic vector autoregression (VAR) and stock market timing. First, we study news attention within an otherwise standard macroeconomic VAR. We find news attention to the "recession" topic has economically large and highly significant predictive power for future output and employment, after controlling for a variety of common VAR components such as stock prices, interest rates, and measures of economic uncertainty. That is, news coverage of recession risk captures distinct useful information about future economic outcomes above and beyond commonly used indicators. This relationship is robust and also holds in a much longer sample stretching back to 1890 for which we have only *WSJ* frontpage text.

Going beyond a single news topic in the VAR poses an overfitting challenge, which we tackle. We seek a VAR that includes any news topics that genuinely influence macroeconomic dynamics and avoids spurious inclusion of irrelevant topics. This is a model selection problem, which we solve with cross-validated group-lasso regression. Our "text-augmented VAR" uses the group-lasso (Yuan and Lin, 2006) to select or remove an explanatory variable in its entirety (rather than selecting, say, only the third lag of a given variable). We find that the optimal cross-validated model selects a single news attention topic, "recession," for inclusion in the VAR.

Modeling economic activity in terms of news-based topics offers economists a new set of tools for understanding the drivers of economic fluctuations. In a standard macroeconomic VAR, interpretation often boils down to a choice of rotation for the error covariance matrix and the resulting impulse responses. With a text-augmented VAR, fluctuations can be directly mapped to textual narratives of economic conditions. We describe an approach to *narrative retrieval* that combines estimated VAR coefficients with topic model estimates and does not require identifying restrictions or economic constraints. It locates the most influential individual articles—i.e., specific narratives—for interpreting the behavior of model-based expectations. For example, we find that large shifts in output growth expectations are associated with news article headlines such as "Consumer Confidence Slides on Fears of Layoffs" and "Stocks Fell Further Amid Concerns Prices Don't Fully Reflect Worsening Global Growth." Using the topic model as a narrative retrieval device, the researcher can map changes in model forecasts to nuanced verbal interpretations of economic events (i.e., news articles). In essence, the estimated model flags articles that the researcher should read thoroughly, without requiring a close manual read of every article in the *WSJ*.

In a second application, we investigate whether news text, as summarized by our topic model, helps detect time variation in the equity risk premium (i.e., the expected excess return of the stock market). Following standard approaches in the literature, we estimate time-varying expected returns with return forecasting regressions. Forecasting with topic attention requires care to avoid look-ahead bias that might arise from LDA (which is estimated from the full sample of text). To avoid such bias, we use Online LDA (Hoffman et al., 2010), which processes documents in sequential batches. This means that Online LDA topic model estimates at a point in time are based only on the text observed up to that point, thus avoiding any look-ahead bias that could pollute subsequent return forecasting regressions.

We conduct out-of-sample return forecasting analysis using LDA topic attention series as predictors. We evaluate the statistical and economic significance of these predictions through the lens of a market timing strategy (Campbell and Thompson, 2008; Kelly et al., forthcoming). The out-of-sample market timing strategy significantly outperforms several natural

benchmarks including a buy-and-hold market strategy, a strategy inspired by *The Economist*'s R-word index, a strategy based on changes in EPU, and two strategies based on the Goyal-Welch predictors. Even though the LDA topics are chosen in an unsupervised manner to reduce the dimensionality of newspaper text, this reduction is so effective at summarizing the information content of news that it generates reliable market return forecasts.

Our analysis is a case study in augmenting economic models to incorporate news attention and retrieve narratives. But these ideas apply to other modeling approaches as well. For example, the implied residuals from an estimated DSGE model can be projected on news attention to extract narratives that aid interpretation of those shocks. More generally, by combining macroeconomic analysis with topic modeling, the researcher accomplishes the task of drawing on vast written text corpora to better understand quantitative economic phenomena. Our results provide a glimpse into the possibilities of using textual data for modeling macroeconomic dynamics or the role of information transmission and media in the macroeconomy. To help facilitate such work, we maintain the interactive website, `www.structureofnews.com`, to allow users to visualize and inspect a wide variety of features from our estimated topic model. In addition, the website allows researchers to download our *WSJ* news attention time series (daily and monthly) for use in their own projects.

Our work contributes to a rapidly growing literature in economics that uses text as data.[3] Topic models have only recently begun to be explored in empirical economics research (the earliest example to our knowledge is Hansen et al., 2017). Papers by Larsen and Thorsrud (2019) and Thorsrud (2020) also apply Latent Dirichlet Allocation (LDA) to Norwegian news data and analyze macroeconomic forecasting models.[4] Complementary research integrates news text into macro-finance analyses using carefully curated researcher inputs in place of statistical models. Chahrour et al. (2021) study the influence of newspapers' sectoral coverage on macroeconomic activity. Rather than directly studying article text, they map article tags from Factiva to firms and their respective sectors in order to measure news coverage. Baker et al. (2016) build indices of economic policy uncertainty from news articles by counting the occurrences of researcher-curated key words, and Baker et al. (2021) analyze the drivers of large stock price moves based on close human reading of business news. In this paper we provide a complete solution for macroeconomic textual analysis to better understand the driving forces of business cycles.[5] This includes i) estimating and selecting an optimal topic model specification from a macroeconomic news corpus, ii) a text-augmented VAR that embeds the topic structure in an otherwise standard empirical macroeconomic model, iii) a selection algorithm

[3] See Gentzkow et al. (2019) for a recent survey.

[4] Subsequent to our paper, Ellingsen et al. (2020) also analyze *WSJ* news for macroeconomic forecasting.

[5] While we use LDA and analyze the full text of all *WSJ* articles, contemporaneous work of Cong et al. (2020) use word2vec methods and Manela and Moreira (2017) use support vector machines, with both of these papers analyzing abstracts of front page articles.

for deciding on the appropriate topics to include in the VAR, and iv) a means of extracting underlying articles that provide narrative interpretations of business cycle fluctuations.

We also contribute to a strand of literature that uses text data to forecast stock market returns. García (2013) shows that news sentiment forecasts stock market returns especially during recessions. Manela and Moreira (2017) show that news-implied volatility forecasts returns and economic disasters. Ke et al. (2019) and Kelly et al. (2021) propose novel machine learning methodologies that convert raw news text into forecasts for stock returns and macroeconomic indicators, respectively. Boudoukh et al. (2018) explain contemporaneous idiosyncratic returns for individual stocks using news text. We contribute to this literature by showing that topic models distill reliable market timing signals from news text. We analyze "online" methods for forecasting with topic models while avoiding look-ahead bias, and for interpreting market timing gains through the lens of economically interpretable narratives.

Section 2 briefly discusses our model structure and estimation. Section 3 describes the estimated structure of economic news underlying *WSJ* text. Section 4 documents the close correspondence of news text and numerical economic data. Section 5 analyzes the role of news text in a macroeconomic VAR. Section 6 analyzes stock market timing with news text, and Section 7 concludes.

2 Topic Model and Estimation

Our model for the structure of economic news text follows the Latent Dirichlet Allocation (LDA) topic modeling approach of Blei et al. (2003). We provide a brief summary of the model and its estimation, and refer interested readers to the original paper for additional detail.

We study the *WSJ* text corpus in its "bag-of-words" form, represented numerically as the article-term matrix, $\mathbf{w}$. This is a $T \times V$ matrix where row indices correspond to the list of T different articles in the corpus, and column indices correspond to the vocabulary of V unique terms in the corpus. Its individual elements, $w_{t,v}$, count the number of times that the v^{th} term appears in article t.

2.1 LDA

While the bag-of-words representation is a dramatic reduction in complexity compared to the raw text, it remains an extraordinarily high-dimensional object. LDA seeks a tractable thematic summary of $\mathbf{w}$ that reduces the dimensionality of the historical *WSJ* to a scale that can be digested and interpreted by a human reader in one sitting. To achieve this summary, it imposes explicit parametric assumptions (a multinomial term count distribution with Dirichlet

priors) and imposes a factor structure on expected term counts. Specifically, LDA assumes that the V-dimensional vector of term counts for a given article t, denoted w_t, are distributed according to a multinomial distribution

$$w_t \sim \text{Mult}(\Phi'\theta_t, N_t), \tag{1}$$

where N_t is the total number of terms in article t and governs the scale of the multinomial distribution. In other words, expected term counts are summarized by a comparatively low dimension set of parameters, θ_t and $\Phi = [\phi_1, ..., \phi_K]'$. The k^{th} "topic" in the text is defined by the V-dimensional parameter vector ϕ_k, where $\phi_{k,v} \geq 0$ for all v and $\sum_v \phi_{k,v} = 1$. That is, a news topic is a probability distribution over terms. The set of terms that take especially high probabilities in ϕ_k convey the thematic content of the topic. The model's dimension reduction is achieved by setting the total number of topics K much smaller than the size of the vocabulary.

While the topics, ϕ_k, describe the common themes in the corpus as a whole, LDA treats an individual article as a mixture of topics. The article-specific parameter vector $\theta_t = (\theta_{t,1}, ..., \theta_{t,K})'$ is also a probability vector, with $\theta_{t,k} \geq 0$ for all k and $\sum_k \theta_{t,k} = 1$. That is, θ_t describes how article t allocates its attention across topics. LDA embeds a factor structure in which topics (ϕ_k) serve as common factors and θ_t captures articles-specific exposure to those factors.

In principle, one could estimate ϕ and θ via maximum likelihood according to (1). However, in most text applications this is computationally unrealistic. Instead, we use Bayesian methods and approach estimation with the collapsed Gibbs sampler proposed by Xiao and Stibor (2010). We describe this approach and how it generates estimates of the topics, $\hat{\phi}_k$, and estimates of attention allocation across topics, $\hat{\theta}_t$, in Appendix B.

2.2 *The Wall Street Journal* Data Set

The data set we use is among the most extensive text corpora of business news studied in the economics literature to date. It consists of all articles published in *WSJ* from January 1984 through June 2017, purchased from the Dow Jones Historical News Archive. This represents the longest history of full text articles in digital text format available for purchase from Dow Jones & Company. As a comparison, Manela and Moreira (2017) extract a news-based analogue to the VIX index using the *WSJ* over a long sample, but their data only includes abstracts of front page articles. Baker et al. (2016) study a large collection of newspapers, but restrict their analysis to counting occurrences of a small, pre-defined key term list. In contrast, our analysis leverages the richness of full newspaper text.

We take a number of steps to homogenize the data sample and reduce confounding effects of organizational changes to the *WSJ* over time. First, the Dow Jones Historical News Archive contains data as far back as 1979, but data prior to 1984 is limited to article abstracts. We therefore omit pre-1984 data to maintain consistency in the definition of what constitutes an article throughout our sample. Next, over its history, *WSJ* has initiated (and sometimes later abandoned) a number of non-core sections such as "Personal Journal" (initiated 2002), "Weekend Journal" (initiated 2005), and "Off Duty" (initiated 2012). To help maintain consistency in topical content over time, we exclude articles appearing in sections other than the three core sections ("Section One," "Marketplace," and "Money and Investing") that are available over our full sample. Because our interest is in economic news, we also exclude articles with subject tags corresponding to predominantly non-economic content such as sports, leisure, and arts, and exclude articles that contain data tables with little supporting text.

Next, we transform the dataset from a collection of raw article text files into numerical term counts for statistical modeling. Our vocabulary of "terms" includes all uni-grams and bi-grams occurring in our data set after applying a mild set of term filters and lemmatization of derivative words. Appendix A provides a step-by-step description of the data processing procedure we use to transform raw article text into term counts.

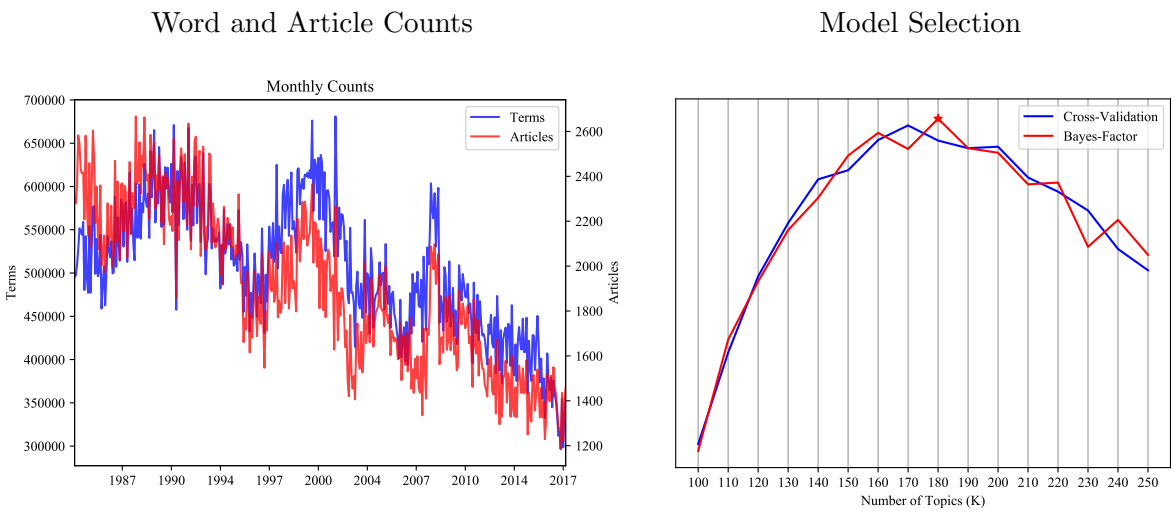
Our final data set consists of 763,887 articles with a vocabulary of 18,432 unique terms. The left panel of Figure 1 shows the size of the data set over time in terms of monthly article count and monthly term count. The structure of the topic model absorbs some of the secular changes in news production by modeling news term proportions rather than term count levels. It is worth noting that, our homogenization efforts notwithstanding, the *WSJ* is an evolving product that undergoes structural shifts over the course of our sample.

2.3 Topic Model Specification Choice

LDA is an unsupervised machine learning technique that requires the econometrician to choose only two inputs: The term count data set and the number of topics, K. We take a data-driven approach to select K. In particular, we estimate a variety of models with K ranging from 50 to 250 topics in increments of ten. We select the model specification from this set as the one with the highest Bayes factor.[6] For robustness we also approach model selection with ten-fold cross-validation. In particular, we partition the *WSJ* sample into ten folds with equal numbers of articles. We re-estimate each K-topic model on the data ten times, one time excluding each of the ten validation samples, then evaluate its fit on the left-out sample. We

[6]The Bayes factor is the ratio of posterior probabilities for the alternative model versus the null model. When selecting among topic models with different K, we compare all models to the same null model so that the denominator is constant across K. Thus, our Bayes factor criterion is equivalent to selecting the model with the highest posterior probability.

Figure 1: *WSJ* Observation Counts and Model Selection



Note. The left panel plots the post-processing article count and the total term counts aggregated over all articles each month. The right panel plots the cross-validated model fit (average log likelihood over ten cross-validation folds) and the Bayes factor for models with topic count K ranging from 50 to 250 in increments of 10 (we omit labels on the right panel vertical axis as the model fit comparisons are relative and specific values are uninformative).

calculate aggregate goodness-of-fit for each model as the average log likelihood value over the ten validation samples.

The right panel of Figure 1 plots goodness-of-fit across candidate specifications. Bayes factor and cross-validation analyses indicate that a 170–180 topic specification approximately optimizes our specification criteria. Manual inspection of estimates for various choices of K also indicates that $K = 180$ is a sensible choice. A model with fewer topics (e.g., 50) produces mixed topics that contain multiple separate themes. A larger model (e.g., 250 topics) delivers similar interpretability as that for the 180-topic model, but with a number of overly specific topics that capture one-off events. Finally, all models in the neighborhood of $K = 180$ look very similar in their topic composition. Appendix F provides further visual evidence that the 180-topic model strikes a balance between interpretability and parsimony.

2.4 Online LDA for Avoiding Look-ahead Bias

In many analyses common in finance and economics, researchers wish to quantify the degree to which variables observable at a point in time can forecast future outcomes. A canonical example is market timing—forecasting future stock market returns using only information available in a prior period. For example, below we study one such application, and find that

topic attention observed in month t can forecast stock market returns at $t+1$.

A common concern in forecasting analysis is look-ahead bias. If the variables used for prediction are constructed using information available only in future periods, then forecasting regression estimates could be biased. In the LDA topic model context, if we use the entire sample to estimate the topics Φ and attention allocation across topics θ_t, then $\hat{\theta}_t$ is informed in part by term counts that realize after time t.[7]

To overcome this bias, we turn to an Online LDA (oLDA) estimator from Hoffman et al. (2010) that constructs topic model estimates in an entirely backward-looking fashion. While oLDA was originally developed to overcome computational challenges associated with datasets too large to fit in memory, we take advantage of its ability to process text data that arrives sequentially, and emphasize this advantage for out-of-sample forecasting analysis.

The key benefit of oLDA for our purposes is that it processes documents in sequential batches. Therefore, time t topic attention estimates are based only on data from periods 1 through t. As a result, time t topic attention is a valid input for forming out-of-sample forecasts of $t+1$ outcomes without look-ahead bias.[8] Moreover, unlike rolling re-estimation which faces the issue that topics in one rolling sample are delinked from topics in a prior sample, oLDA preserves a coherent topic interpretation over time (while accommodating evolving topics). We refer readers to Hoffman et al. (2010) for additional detail.

3 The Structure of Economic News

In this section we dissect the *WSJ* news topic model estimates. We focus here on the 180-topic specification based on a single estimation using the full sample, and discuss oLDA estimates in Sections 5 and 6. Note that unlike other dimension reduction techniques such as principal component analysis, LDA does not provide a natural ordering of topics, thus our presentation of topics is based on expositional convenience. It is also important to recognize that due to the richness of information contained in news text, our 180-topic model is higher dimensional than typical economic models, and it is impossible to report the full scope of estimates in this article. To give readers the ability to explore all facets of our model, we have built an interactive website, `www.structureofnews.com`, to allow users to inspect our estimated topic model in fine detail.

[7] A standard finance analogy is using PCA to estimate return factors from a matrix R of N stock returns over T periods. The factor estimates for any date $t < T$ are indirectly informed by data after date t because the PCA factor loadings use information from the full sample.

[8] To estimate oLDA we use the *gensim* python package. *gensim* uses an alternative estimation procedure—variational inference as opposed to Gibbs sampling, however, manual inspection of the resulting topics suggests that both estimation procedures yield comparable topic attention.

3.1 Topic Key Terms

First, we present estimates for each term cluster—that is, the $V \times 1$ term probability vector ϕ_k—that defines a given topic k. The most common terms in the vocabulary appear with high frequency in many topics, and this is naturally reflected in LDA estimates. In order to best identify the *unique* semantic content of each topic, we rescale the topic-term weights by the inverse of the term-frequency, f_v:

$$\tilde{\phi}_{k,v} = \frac{\hat{\phi}_{k,v}}{f_v}.$$

This scaling emphasizes terms that have an unusually large weight in topic k (terms like "pharmaceutical" or "iron ore") and downplays words that are common to the corpus overall (terms like "price" and "company"). Sorting the elements of $\tilde{\phi}_k$ identifies the terms that are most diagnostic of the thematic content in topic k.

Appendix Table 7 lists the *key terms* for each topic k, defined as the top ten vocabulary items based on the sorted $\tilde{\phi}_k$ vector. These key terms are entirely unsupervised as they are estimated from the data without any guidance from article labels or from the researcher. We manually assign a label to each topic based on our reading of the key term lists. Topic labels serve as a shorthand for referencing topics throughout the paper.

There are several features of the key term lists that stand out. First, they show that topics represent coherent concepts with clear interpretability. For example, the first key term list from the table is

> *bonus, base salary, total compensation, pay package, compensation package, compensation committee, restrict stock, tyco, executive compensation, bonuses.*

This list is easily recognized as an "executive pay" topic. The list is un-mixed; each term has a direct narrative link to the topic label. The purity of the topic extends beyond the top ten terms. For example, key terms 11-25 for this topic are

> *stockoption, executive pay, stock option, compensation, annual salary, tyco international, exercise option, severance package, severance payment, exercise price, salary, retention, vest, proxy statement, severance.*

This term purity is representative of our estimated model more broadly. No topics appear to mix themes. To draw a contrast, when we estimate a smaller model with $K = 50$ topics, we find topics that begin to lump multiple concepts together. For example, the 50-topic model generates the following key term list for one of its topics:

creditor lawsuit, unsecured creditor, ual corps, chapter company, texas air, bankruptcy court, federal bankruptcy, bankruptcy code, continental airline, usair, load factor, flight attendant, reorganization plan, american airline, northwest airline, chapter file, amr corps, bankruptcy protection, major airline, corps unit, airline unit, pilot association, chapter bankruptcy, world airways, twa.

Evidently, this topic combines two distinct topics—"airlines" and "bankruptcy"—that we identify in the $K = 180$ specification. While this mixing is understandable given the string of airline bankruptcies in our sample period, mixing confounds the model's ability to identify important non-bankruptcy airline news such as sharp drops in flight demand following terrorist events. Examples like this abound in smaller models.[9] The general absence of mixing in the $K = 180$ specification is a validation that the larger model achieves a successful separation of distinct subjects.

Second, we see that most topics represent news subjects, as opposed to appraisals of good versus bad news for a given subject. Examples of appraisal-free subject topics include "airlines," "Federal Reserve," and "China." Some subject topics carry an implicit appraisal, usually of bad news, such as "terrorism," "natural disasters," and "recession." A small number of sentimental topics also appear, such as the "problems," "concerns," and "positive sentiment" topics. We interpret these as "modifier" topics that, when combined in a news article with a given subject topic, convey a signed or directional appraisal of the subject's current events.[10]

A larger K results in more specific subject categories. An advantage of having fine-grained topics is that, if a broader notion of a topic is desired, finer topics can be agglomeratized into broader metatopics. In fact, we find that our 180 topics cluster into an intuitive hierarchy of increasingly broad metatopics. Figure 2 illustrates the topical hierarchy in our model in the form of a dendrogram. We estimate this hierarchy with recursive agglomeration (Murtagh and Legendre, 2014) based on the semantic distance between topics (defined as the distance in their $\hat{\phi}_k$ vectors). Like the estimation of topics themselves, the connectivity of topics in the

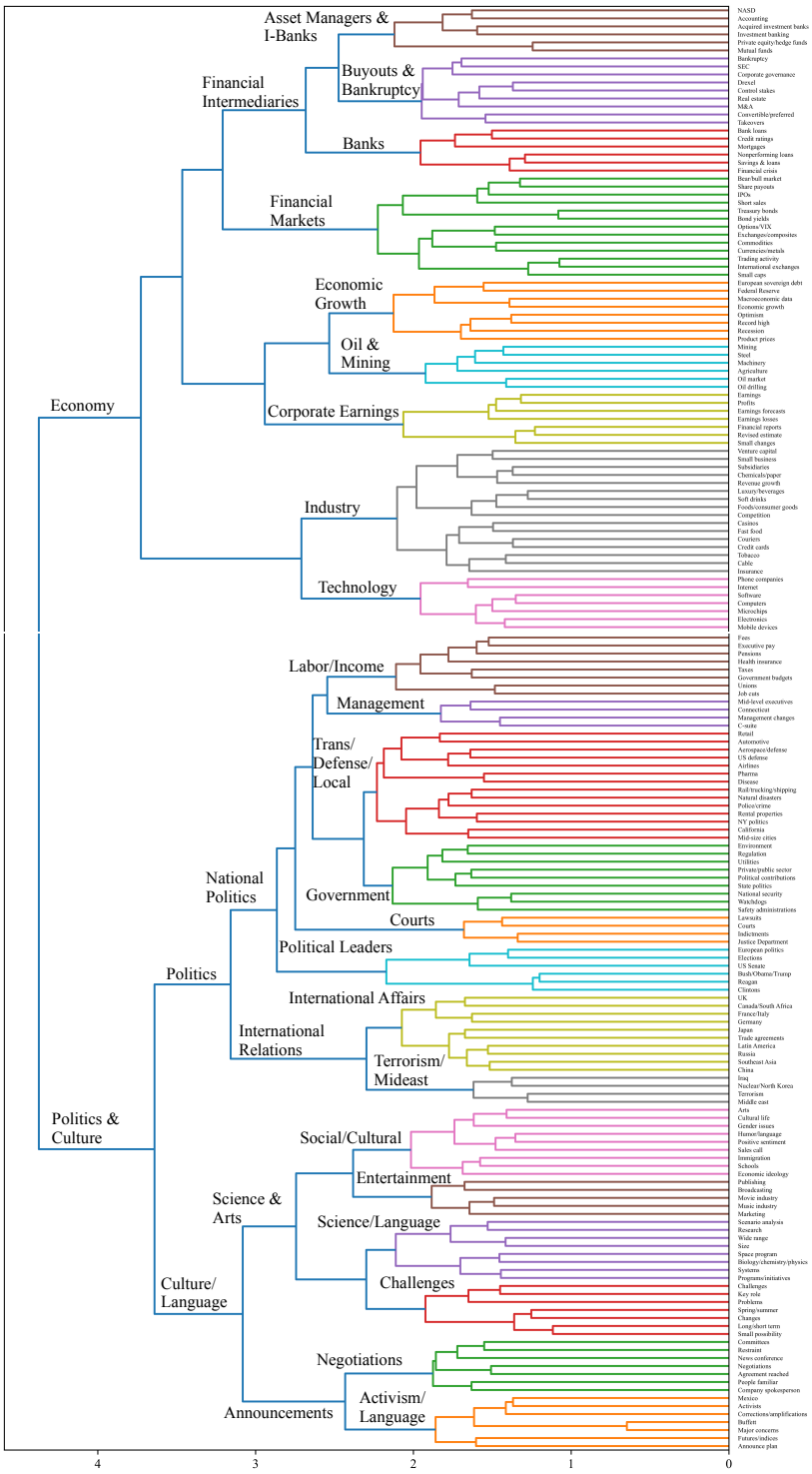
[9] Another example in the $K = 50$ specification is a topic with key terms

ipo market, sachs group, ipo, csfb, stearns cos, berkshire, buffett, berkshire hathaway, suisse group, ipo price, brother hold, investmentbanking business, stanley group, goldman, hathaway, group credit, firstday, weill, witter discover, investmentbanking, stanley dean, goldman sachs, sachs, warren buffett, credit suisse,

which is a mixture of the "IPO," "investment banking," and "Buffett" topics in the $K = 180$ model.

[10] There is also a very small number of topics that represent shifts in language usage over time, but do not have any economic meaning. For example, there is a "corrections/amplifications" topic {key terms: *article correction, amplifications, incorrect, misstated*} and a "news conference" topic {key terms: *press conference, told reporter, apology, issue statement*}. Both of these have a clear positive time trend. Language topics are valuable for de-noising news text by absorbing common variation in language that is distinct from genuine news content, and that would otherwise be absorbed into a content topic and act as a source of noise in our subsequent analysis.

Figure 2: Hierarchical Taxonomy of *WSJ* News



Note. Hierarchical agglomerative clustering dendrogram based on $\hat{\phi}_k$ similarity among the 180 topics listed on the right.

dendrogram is entirely data driven, and we only supply metatopic labels based on our manual read of the clusters. While this figure focuses on the high level metatopic hierarchy, interested readers can find the metatopic assignment of all 180 individual topics in Appendix Table 7.

Figure 2 provides a full taxonomy of *WSJ* news. At the broadest level, news is classified into either "economy" topics (top half of the dendrogram) or "politics and culture" (bottom half) topics. Within "economy," topics split into broad metatopics such as "financial intermediaries," "economic growth," and "industry." The "politics and culture" branch includes topic clusters such as "international relations," "national politics," and "science and arts." As emphasized by Quinn et al. (2010), an intuitive metatopic hierarchy is a useful check on the semantic validity of a model. Furthermore, it gives us the ability to analyze attention to news topics at various levels of granularity.

3.2 Quantifying News Attention

With an understanding of the topic structure in place, we turn to the second main output of the model—the estimate of news attention paid to each topic. Attention allocation is a convenient quantitative transformation of news composition that can serve as numerical input to subsequent empirical investigations of economic hypotheses. Our estimates describe how allocation of media attention across topics evolves over time.

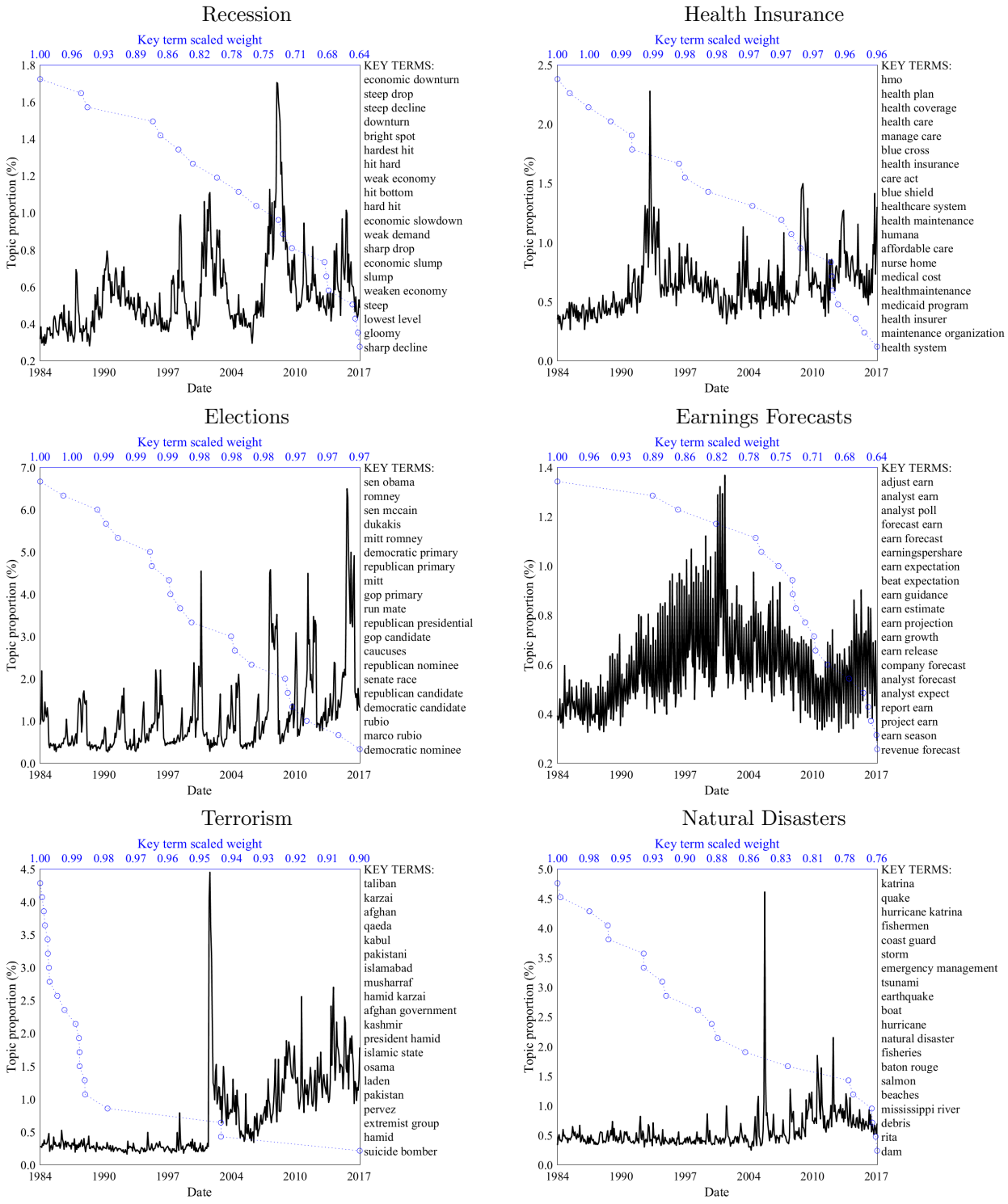
3.2.1 News Attention by Month

Topic attention is estimated at the article level but can be summed to any level of aggregation according to Equation (6). We focus on aggregate *WSJ* topic attention at the monthly frequency by summing attention estimates over all articles published in the same calendar month. Figure 3 shows the time series variation in attention for a subset of six illustrative topics.[11] The black line, corresponding to the left vertical axis, shows the attention to a given topic as a percent of total monthly *WSJ* news production. The right vertical axis lists the key terms, and the blue line shows the weight of each term in the topic (scaled by the maximum term weight within that topic), corresponding to the top horizontal axis.

The topic attention plots highlight a number of stylized facts about the composition of business news. First, news attention is generally persistent. This is clearly evident in the top two topics, "recession" and "health insurance," which exhibit prolonged waves of high and low attention. "Recession" is a prime example of a recurrent topic that is active throughout the sample. As is the "health insurance" topic, which becomes highly prevalent during debate of the Clinton Health Plan proposal (peaking around President Clinton's speech to Congress

[11] Appendix Figure 12 plots an additional 24 illustrative topic attention time series, and the website for this research project, www.structureofnews.com, includes time series plots for all 180 topics.

Figure 3: Topic Attention Proportions



Note. The black line shows the topic attention as a percent of total monthly *WSJ* news production. The right vertical axis lists the topic's key terms, and the blue line shows the weight of each term in the topic (scaled by the maximum term weight within that topic).

in September 1993), the proposal and passage of Obamacare in 2008–2010, and around the debated repealing of Obamacare around the 2016 presidential election. The middle two plots in Figure 3 are examples of seasonal topics. "Elections" coverage has a clockwork pattern peaking every four years with a secondary peak every two years, and the "earnings forecasts" topic which spikes ahead of each quarterly earnings announcement season. The last two plots are examples of emergent topics. Usually dormant, they are punctuated by intense coverage around particular events. The "terrorism" topic is a prototypical regime shift: It receives minute attention for the first half of our sample, spikes dramatically at 9/11, and remains high henceforth. Similarly, the "natural disasters" topic draws very little attention during most of the sample, but rises sharply in August 2005 (Hurricane Katrina), then reverts but remains slightly elevated for much of the remaining sample, presumably due to increased news attention on climate change (Engle et al., 2019).

With 180 topics, it is difficult to isolate the focus of attention at each point in time in a single line plot overlaying many time series. To make these news attention series easier to grasp, in Appendix D we report the dominant topics in each month and in Appendix E we plot attention at the meta topic level.

4 Matching News Attention To Economic Activity

We next validate our topic model estimates by investigating how faithfully the topical content of *WSJ* news captures the state of the economy. We present a descriptive analysis documenting the high correlation between news text and a wide range of numerical economic time series.

4.1 Selection Via Lasso Regression

As a first step toward establishing that news attention beneficially measures the state of the economy, we document that diverse numerical measurements of economic activity all share a close correspondence with thematically related topical content of news text.

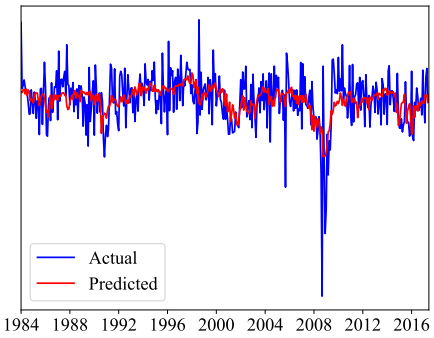
To this end, we devise a generic means of evaluating the alignment between a candidate numerical measure of economic activity, x_t, and the set of estimated topic attention proportions, $\hat{\theta}_t = (\hat{\theta}_{1,t}, ..., \hat{\theta}_{K,t})$. Our criteria for this analysis are two-fold. First, we evaluate accuracy of the text-based explanation in terms of regression fit, and second, we evaluate whether the most influential news topics in the regression are thematically related to x_t.

With $K = 180$ topics, the attention vector is high-dimensional relative to the number of monthly time series observations, so OLS regression of x_t on $\hat{\theta}_t$ will be difficult to interpret and suffer from overfit. Therefore we use lasso regression and set the lasso penalty parameter such that exactly five out of the 180 coefficients take non-zero values. We choose a fixed five-variable

Table 1: Reconstructing Macroeconomic Time Series

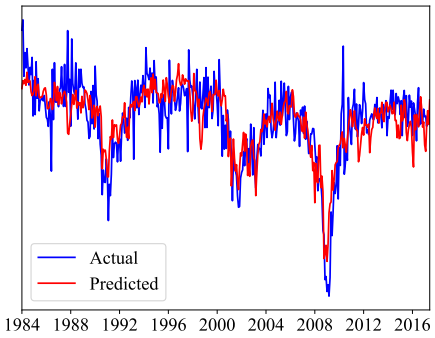
Industrial Production Growth

Topic	Coeff.	p-val.
Recession	-0.38	0.00
Oil market	-0.17	0.00
Southeast Asia	0.11	0.10
Health insurance	0.06	0.93
Clintons	0.03	0.40
R^2	0.21	



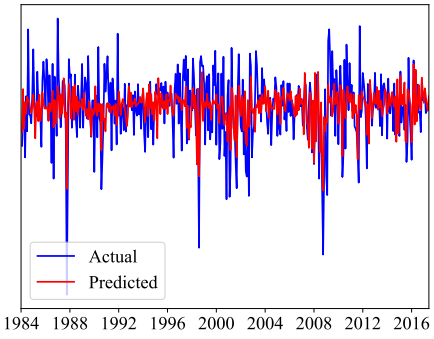
Employment Growth

Topic	Coeff.	p-val.
Recession	-0.61	0.00
Rail/trucking/shipping	0.22	0.01
Bush/Obama/Trump	-0.15	0.09
Iraq	-0.14	0.01
Clintons	0.12	0.01
R^2	0.59	



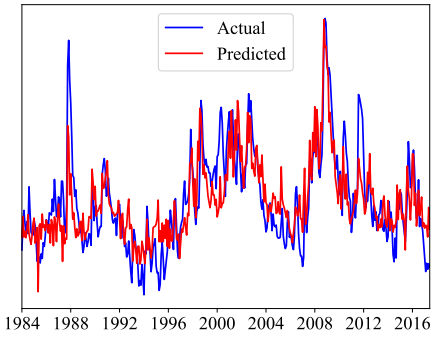
Market Returns

Topic	Coeff.	p-val.
Recession	-0.35	0.00
Problems	-0.19	0.00
Convertible/preferred	-0.16	0.21
Record high	0.12	0.08
Options/VIX	-0.11	0.76
R^2	0.25	



Market Volatility

Topic	Coeff.	p-val.
Recession	0.48	0.00
Options/VIX	0.24	0.00
Problems	0.22	0.00
Small business	-0.18	0.00
Investment banking	0.13	0.00
R^2	0.63	



Note. Five-regressor active-set regression estimates based on lasso selection with p-values adjusted for post-selection inference.

regression to ensure clear interpretability of results, to have uniformity and comparability across the different x variables that we study, and to limit the undue influence of data mining on our conclusions.

Because lasso penalization serves in part as a model selection device, standard errors on

the non-zero coefficients must be adjusted to account for model search. Our reported p-values use the post-selection inference procedure of Tibshirani et al. (2016). We scale regression coefficients to reflect the dependent variable's response in standard deviation units to a one standard deviation change in the regressor.

To assess the breadth of economic summary achieved by studying news text, we examine a diverse set of numerical economic data series. The gamut of variables that we study includes macroeconomic aggregates, corporate financing activities, measures of industry-level risks, and uncertainty about a range of economic policies.

4.2 Macroeconomic Aggregates

We begin by studying monthly data on aggregate output, employment, stock market returns, and stock market volatility. The top left of Table 1 reports results for the lasso regression of log industrial production growth on the 180 news attention time series. Among the five selected topics, news attention to the "recession" and "oil market" topics significantly negatively correlates with output growth. The coefficient estimates are economically large. A one standard deviation increases in "recession" attention associates with a 0.38 standard deviation drop in output growth.

In fact, "recession" attention is by far the strongest explanatory variable for all four economic aggregates in Table 1. A unit standard deviation rise in "recession" attention associates with a 0.61 standard deviation drop in log employment (non-farm payrolls) growth. Employment growth also has a significant positive coefficient on "rail/trucking/shipping" attention and a negative coefficient on "Iraq" attention, though the impact of these topics is small compared to "recession" attention.[12]

The lower left panel regresses returns on the value-weighted US stock market index on AR(1) innovations in topic attention.[13] "Recession" attention is the largest determinant of market returns with a coefficient of -0.35. The second largest effect comes through attention to the "problems" topic, which describes adverse conditions more generally {key terms: *problem face, biggest problem, big problem, matter worse*}.

"Recession" and "problems" attention are also significant positive drivers of market volatility, defined as the realized standard deviation of daily market returns within each month. They both enter with large positive coefficients, reflecting the strongly countercyclical nature

[12]The lasso regression juxtaposes the employment experience under various presidential administrations, though these are insignificant effects at the 5% level. This result also shows how the choice of topic number can influence second-stage economic analysis. A finer decomposition of the *WSJ* into 300 topics produces separate topics for each presidential administration, while a coarser decomposition of 50 topics groups all administrations together.

[13]Market returns have close to zero serial correlation and theoretically behave like a martingale difference sequence, making it natural to likewise represent the regressors as time series innovations.

Table 2: Reconstructing Stock Market Returns

Topic Model & CFNAI			Topic Model & FRED-MD		
Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
Recession	-0.35	0.00	Recession	-0.36	0.00
Problems	-0.19	0.00	Problems	-0.18	0.00
Convertible/preferred	-0.16	0.20	Convertible/preferred	-0.14	0.24
Record high	0.12	0.08	Record high	0.13	0.07
Options/VIX	-0.11	0.76	IPNCONGD (FRED-MD)	-0.10	0.98
Full R^2	0.25		Full R^2	0.25	
Benchmark R^2	0.02		Benchmark R^2	0.09	

Note. Five-regressor active-set regression estimates based on lasso selection with p-values adjusted for post-selection inference. Full R^2 corresponds to the fit with both topics and benchmark variables. Benchmark R^2 corresponds to the fit with only benchmark variables.

of market volatility. It also shows that "small business" topic tends to associate with low volatility regimes in our sample.

For all four macroeconomic variables, five news topics explain a significant fraction of their variation measured by R^2. News attention explains 25% of the variation in stock market fluctuations. To illustrate, Table 2 compares news attention to other business cycle leading indicators. First we regress stock returns on AR(1) innovations in the four sub-indices of the Chicago Fed National Activity Index (CFNAI). These indices measure production/income, employment/unemployment/hours, personal consumption/housing, and sales/orders/inventory. CFNAI indices jointly explain only 2% of the variation in contemporaneous monthly stock returns. Table 2 also shows that when we include these along with the news attention series in our five-regressor lasso model, none of the CFNAI indices are selected.

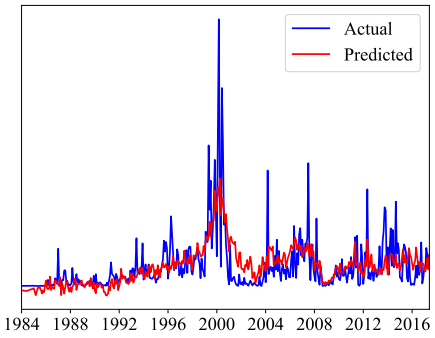
As another benchmark, we regress market returns on AR(1) innovations in the 101 macroeconomic variables in the FRED-MD database.[14] To remain comparable with our news attention regressions, we likewise impose a lasso penalty to select exactly five non-zero coefficients among the FRED-MD variables. We find an R^2 of 9%, which exceeds that from the CFNAI regression, but is less than half of the news-based result.[15] When we include the 101 FRED-MD series alongside the news attention series in the lasso model, only one FRED-MD variable is selected, industrial production growth of non-durable consumption goods (IPNCONGD),

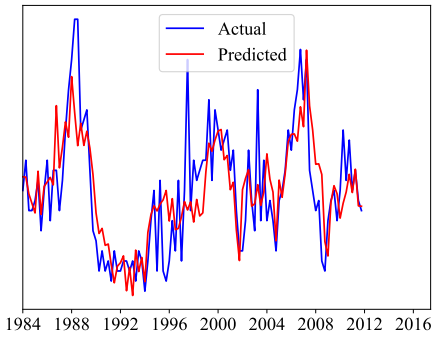
[14] Available at https://research.stlouisfed.org/econ/mccracken/fred-databases/. We exclude FRED-MD interest rate and exchange rate variables. These are asset prices and have mechanical links to stock market returns.

[15] This result is robust to a variety of regression specifications. The explained variation in returns is essentially unchanged if we use other constructions of macro variable innovations, or if we control for lagged as well as contemporaneous values. Even if we include interest rate and exchange rate variables, the R^2 rises only to 17%.

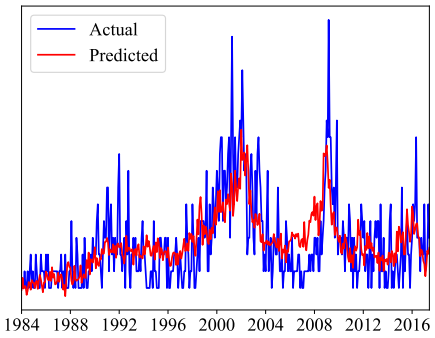
Table 3: Reconstructing Financial Activity

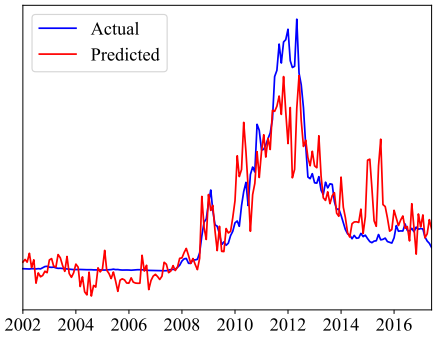
Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
	IPO Volume			LBO Volume	
IPOs	0.26	0.00	Takeovers	0.31	0.28
Venture cap.	0.22	0.00	Insurance	-0.28	0.02
Bankruptcy	-0.14	0.25	Control stakes	0.27	0.03
Internet	0.13	0.08	Key role	0.23	0.70
M&A	0.09	0.72	Job cuts	-0.18	0.13
R^2	0.44		R^2	0.58	





	Bankrupty Filings			European CDS Spreads	
Recession	0.48	0.00	Euro. sov. debt	0.58	0.00
Venture cap.	0.25	0.01	Govt. budgets	0.17	0.61
Accounting	0.24	0.01	Middle east	0.16	0.08
Small caps	0.12	0.00	Mobile devices	0.09	0.06
Machinery	-0.06	0.31	News conference	0.07	0.98
R^2	0.42		R^2	0.78	





Note. Five-regressor active-set regression estimates based on lasso selection with p-values adjusted for post-selection inference.

though it is insignificant and enters with a counter-intuitive negative sign.

4.3 Financing Activity

As a second and more specific example, we analyze how news attention tracks fluctuations in financing activity. We choose x_t variables that help illustrate how specific news topics line up with detailed differences across economic activities.

First, we study monthly dollar volumes of IPOs and LBOs. Table 3 shows that news attention closely tracks fluctuations in both series with an R^2 of 44% and 58%, respectively. More interestingly, the news topics that associate most strongly with each financing variable are also those with the most similar thematic content. The three significant topics in the IPO volume regression are "IPOs," "venture capital" which is the typical form of financing prior to IPO, and "internet" which is the industry that saw the greatest IPO activity during our sample period. For LBO volume, "takeovers," "control stakes," "key role," and "job cuts" topics capture key LBO concepts such as transition from public market to private ownership and control, management changes ("key role"), and eliminating jobs to reduce costs and improve production efficiency.

Table 3 also looks at the credit risk side of financing, measured first as the monthly count of bankruptcies among US public companies. News attention explains variation in bankruptcy intensity with an R^2 of 42%, and the four significantly associated topics ("recession," "accounting," "venture capital," and "small caps") are all themes that commonly coincide with debt default. The last panel introduces an international dimension to credit risk in the form of average CDS spreads among Eurozone sovereigns (available beginning in 2002). A one standard deviation rise in the lone significant news topic, "European sovereign debt," associates with a 0.58 standard deviation increase in European sovereign credit spreads. The time series plot of actual spreads versus spreads reconstructed from news shows the rapid rise and fall of CDS spreads. This data series amounts to only a handful of distinct observations around the financial crisis. Yet despite this, estimated news topic attention is informative enough that lasso regression hones in on the correct narrative account of these events. While other news topics related to financial crises and subprime mortgages follow similar time series patterns, the penalized regression bypasses these in favor of the European sovereign debt narrative.

4.4 Industry Volatility

Our next example investigates how well news attention accounts for patterns in industry-level stock volatility. We use the 49 industry categories from Ken French's website, and measure monthly volatility as the standard deviation of daily industry returns within the month. All equity portfolios tend to share a large common time series component associated with overall market volatility (Herskovic et al., 2016). To hone in on industry-specific volatility

Table 4: Reconstructing Industry Volatility (Innovations)

Topic	Coeff.	p-val.	Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
Automotive			Banking			Pharmaceuticals		
Automotive	0.16	0.03	Nonperform. loans	0.26	0.00	Clintons	0.20	0.01
Fees	0.14	0.11	Mortgages	0.20	0.00	News conf.	-0.14	0.48
Mutual funds	-0.13	0.33	Options/VIX	-0.18	0.00	Pharma	0.14	0.10
Corporate govt.	0.12	0.59	Chemicals/paper	-0.15	0.10	Earnings fcts.	0.12	0.35
Financial crisis	-0.08	0.87	NASD	-0.09	0.76	Health ins.	0.10	0.62
R^2	0.10		R^2	0.17		R^2	0.11	
Computer Hardware			Oil and Gas			Tobacco		
Retail	0.20	0.00	Oil market	0.34	0.00	Tobacco	0.17	0.34
Computers	0.17	0.01	Earnings losses	-0.15	0.05	Problems	-0.10	0.43
Software	0.10	0.18	Options/VIX	-0.15	0.87	Mexico	0.10	0.98
Fees	-0.09	0.72	Cultural life	0.14	0.08	Terrorism	-0.09	0.41
Russia	-0.08	0.93	Elections	-0.13	0.46	Lawsuits	0.09	0.59
R^2	0.09		R^2	0.21		R^2	0.08	

Note. Five-regressor active-set regression estimates based on lasso selection with p-values adjusted for post-selection inference.

patterns that are distinct from the aggregate market volatility analysis in Table 1, we perform two adjustments to the raw industry volatility data. First, we orthogonalize each industry volatility series against the first principal component of the industry volatility panel. Then, we construct the AR(1) innovations in the adjusted series.[16]

Table 4 reports results from five-regressor lasso regressions of industry volatility innovations on AR(1) innovations in news topic attention. For the sake of brevity, the table reports results for nine of the 49 industries that best illustrate how specific aspects of news attention align with specific industry behaviors. Indeed we find that the influential topics in each regression are close thematic counterparts for the respective industry. For example, banking sector volatility is highest amid news attention to "nonperforming loans" and "mortgages" topics. Volatility in the computer hardware industry (which includes firms like Apple, Dell, and Hewlett Packard) is most tightly linked to "retail," "computers," and "software" topic attention. In addition to the "pharma" news topic, volatility in the pharmaceuticals sector is associated with attention to the "Clintons" topic, reflecting the important role of healthcare reform in the policy platforms of both Bill and Hillary Clinton, as well as attention to the

[16]Without the principal component adjustment, we essentially recover the results seen for market volatility in Table 1. While the principal component adjustment is necessary for our conclusions, the results from our analysis are largely unchanged if we use levels of orthogonalized industry volatility rather than AR(1) innovations.

Table 5: Reconstructing Macroeconomic Time Series With Interactions

Industrial Production Growth			Employment Growth		
Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
Recession×Oil market	-0.25	0.00	Inv. banks×Bush/Obama/Trump	-0.28	0.94
Inv. banks×Bush/Obama/Trump	-0.18	0.17	Recession×Fees	-0.22	0.24
Job cuts×Recession	-0.09	0.73	Publishing×Recession	-0.19	0.51
Recession×Corporate governance	-0.07	0.94	Recession×Bankruptcy	-0.17	0.04
Short sales×Recession	-0.02	0.11	Credit ratings×Recession	-0.07	0.60
R^2	0.24		R^2	0.63	

Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference, using individual topic attention regressors as well as their pairwise interactions.

"health insurance" topic. Occasionally we find unintuitive but significant topic assignments such as "cultural life" in the oil and gas industry regression, though these examples are rare.

4.5 Exploring Ultra-high Dimensionality

In the preceding sections, we approach the reconstruction of numerical economic time series with linear models of topic attention time series. The view we propose, that the state of the economy is describable as time-varying configurations of topic attention, is likely to be better captured in regressions that allow for (potentially complex) interactions among topics. We explore this by extending the five-regressor lasso regressions for industrial production and employment in Table 1 to include not only the 180 topic attention time series as regressors, but also all pairwise topic interactions.

Table 5 reports the results. Interestingly, we find that no individual topics are selected for the model. Instead, *all* of the selected regressors are topic interactions, and the majority of them (eight of ten) include the recession topic. These results suggest that macroeconomic outcomes can be accurately captured not just by discussion of economic downturns, but with specific attributes of downturns such as conditions in commodities markets and attention to corporate bankruptcy. More broadly, these results raise the possibility that more flexible non-linear models from the machine learning toolkit, such as tree models which tend to be more adept than regressions for modeling covariate interaction, are likely to be valuable tools for measuring the state of the economy.

5 News Attention and Macroeconomic Dynamics

Thus far we have analyzed contemporaneous correlations between news narratives of economic time series. This helps validate the conjecture that news can be a useful quantitative tool for

summarizing the state of the economy and helps demonstrate the interpretability of quantitative analyses that use topic model estimates. However, it has little to say about whether news text conveys novel information that is distinct from information in standard numerical macroeconomic indicators, or whether news text helps in modeling longer term macroeconomic trajectories.

5.1 Macroeconomic VAR

We investigate this question by studying the role of news attention in a macroeconomic VAR. We build on the five-variable monthly VAR specification studied by Baker et al. (2016, BBD), which includes (in order) their economic policy uncertainty (EPU) index, log value of the S&P 500 index, Federal Reserve funds rate, log employment, and log industrial production. The BBD VAR is a natural starting point for our analysis given that paper's large impact in macroeconomics and given the fact that news text is an input to the EPU index.

In our baseline formulation, we augment the VAR to include news attention to the "recession" topic, based on its central role in the lasso regressions of Table 1, in place of EPU (and we include three lags of all variables).[17] We estimate the VAR using data from 1985–2017.[18] We then plot output and employment impulse response functions for a shock to "recession" news attention.[19] We define a news impulse as a shift in "recession" attention from its 5^{th} percentile to its 95^{th}. We use the same definition of an impulse for EPU in our comparisons below (following the impulse definition in BBD).

Panel A of Figure 4 plots the results. A shock to "recession" news attention generates a 1.99% drop in industrial production after 17 months. Employment declines by 0.92% after 20 months. These effects are highly statistically significant, as shown by the 90% confidence bands.[20] The estimated responses are large in economic magnitude. As a benchmark, Panel B compares our baseline VAR specification (solid black line) to the impulse response for EPU (solid gray line) in the VAR model of BBD.[21] The response of output to EPU (−0.98%) is roughly half as large as the response to "recession" attention, and the employment response

[17] Section 5.5 shows that this is a statistically optimal choice from the full set of topics.

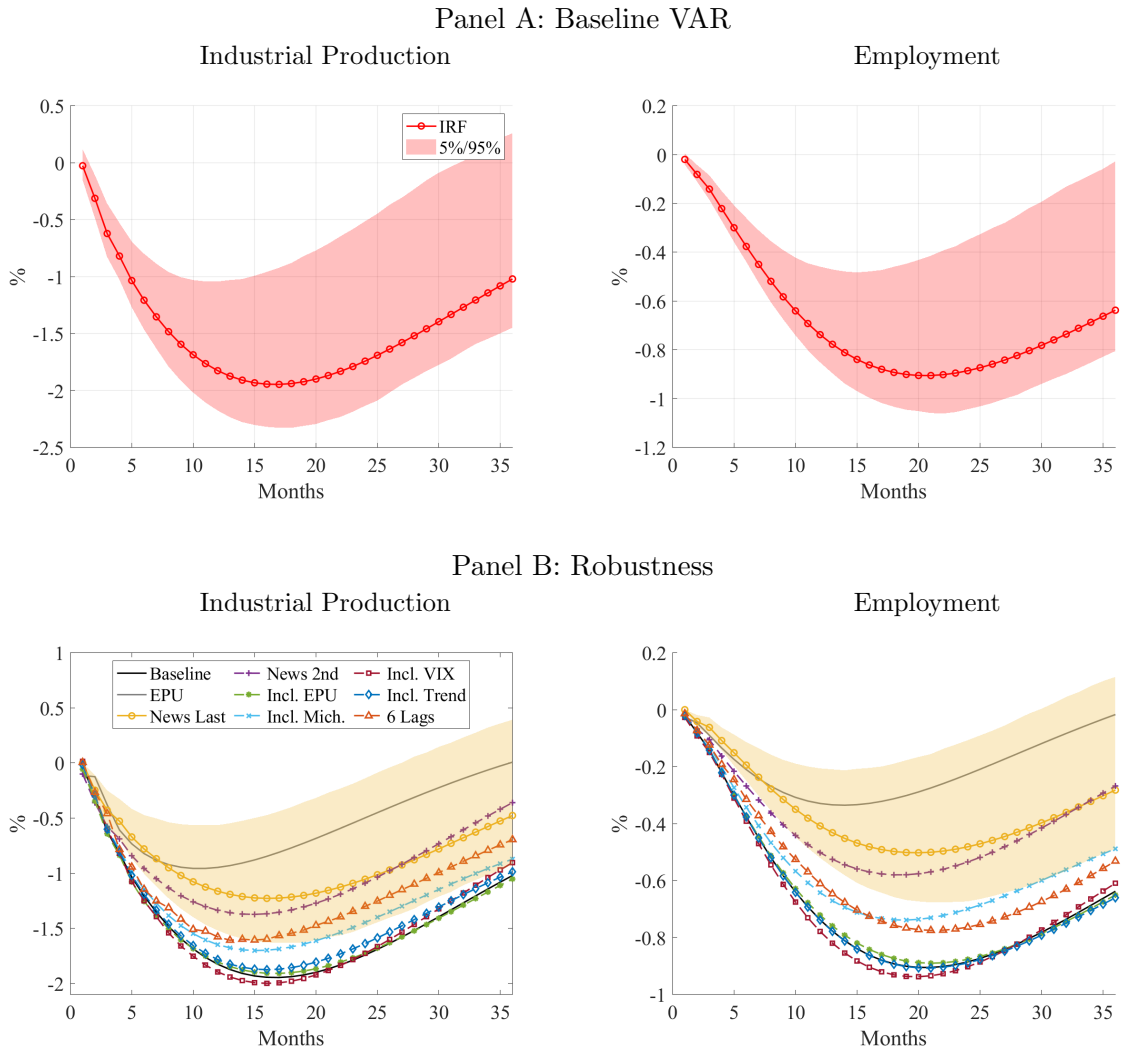
[18] We update BBD's publicly available data (which runs from 1985–2014) through June 2017 to correspond with our *WSJ* sample. We also replace their S&P 500 index data, which is constructed from the average daily index value within the month, with the value of the S&P 500 on the last day of the month. Using average daily values within the month causes additional serial correlation in log changes in the series (an increase from 6% with month-end values to 26% with average daily values). This change has only minor quantitative consequences for our results as well as the results of BBD.

[19] We define orthogonal shocks based on the Choleski decomposition given the BBD variable ordering plus the "recession" topic (ordered first in the vector).

[20] We compute bootstrap confidence intervals following Kilian (1999).

[21] This is the same five-variable specification of BBD, with EPU ordered first, and excluding "recession" attention.

Figure 4: Output and Employment Responses to "Recession" News Attention Shock



Note. Responses of output and employment to a "recession" attention impulse in various VAR specifications, 1985–2017. The red shaded regions in the top plots are 90% confidence bands. This gold shaded regions in the bottom plots are 90% confidence bands when news attention is ordered last in the VAR (corresponding to the gold impulse response curves).

(−0.34%) is roughly one third as large.

Panel B demonstrates the robustness of our baseline VAR conclusions to alternative specifications. First, we consider ordering news attention last in the VAR while preserving the order of other covariates. In this case, the impulse responses reflect the macroeconomic impact of "recession" news coverage after controlling for contemporaneous responses to all other variables. This has a notable impact on our estimates, reducing the maximum output response from −1.99% to −1.25% (shown in gold, surrounded by the 90% confidence band also in gold).

The maximum impact on employment is −0.51%, versus −0.92% in the baseline specification. The effects of "recession" news attention, though smaller when it is ordered last in the VAR, remain highly significant and larger than the impulse responses to EPU.

If we instead alter the order of variables so that "recession" attention is second (after the S&P 500 index but before the Fed funds rate, employment, and industrial production), we see similar impulse responses to the case where news is ordered last. The difference in "baseline" and "news 2nd" specifications highlights that media outlets and financial markets have a partially synchronized response to news arrival (consistent with the high R^2 for the stock market return regression in Table 1).

When we simultaneously control for EPU (which we order second in the VAR after "recession" attention), the "recession" attention impulse responses are essentially unchanged from the baseline. Replacing EPU with the Michigan Consumer Sentiment Index has a small effect on the output response to "recession" news (changing from −1.99% to −1.74%) and a small effect on the employment response (changing from −0.92% to −0.76%). Likewise, accounting for fluctuations in economic uncertainty via inclusion of the VIX index, introducing a time trend, and incorporating six rather than three lags in the VAR leave results qualitatively unchanged.

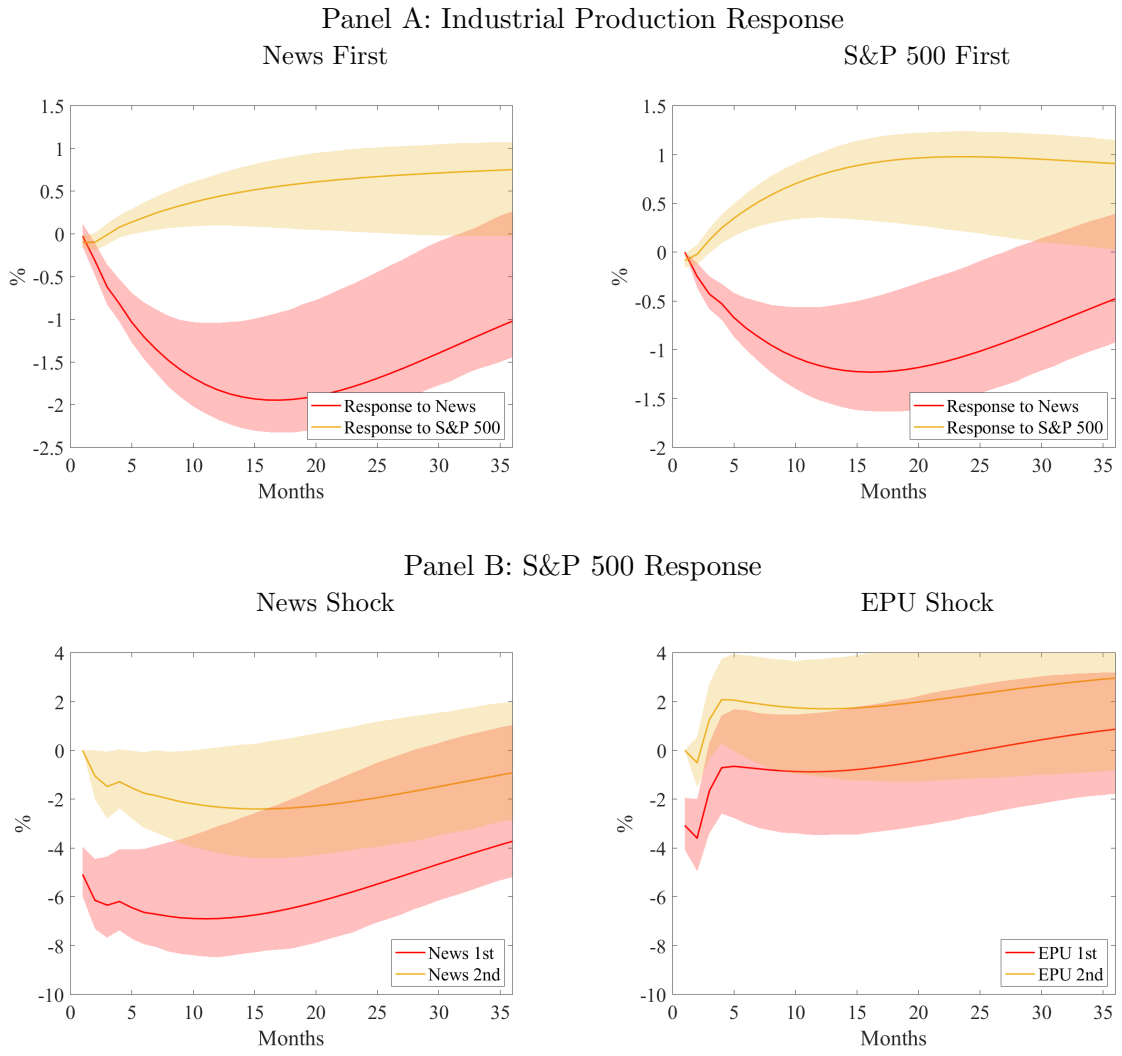
5.2 News Attention and Stock Market Dynamics

A large literature beginning with Fama (1990) argues that stock price fluctuations are driven by changing expectations about future macroeconomic outcomes. Section 4.2 documents a strong contemporaneous correlation between stock market returns and news attention (a correlation unmatched by any FRED-MD macro series). In light of this, a natural interpretation of this high correlation is that news text, like the stock market, reflects agents' expectations of the future macroeconomy, and these expectations are not well captured by time series patterns in numerical macroeconomic data.

An intriguing implication of the VAR analysis is that recession attention and the stock market also contain notably *different* information about future macroeconomic outcomes. The top two panels of Figure 5 compare industrial production impulse responses for shocks to news attention and stock valuations.[22] The top left shows the baseline VAR (with news ordered first and the S&P 500 second) and the top right shows the VAR with S&P 500 ordered first and news second. Both cases tell a similar story: "Recession" attention contains information about the future macroeconomy that is not fully reflected in stock prices. It has a large negative impact on future output, even when ordered after the S&P 500. Moreover, "recession" news has a

[22]The S&P 500 impulse response is defined as the 95^{th} percentile of log changes in the index. In other words, the shock corresponds to a one month stock market return of 6.9%

Figure 5: Stock Market and "Recession" Attention

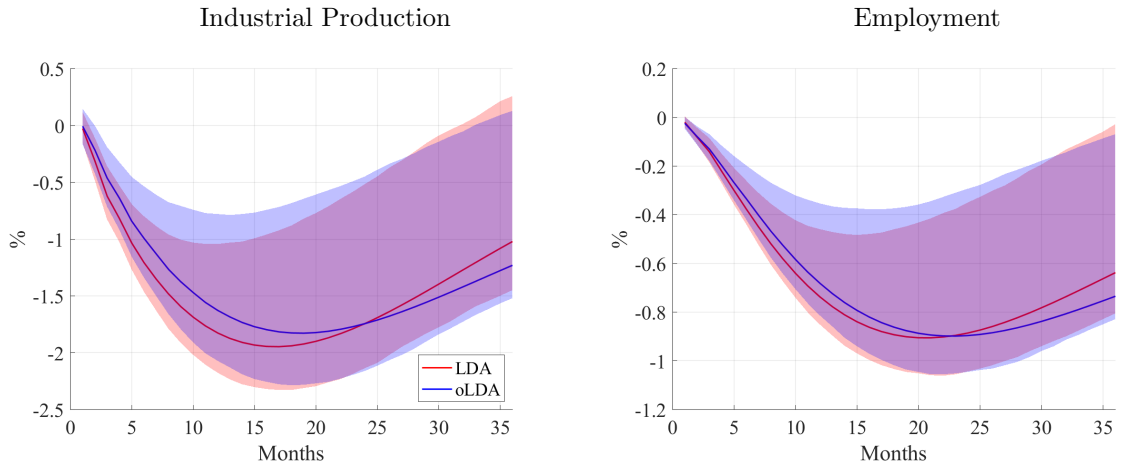


Note. Panel A plots the response of output to an impulses in "recession" attention and stock market value. Panel B reports the response of stock market value to an impulse in "recession" attention in our baseline VAR (left plot) and an impulse in EPU in the BBD VAR (right plot), and includes specifications for impulse variables ordered either before or after the S&P 500 in their respective VARs. Sample is 1985–2017.

stronger association with future output than the stock market does (regardless of ordering), which is a surprising and impressive fact given the Fama (1990) result.

The bottom left panel of Figure 5 further illustrates the dynamic association between recession news and stock prices when news is ordered first. Upon arrival of a positive "recession" attention shock, prices initially drop by about 5%. But this response is incomplete. Over the next year prices drop further, reaching 7% below their pre-shock levels. When news is ordered second, the total response is smaller and borderline insignificant, but still economically large

Figure 6: Responses to "Recession" News Attention Shock: Avoiding Look-ahead Bias with Online LDA



Note. Responses of output and employment to a "recession" attention impulse for LDA vs. oLDA, 1985–2017. The red line represents the LDA point estimates and the red shaded regions are 90% confidence bands. The blue line represents the oLDA point estimates and the blue shaded regions 90% confidence bands.

at -2.4% over 18 months.

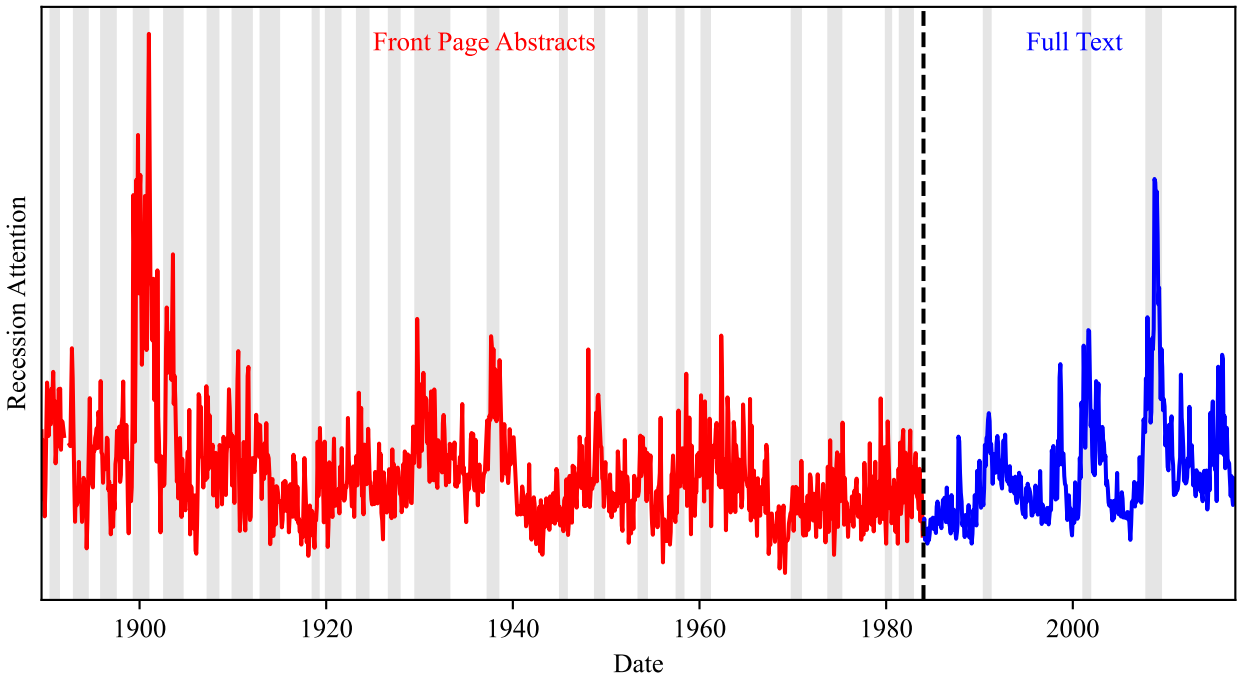
The bottom right panel of Figure 5 helps illustrate the large economic magnitude of stock market responses to "recession" attention by comparing them to EPU impulses based on the VAR in BBD. When EPU is ordered first in the VAR, EPU has a negative and prolonged impact on stock prices, though the magnitude is about half as large as the impulse response to "recession" attention. The differences in these effects are statistically significant at the 90% level. Furthermore, if EPU is ordered second in the VAR (after the S&P 500), the stock market response to EPU is zero at first, then becomes positive.

5.3 Avoiding Look-ahead Bias with Online LDA

Next, we repeat our VAR analysis using the oLDA topic model estimates described in Section 2.4. We extract the topic most similar to the "Recession" topic in our main specification by comparing the top word lists and rerun the VAR exercise with the online version of the "Recession" topic.

Figure 6 reports oLDA (no look-ahead) impulse response estimates and compares them with the full sample LDA impulse responses. We find that the effects of the oLDA "Recession" news attention shock on both industrial production and employment are only slightly attenuated relative to the full sample LDA estimates. For example, using the oLDA topic, output is 1.27 percent lower at 15 months, compared with 1.94 percent using full sample LDA.

Figure 7: Recession Attention and NBER Recessions, 1890–2017



Note. Time series of recession attention and NBER recessions, 1890–2017. The red line corresponds to the frontpage abstracts sample and the blue line the main full text sample. The grey shaded regions correspond to NBER recessions.

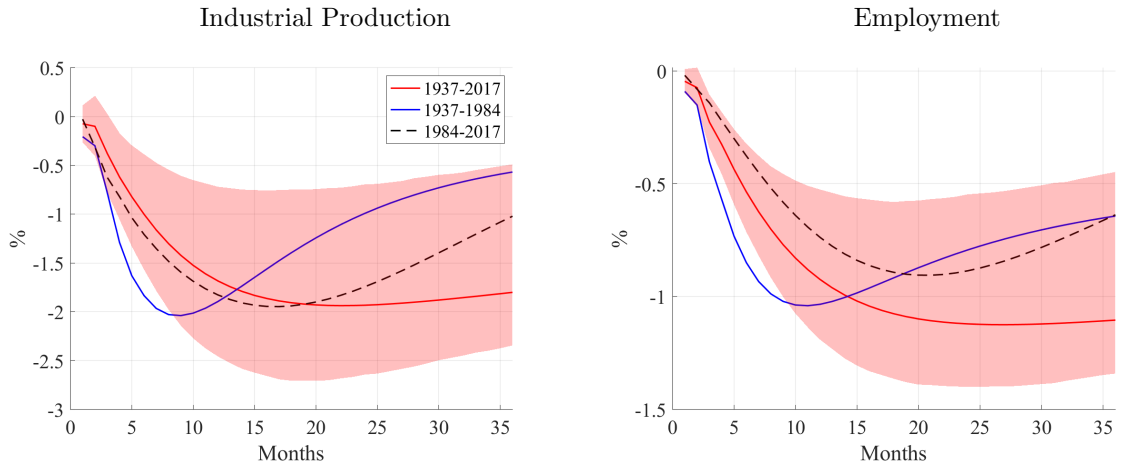
Employment responses at 15 months decline from 0.86 percent with LDA to 0.76 percent with oLDA. This attenuation could result from either a reduction in look-ahead bias, or from the relative noisiness of oLDA estimates based on shorter samples. In any case, the differences are quantitatively minor. From a parameter uncertainty perspective, at short horizons LDA and oLDA estimates are distinct, but at longer horizons they are statistically indistinguishable.

5.4 Over a Century of Business News: 1890–2017

Our analysis thus far starts in 1984, which is when our full text sample starts. But, the *Wall Street Journal* has been in print since 1889. A natural question is whether the conclusions we draw from the 1984–2017 sample extend back to earlier periods.

To answer this question, we leverage the sample of WSJ frontpage titles and abstracts from Manela and Moreira (2017). Prior literature indicates that LDA does not perform well on short documents such as titles and abstracts (Tang et al., 2014). Hence, we use the topic model learned from our main sample (full text articles from 1984–2017) to extend the topic attention series back to 1890.

Figure 8: Responses to "Recession" News Attention Shock: 1939–2017 Sample



Note. Responses of output and employment to a "recession" attention impulse, 1939–2017. We estimate news attention over the 1939–1984 period by applying our rich topic model to the WSJ frontpage abstracts used by Manela and Moreira (2017). The shaded regions are 90% confidence bands.

To form estimates of attention using the abstracts corpus we first constrain its vocabulary to that of the full text corpus. We then use the topic model trained on our main sample to calculate topic attention in the abstracts corpus from 1890–1983. Next, we append the fitted topics for the pre-1984 sample with the estimates from our main sample to arrive at a unified time series of topic attention from 1890–2017.

The major benefit of extending our sample is that we capture more than 20 recessions (versus just three in our main sample). Figure 7 plots the time series of "recession" attention over this considerably longer sample. The unified series clearly shows that the *WSJ* allocates more space to recession coverage during NBER recessions. Interestingly, recession attention often rises before an NBER recessions starts and declines before the recession ends. It also shows that recession attention estimates are noisier in the earlier sample of frontage abstracts compared to the later sample that includes full article text for all pages.

Given the longer attention series, we replicate our VAR analysis using data from 1939–2017. 1939 is the first year for which all of the VAR component series are available (non-farm payroll employment has the shortest history, and begins in 1939). Figure 8 reports the impulse-response functions for these VAR estimates. Comparing the impulse responses to those in Figure 4, which are based on the more recent sample, two important patterns emerge. First, we find that the short-term response of both industrial production and employment to a recession news shock is quite similar across samples. In particular, 20 months after an increase in recession news attention, industrial production and employment decline by about two and

one percent, respectively.

Second, we find that the long-run response estimates are more persistent in the longer sample. Estimates based on the more recent sample show considerable long-run reversion. Impulse responses of both output and employment revert to about half of their peak effects. By contrast, the longer frontpage abstracts sample responses show little reversion at 36 months. For robustness, we also report VAR results based on the 1937–1984 sample, and find similar estimates to those from our main 1984–2017 sample.

5.5 Attention Selection in a Text-augmented VAR

Our VAR specification uses recession attention as the only news topic variable in the system. We next show that this is not an arbitrary choice, but is in fact a statistically optimal specification in terms of its tradeoff of model fit and parameter parsimony.

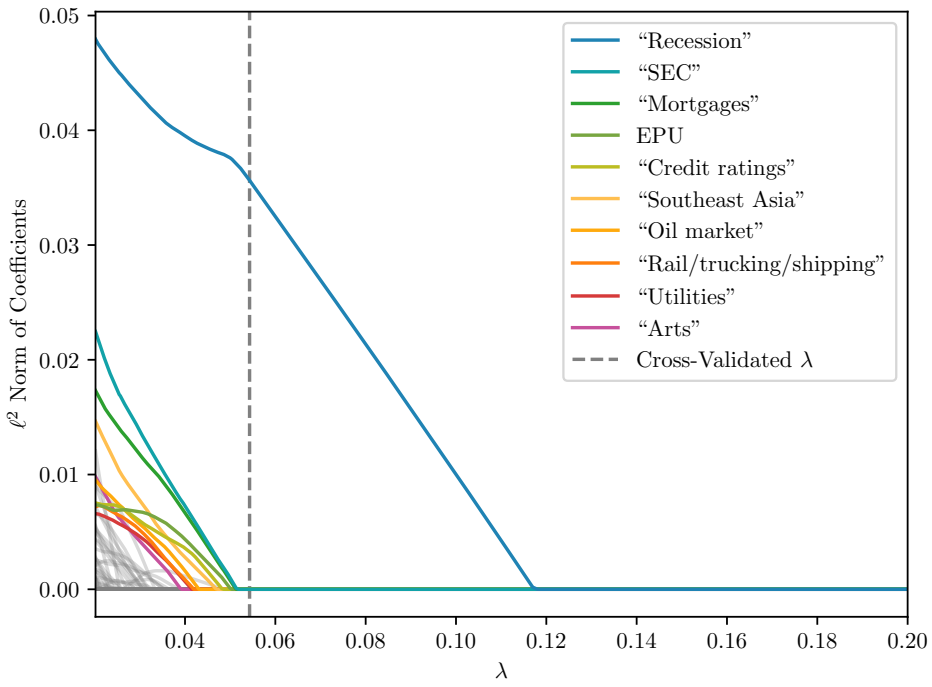
Due to the large size of the topic model, including all news attention series in the VAR is obviously problematic due to "degrees-of-freedom problems" (Bernanke et al., 2005). Instead, we seek a VAR which includes news topics that genuinely influence macroeconomic dynamics and avoids spurious inclusion of irrelevant topics. This is a model selection problem, which we solve with a standard tool: cross-validated lasso regression.

We begin with the four core macroeconomic variables in the VAR specification above (S&P 500 index, Fed funds rate, employment, and industrial production) which we denote as y_t and fix as left side variables in a multivariate regression. We then consider a collection of predictors that includes the 180 news attention series as well as EPU, VIX, and the Michigan consumer sentiment index, which we arrange as $x_t = (x_{1,t}, ..., x_{183,t})'$. The right side variables are three lags each of y_t and x_t:

$$y_t = c + \sum_{l=1}^{3} (\Omega_l y_{t-l} + \Gamma_l x_{t-l}) + \epsilon_t. \tag{2}$$

Our modeling approach uses a lasso penalty to select robust predictors of y_t. We prefer to select or remove a right side variable in its entirety (rather than selecting, say, only the third lag of a given variable, or selecting a variable to explain one element of y_t but not another). The group-lasso of Yuan and Lin (2006), recently used in finance by Freyberger et al. (2020), is ideally suited for this task. Group-lasso assigns non-zero coefficients to only those predictor groups that most reliably forecast y_t, where the notion of "reliable" is determined by the penalty parameter λ. We use 183 groups, one for each variable in x_t (for example, the group for $x_{j,t}$ includes all terms on the right-hand-side of (2) associated with $x_{j,t-1}, x_{j,t-2}$, and $x_{j,t-3}$).

Figure 9: Group-Lasso VAR Selection



Note. ℓ^2 norm for all variables considered in a group-lasso VAR as a function of the penalty parameter λ. The legend shows the ten predictors that survive the highest levels of penalization. Variable names in quotation refer to news attention series.

We include lags of y_t without penalization.[23]

Figure 9 summarizes the importance of each predictor for forecasting macroeconomic outcomes. It shows the ℓ^2 norm of coefficients for each predictor variable as a function of the penalty parameter. When $\lambda \to 0$ there is no selection and all variables are included.[24] As λ rises, parameters become more heavily penalized, and weak predictors begin to drop out. At high levels of λ, only the most potent macroeconomic predictors receive non-zero coefficients.

The legend reports the ten predictor variables that survive heavy penalization, ordered in terms of statistical importance from top to bottom. One predictor stands out above all others—attention to "recession" news. Remarkably, its predictive contribution dominates

[23]Specifically, our group-lasso estimation objective is

$$\min_{\{\Omega_l,\Gamma_l\}_{l=1}^3} \sum_{i=1}^{4}\sum_{t=5}^{T}\left(y_{i,t} - \sum_{j=1}^{4}\sum_{l=1}^{3}\omega_{i,j,l}y_{j,t-l} - \sum_{m=1}^{183}\sum_{l=1}^{3}\gamma_{i,m,l}x_{m,t-l}\right)^2 + \lambda\sum_{m=1}^{183}\sqrt{\sum_{j=1}^{4}\sum_{l=1}^{3}\gamma_{i,j,l}^2}$$

where $\omega_{i,j,l}$ and $\gamma_{i,j,l}$ are the (i,j) elements of Ω_l and Γ_l. All variables are variance-standardized prior to model estimation to ensure comparability of coefficient estimates.

[24]When $\lambda = 0$, the model has more regressors than observations and the estimator is not defined, hence the first estimates on the left of Figure 9 correspond to a small positive value of λ.

EPU, VIX, and the consumer sentiment (of these, only EPU makes the list of top ten predictors). At lower levels of penalization, other news variables that enter the model include "SEC (Securities Exchange Commission)," "mortgages," and "credit ratings."

Figure 9 establishes that the recession topic is the appropriate choice for inclusion in the VAR if we were to choose *just one* topic. But, if we were to use the group-lasso to decide on a model specification, what is the *best* model? Or, equivalently, which value of λ is most appropriate? A standard criterion for selecting among λ values is cross-validation, which is designed to identify the model with the lowest expected out-of-sample forecast error. Specifically, we choose the λ corresponding to the model with lowest mean squared forecast error in ten-fold cross-validation.

The optimal cross-validated model corresponds to $\lambda = 0.054$, depicted as a vertical line in Figure 9. This model selects a single news attention topic, "recession," for inclusion in the VAR. It is based on this result that our text-augmented VAR specification uses "recession" attention alone, as it represents a statistically optimal set of news topics for VAR prediction.[25]

5.6 VAR Discussion

Combining macroeconomic analysis with topic modeling, we draw on a vast corpus of news text to better understand quantitative economic phenomena. Our VAR demonstrates a large incremental role for news attention in understanding economic dynamics. Our interpretation of this result is that news media reflects agents' perceptions of the state of the economy, including (but not limited to) their expectations of future macroeconomic conditions. It should not be surprising that news media can offer a representation of economic conditions that is unspanned by commonly studied numerical macroeconomic data. News contains information about our high dimensional perceptions of the world around us, distilled into narratives through the complex process of human understanding.

Figure 4 provides evidence of economic fluctuations emerging from news. This news is an amalgamation of at least four phenomena. First, business news provides a summary of expectations about future productivity. The news-driven business cycle literature has made theoretical strides in understanding how changes in expectations induce booms and busts even absent concurrent changes in fundamentals (Beaudry and Portier, 2004, 2007, 2014). Second is the closely connected role of noise in business cycle dynamics (as in Lorenzoni,

[25] A technical point regarding our construction of bootstrap standard errors in Figure 4 is that they are not adjusted for lasso model selection. In Appendix G we describe and report impulse response standard errors accounting for post-selection inference. While post-selection inference widens our confidence bands, the effect is small and does not change our conclusions about the statistical significance of news attention impulse responses. Hence, to simplify exposition in the main text, we stick with standard bootstrap confidence intervals without selection adjustment.

2009; Angeletos and La'o, 2010). Inevitable imperfections in macroeconomic expectations (as summarized by news media) can induce volatility in economic activity. A third and also closely related component of news text is the sentiment of economic agents, akin to Keynesian "animal spirits," which drive fluctuations as in Angeletos and La'o (2013). Frequent *WSJ* interviews of influential asset managers, corporate executives, and policymakers are a likely vehicle through which sentiments become reflected in news text. Fourth, news production is determined by media firm incentives, which are in turn influenced by consumer demand (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2010). As such, news text also reflects media slant, which can in turn influence economic dynamics.

While we conduct our macroeconomic analysis in the context of a VAR, this is just one example for how to combine an economic model with news attention to aid inference and interpretation. Other exciting applications include using text narratives to better understand survey revisions and forecast errors (Coibion and Gorodnichenko, 2015; Bordalo et al., 2020) or to interpret residuals implied from DSGE models.

5.7 Narrative Retrieval

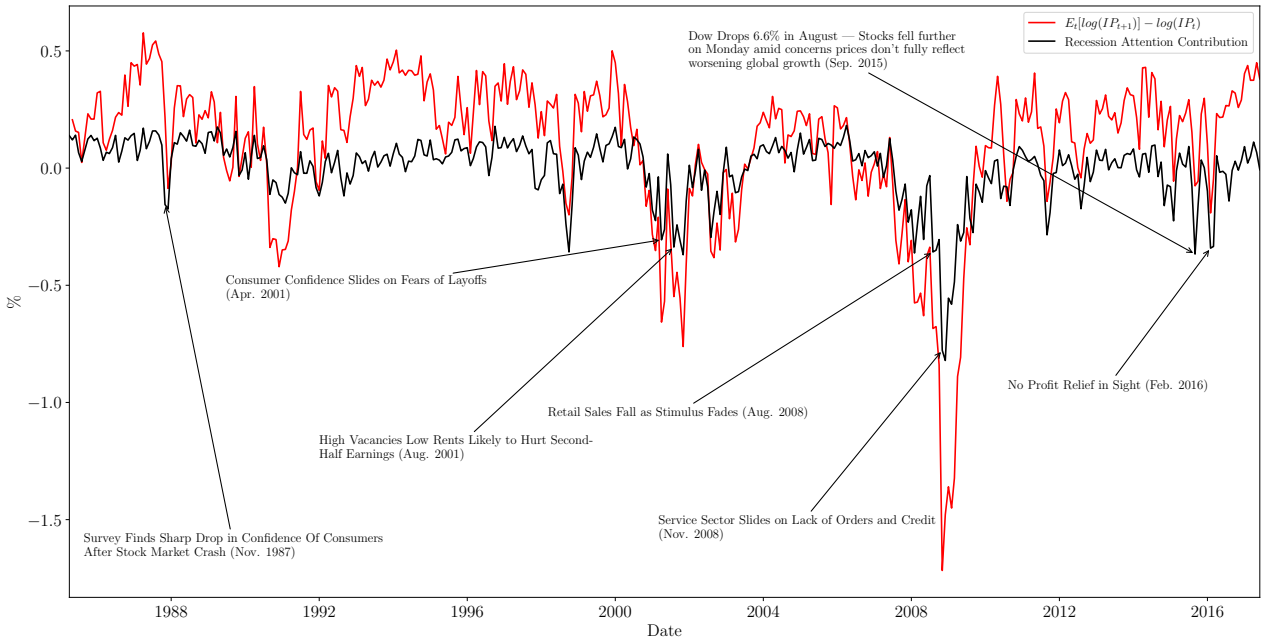
Empirical macroeconomics research has long wrestled with the issue of shock identification (see Ramey, 2016, for a survey discussion). A common approach to model interpretation imposes identifying restrictions on the residual covariance matrix in a VAR, then studies impulse response functions for each variable in the system. Our analysis in Figure 4 is one such example.

A great advantage of embedding news text in macroeconomic models is that fluctuations can be directly mapped to narratives of economic conditions in the underlying article text. This offers a new approach to model interpretation that complements traditional methods. Specifically, we propose a scheme for *narrative retrieval* in text-augmented VAR models. It relies on only two inputs—the estimated VAR coefficients and topic model parameters—without requiring the researcher to impose additional identifying restrictions or economic constraints.

First, we trace future macroeconomic outcomes to prevailing levels of news attention through the estimated VAR coefficients. From here, we further trace economic fluctuations to specific textual narratives using topic model estimates. In each month t, we identify the *WSJ* article published during the month that is most representative of a given VAR topic, k. This representative article is defined as the article i with the largest proportion of its content (θ_i) allocated to topic k.[26] This chain of estimates identifies the most influential individual

[26] Our VAR uses a single topic attention series, but this narrative retrieval approach generalizes to a VAR with any number of topic elements.

Figure 10: Narrative Retrieval for Industrial Production Growth Forecasts



Note. The figure shows one-month-ahead growth rate expectations from the baseline VAR (red line) and contribution to the expectation coming from "recession" attention (black line). We annotate seven of the largest shifts in "recession" attention-based expectations. Annotations are headlines of articles that have the highest allocation to the "recession" topic in a given month.

articles—i.e., specific narratives—that underlie model-implied macroeconomic expectations.

Our narrative retrieval procedure is a natural way to flag articles that coincide with large shifts in model predictions. It points the researcher to the most diagnostic articles for deep reading, bypassing the labor intensive process of reading every document in the underlying corpus. And it automates the integration of narratives into macroeconomic models, rather than relying on manual narrative selection (as in Romer and Romer, 1989).

We demonstrate this approach in the baseline VAR of Figure 4, and study the narrative determinants of output expectations. In Figure 10, the red curve shows one-month-ahead expected output growth based on the estimated VAR. The black curve shows how "recession" news attention contributes to this expectation. This is another visualization of the large impact that news has on output dynamics to complement the impulse responses shown earlier.

Next, we annotate seven of the largest monthly drops in expected output growth coming from shifts in "recession" attention. Annotations show the headline of the article with the highest attention allocation to "recession" at the time of the drop. These articles are the biggest contributors to month-to-month variation in "recession" attention.

Reading the flagged articles carefully, we find output growth expectations are underpinned

by narratives that support theories discussed in Section 5.6. An informative example is the April 2001 article "Consumer Confidence Slides on Fears of Layoffs." The author, Greg Ip, notes that consumer confidence surveys have recently soured and offers an interpretation:

> "mostly because consumers have been more pessimistic about the future as layoff announcements have mounted, energy costs have risen, and stock prices have fallen."

This is a narrative of deteriorating consumer demand born from declining expectations for future growth, consistent with news-driven business cycle theory. The article then goes on to quote differing viewpoints of two experts, suggestive of the mechanism described by Angeletos and La'o (2013) in which macroeconomic expectations emerge from heterogeneous beliefs that propagate across agents.

A number of other articles also highlight narratives about economic growth expectations (November 2008, September 2015) and corporate profit expectations (August 2001, August 2008, February 2016) as primary drivers of VAR forecasts. The September 2015 article "Dow Drops 6.6% in August — Stocks Fell Further on Monday Amid Concerns Prices Don't Fully Reflect Worsening Global Growth" is another informative example. Why does news text provide such large incremental predictive information above S&P 500 valuations? This article suggests a rationale in the form of heterogeneous belief propagation. Despite the fact that "the Dow lost 6.6% in August, its largest one-month percentage decline since May 2010," the Dow fell another 0.7% on Monday, September 1, 2015. The narrative put forth by journalists Saumya Vaishampayan and Christopher Whittall is:

> "Monday's selling was driven by intensifying concerns that stock prices don't fully reflect the deterioration in recent months of the global economic outlook, traders said. ... While many traders and analysts remain upbeat on the prospects for the U.S. economy and generally optimistic about investing in the shares of large U.S. companies, market sentiment is broadly cautious following the sharp price swings of the past two weeks.... Some portfolio managers are heeding signs that growth is slowing in the economies of many developing nations and that economic gains are likely to be slow in coming elsewhere. 'Emerging markets don't look that great, the Federal Reserve is going to raise rates, and Europe is still gradually emerging from recession,' said David Lebovitz, global market strategist for J.P. Morgan Asset Management."

The November 1987 article "Survey Finds Sharp Drop in Confidence Of Consumers After Stock Market Crash" is yet another informative data point. This article's narrative illustrates how the sudden and extreme stock price drop in October 1987 dented consumer confidence:

> "Consumer Sentiment stood at 92.5 before the panic selloff on Wall Street Oct. 19 but plummeted to 82.4 for the rest of the month. ... 'Needless to say the current situation is volatile,' the surveyors said. 'Whether a recession eventually develops can't be determined from these data (because) the response by consumers has yet to move past their initial reactions. Nonetheless the data indicate weakening sales and a much larger potential downside risk.' ... The surveyors conceded that 'caution must be used in the interpretation of these results because the post-crash sample was so small.' ... Mr. Curtin[27] said consumers still haven't assessed the personal toll such as reduced income or lost jobs that the crash may ultimately exact."

In the context of the text-augmented VAR, this narrative is associated with a drop in expected one-month industrial production grown of roughly 0.25%. The causes of the 1987 crash are hotly debated to this day,[28] but it is generally agreed that the crash was a transitory incident with little effect on subsequent macroeconomic fundamentals (Barro, 1990). All the more impressive, then, is how the uncertainty emphasized in this article harms economic expectations in the immediate aftermath of the crash. The decline in VAR-based output expectations recovers within two months of the crash as recession narratives associated with the crash quickly disappear. This episode resembles the effects of a prototypical noise shock in noisy business theories (Lorenzoni, 2009; Angeletos and La'o, 2010).

6 Stock Market Timing

In this section we investigate the association between news attention and aggregate market return dynamics. Predictable variation in returns may arise from at least two sources; market inefficiency and time-varying discount rates. In inefficient capital markets, some information contained in the newspaper may not be fully priced by the stock market (Fama, 1970). Predictability can also arise if news attention to specific topics is indicative of systematic behavioral biases of investors (Shiller, 2003). Alternatively, even in efficient markets, rational time variation in discount rates can lead to return predictability (Cochrane, 2011). For example, if heightened attention to recession news indicates an elevated disaster probability, such periods should be followed by higher expected returns (e.g. Wachter, 2013).

When considering return predictability based on news attention, we are cognizant that full sample topic estimates are susceptible to look-ahead bias. To address this, in addition to our

[27] Richard T. Curtin, director of consumer surveys at the University of Michigan's Institute of Social Research

[28] Potential explanations range from the prevalence of portfolio insurance strategies (Brady et al., 1988) to proposed and eventually abandoned legislation on corporate takeovers (Mitchell and Netter, 1989).

full sample LDA attention estimates, we also consider a time series of oLDA topic attention estimates (introduced in Section 5.3), which is free of look-ahead bias. When using oLDA we use the same number of topics as in the full sample LDA specification.

For both topic models we form a lasso-based forecast of market returns selecting five regressors as in the main specifications. We detrend monthly attention estimates because, as Figure 3 shows, many topics such as recession and terrorism are increasingly covered by the newspaper. Following Kelly et al. (forthcoming), returns are standardized by their trailing 12-month standard deviation, while we standardize the signals by the trailing full sample standard deviation.

We train the lasso model out-of-sample starting with half the sample as training data and then evaluate one month out-of-sample before extending the training sample with the new month of data and repeating the procedure. This expanding sample procedure results in 174 months of test sample data. We start our sample in 1988 – excluding the first four years from the sample as oLDA requires a burn-in period to converge. This period corresponds to the point where the log-likelihood estimates for oLDA stabilize.

To evaluate the economic significance of the forecasting model, we convert our return forecasts into an out-of-sample market timing strategy. Each month t, the strategy takes a position in the market that is proportional to the one-month-ahead out-of-sample return forecast as of month t.

We compare our news-based timing strategies to several natural benchmarks. One is a naive buy-and-hold strategy that holds the market in constant proportion throughout our sample. We also consider trading strategies for two alternative text-based measures: (i) inspired by *The Economist*'s R-word index, we consider the frequency of mentions of the word "Recession" in each month, (ii) the economic policy index used previously in Section 5.1. We detrend both of these text-based measures in the same way we detrend the LDA attention series. Lastly, we consider a trading strategy using the 15 monthly predictors from the "kitchen sink" forecasting regression in Welch and Goyal (2008). For all but the buy-and-hold strategy, we regress next month's stock market returns on the predictors during a training period. We then form a predicted return $\hat{r}_{t+1}$ based on month t predictors. Finally, we hold $\hat{r}_{t+1}$ units of the index to the next period, yielding a trading strategy return of $\hat{r}_{t+1} \times r_{t+1}$. In the case of the Welch and Goyal (2008) kitchen sink regression, we also consider a lasso-based timing strategy in analogy to the lasso regression step underlying the news attention trading strategy.

Table 6: Out-of-Sample Performance of Market Timing Strategies

	Buy-and-hold	LDA	oLDA	R-word	EPU	GW	GW Lasso
Sharpe Ratio	0.71	0.99	1.04	0.65	0.53	0.53	0.62
Expected Return	2.72	3.79	4.01	2.50	2.04	2.03	2.40
t-Stat	2.69	3.76	3.97	2.48	2.02	2.01	2.37
Information Ratio		0.29	0.31	-0.46	-0.50	-0.18	-0.07
α		2.30	2.87	-0.20	-0.51	0.61	1.55
α t-Stat		2.67	3.06	-1.54	-1.43	0.69	1.58
Max Loss	-47.09	-47.31	-25.89	-46.98	-77.66	-65.76	-64.31
Skew	-0.56	0.20	1.04	-0.66	-1.41	-0.71	0.05

Note. Reports annualized performance measures for the different market timing strategies. Buy-and-hold corresponds to a strategy that holds the market. LDA corresponds to a lasso strategy using our original full sample LDA attention series. oLDA corresponds to a lasso strategy which uses oLDA in-place of full-sample LDA. R-word uses a prediction based on mentions of the word "Recession" in the *WSJ*. EPU uses a prediction based on the economic policy uncertainty index. GW uses the 15 predictors from the Welch and Goyal (2008) "kitchen sink" regression. GW lasso uses a comparable lasso procedure to oLDA with the same 15 predictors from GW. All strategies except buy-and-hold, hold the forecasted market return to the following month.

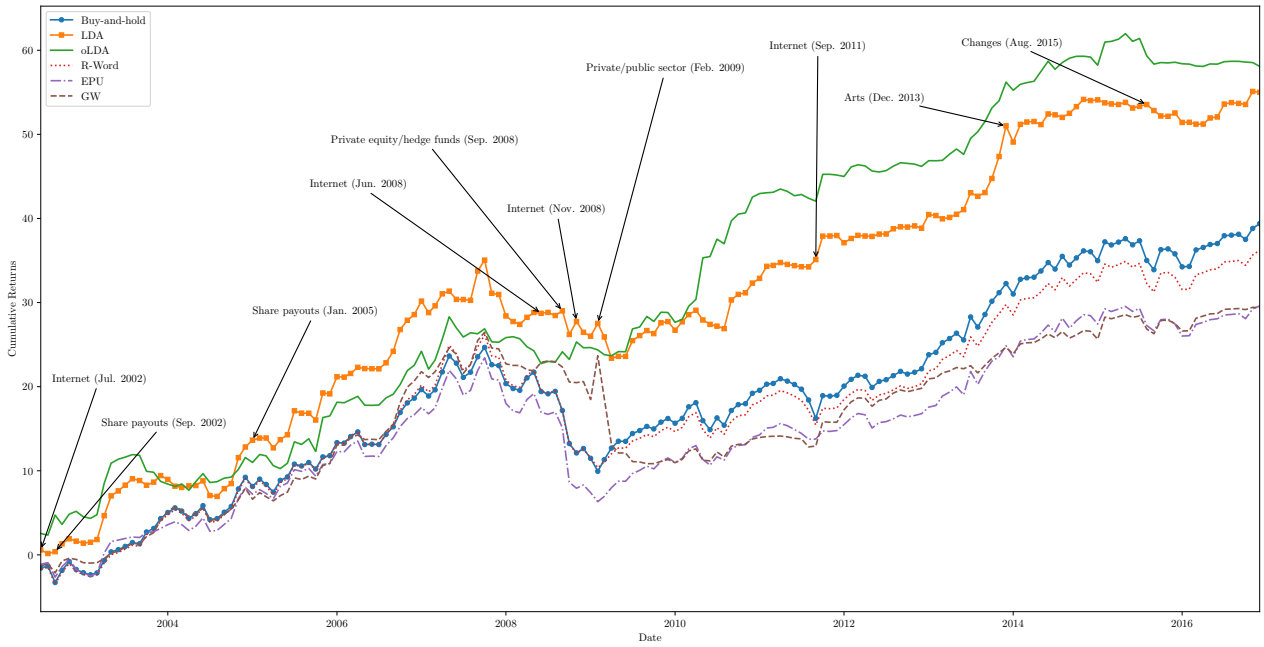
6.1 Results

Table 6 reports the main results on market timing with news attention. We report annualized performance statistics for the different strategies, which are commonly used to compare strategies by the market timing literature (e.g., Kelly et al., forthcoming). For each strategy, we report the Sharpe ratio, expected return, and the corresponding t-statistic for the expected return. To study the excess performance of each strategy relative to the buy-and-hold benchmark, we next report the information ratio (IR), alpha and alpha t-statistic from a time series regression of each strategy on the buy-and-hold returns. Finally, we also report the maximum loss, or the lowest return earned by the strategy, and its skewness.

We find that news attention provides a useful signal for market timing. Both the LDA and oLDA strategies generate the highest Sharpe ratio, alpha, and t(alpha). At the same time its maximum loss is modestly elevated but comparable to the other benchmarks. The worst performer is the Goyal-Welch kitchen sink regression model. Because the advantage of LDA over the Goyal-Welch strategy could be due to its use of regularization, we also analyze a lasso variant similar to the one we use with LDA. As the table shows, performance of the regularized Goyal-Welch strategy remains below that of the buy-and-hold benchmark, and well below that of LDA.

The ability of our LDA strategy to outperform the market suggests that paying attention to *Wall Street Journal* coverage can be quite lucrative for investors. It also shows that,

Figure 11: Cumulative Returns for Market Timing Strategies



Note. Reports the cumulative return for different market timing strategies over the testing sample. Buy-and-hold corresponds to a strategy that holds the market with static leverage. LDA corresponds to a lasso strategy using our original full sample LDA attention series. oLDA corresponds to a lasso strategy which uses online LDA in-place of full-sample LDA. R-word uses a prediction based on mentions of the word "Recession" in the *WSJ*. EPU uses a prediction based on the economic policy uncertainty index. GW uses the 15 predictors from the Welch and Goyal (2008) "kitchen sink" regression. Returns are standardized with the same standard deviation as the buy-and-hold strategy for interpretability. Labeled are the months with the largest LDA outperformance relative to the buy-and-hold benchmark and their most relevant topic.

despite reducing the dimensionality of the newspaper text in an unsupervised manner, the reduction effectively summarizes the main themes of the newspaper. This reduction enables the forecasting regression model to estimate expected market returns based on high-level topics, rather than specific word usage such as "Recession."

6.2 Narrative Retrieval

Figure 11 reports the cumulative returns for the different strategies. It reveals the points in time when LDA diverges by outperforming the benchmark strategies. For example, while holding the market or pursuing most other strategies would have generated substantial losses in 2008 and in 2011, the LDA strategy reduces its market exposure during these periods.

What does the *WSJ* report during these periods that leads the LDA strategy to expect below average market returns? We next run a variant of the above narrative retrieval analysis but for the market timing application. We aim to understand the news shocks that correspond

most to large differences between our market timing strategy's performance and that of the simple buy-and-hold strategy.

To retrieve the narratives underlying these performance gains, we label the months in Figure 11 with the largest performance improvement in our timing strategy over the buy-and-hold strategy. We denote the change in topic k's attention at time t by $\Delta\theta_{t,k}$, the regression coefficient estimated over the preceding training period by $\hat{\beta}_{t,k}$, and the realized market return by r_t. We select the topic k, which contributes the most to the market timing strategy's performance in month $t+1$, follows:

$$\hat{k}_{t+1} = \text{argmax}_k \Delta\theta_{t,k}\hat{\beta}_{t,k}r_{t+1}. \tag{3}$$

We can then isolate the articles most relevant to this improvement by sorting articles based on their exposure to this topic.

For example, our predictive model identifies increased attention to the "private/public sector" topic as a strong negative signal about future market returns because prior such news preceded stock market declines. This leads the LDA strategy to short the market going into February 2009, and to generate a 3 standard deviation return gap relative to the declining market, which is the largest gap in our sample. By focusing on the articles with the largest attention to the private/public sector topic, we can discern a clear narrative for this outperformance.

In January 2009, the newly-elected Obama administration frantically tries to restore confidence in the US banking sector. It also initiates its key healthcare reform, which would later become the Affordable Care Act. For example, the *WSJ* reports under the headline "Stimulus Package Unveiled — $825 Billion Plan Includes Business Tax Breaks; Senate Releases Cash for Bank Rescues,"

> "The stimulus plan was released hours before the Senate backed Mr. Obama on another key measure by approving the Treasury's call for the release of the second half of the 700 billion financial system rescue package. That program is being replenished as Bank of America Corp. was near an agreement with U.S. officials that would provide it with 15 billion to 20 billion of fresh capital. The plan would be one of the largest single government expenditures in U.S. history and would be equivalent to about 3 of gross domestic product over two years."

Later that month, under the headline "Banks Hit by Nationalization Fears — Financials Plunge As U.S. Considers New Rescue Options," the *Journal* reports,

> "Shares of the biggest names in American banking plunged Tuesday as some investors feared that the government would need to nationalize the most deeply

> wounded financial institutions wiping out stockholders. The hours-old administration of President Barack Obama is expected to move swiftly to try to stabilize the financial system by pumping more capital into weakened banks and buying bad assets. Nationalization appears to be a last resort but other options on the table move the U.S. in that direction."

In February 2009, when this outperformance realizes, confidence in banks worsens and many believe the US government is destined to take over some of the largest banks. The *Journal* reports under the headline "U.S. Seeks to Stem Bank Fears — White House Plays Down Nationalization Talk as Stocks Hit New Bear-Market Low,"

> "The White House tried to knock down speculation that the government is preparing to nationalize several large U.S. banks, but some bankers complained that the Obama administration needs to act even more aggressively to shore up confidence in battered financial institutions."

The budget proposal by the new administration generally promises a larger role for the federal government in the US economy, which likely worsens such fears and further depresses market prices. A front page article titled "Obama Budget Pushes Sweeping Change" describes this policy change as follows:

> "The budget blueprint for fiscal year 2010 is one of the most ambitious policy prescriptions in decades, a reordering of the federal government to provide national health care, shift the energy economy away from oil and gas, and boost the federal commitment to education."

These articles suggest that the notable outperformance in February 2009 stems from an underestimation by investors of the degree to which the financial crisis would prompt the new administration to redraw the boundaries between the private and public sectors of the US economy, and what this policy change implies for shareholders.

7 Conclusion

Understanding the forces that drive fluctuations in the state of the economy is central to economic modeling. The overwhelming majority of empirical research has focused on numerical macroeconomic indicators to approach this problem. We offer an alternative approach that summarizes economic conditions in terms of narratives in business news.

Our approach is motivated by the view that news text is a mirror of the state of the economy. The media sector is an information intermediary that meets information demand of

consumers and investors with verbal descriptions of economic events and their interpretation. Of course, news media is a verbal mirror of the economy that comes with flawed refractions, often in the form of producer and consumer biases, noisy inferences, and speculative sentiment.

We estimate a topic model from the full text of *The Wall Street Journal*. Our estimates provide a taxonomy of themes that are the subjects of attention in financial markets and the broader economy. We measure how much attention is allocated to each theme at each point in time, and then use these measurements as inputs into statistical models of economic fluctuations.

Economic topics identified from *WSJ* text closely coincide with related numerical measures of economic activity, including macroeconomic aggregates like output and employment, financing activity, asset prices, and different measures of economic uncertainty. In the context of a standard macroeconomic VAR, we show that news topic attention strongly influences economic dynamics above and beyond standard numerical economic indicators. In the context of stock market timing, we show that an Online LDA-based strategy outperforms the market as well as several natural alternative strategies.

We propose a new perspective on model interpretation with narrative retrieval. Our approach relies on the model to digest massive text corpora that are beyond human readability and flags articles that are most statistically related to a specific fluctuation of interest. It thus isolates news-based narratives of those events that the researcher can "close-read" for detail and nuance about the drivers of model-based expectations.

Finally, readers may explore our results in greater detail through the interactive website, www.structureofnews.com, where we post all of our topic model output. Through the website, researchers can download our *WSJ* news attention time series for use in their own projects. We intend to post regular model updates as new data becomes available.

References

Angeletos, George-Marios, and Jennifer La'o, 2010, Noisy business cycles, *NBER Macroeconomics Annual* 24, 319–378.

Angeletos, George-Marios, and Jennifer La'o, 2013, Sentiments, *Econometrica* 81, 739–779.

Asuncion, Arthur, Max Welling, Padhraic Smyth, and Yee Whye Teh, 2009, On smoothing and inference for topic models, in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, 27–34, AUAI Press.

Baker, Scott R, Nicholas Bloom, and Steven J Davis, 2016, Measuring economic policy uncertainty, *The quarterly journal of economics* 131, 1593–1636.

Baker, Scott R, Nicholas Bloom, Steven J Davis, and Marco C Sammon, 2021, What triggers stock market jumps?, Technical report, National Bureau of Economic Research.

Barro, Robert J, 1990, The stock market and investment, *The review of financial studies* 3, 115–131.

Beaudry, Paul, and Franck Portier, 2004, An exploration into pigou's theory of cycles, *Journal of monetary Economics* 51, 1183–1216.

Beaudry, Paul, and Franck Portier, 2007, When can changes in expectations cause business cycle fluctuations in neo-classical settings?, *Journal of Economic Theory* 135, 458–477.

Beaudry, Paul, and Franck Portier, 2014, News-driven business cycles: Insights and challenges, *Journal of Economic Literature* 52, 993–1074.

Bernanke, Ben S, Jean Boivin, and Piotr Eliasz, 2005, Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach, *The Quarterly journal of economics* 120, 387–422.

Blei, David M, Andrew Y Ng, and Michael I Jordan, 2003, Latent dirichlet allocation, *Journal of machine Learning research* 3, 993–1022.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer, 2020, Overreaction in macroeconomic expectations, *American Economic Review* 110, 2748–82.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson, 2018, Information, trading, and volatility: Evidence from firm-specific news, *The Review of Financial Studies* 32, 992–1033.

Boydstun, Amber E., 2013, *Making the News: Politics, the Media, and Agenda Setting* (University of Chicago Press).

Brady, Nicholas, Presidential Task Force on Market Mechanisms, et al., 1988, *Report of the Presidential Task Force on Market Mechanisms: Submitted to the President of the United States, the Secretary of the Treasury and the Chairman of the Federal Reserve Board* (US Government Printing Office).

Campbell, John Y, and Samuel B Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *The Review of Financial Studies* 21, 1509–1531.

Chahrour, Ryan, Kristoffer Nimark, and Stefan Pitschner, 2021, Sectoral Media Focus and Aggregate Fluctuations, *American Economic Review* 111, 3872–3922.

Chauvet, Marcelle, and Jeremy Piger, 2008, A comparison of the real-time performance of business cycle dating methods, *Journal of Business & Economic Statistics* 26, 42–49.

Cochrane, John H., 2011, Presidential Address: Discount Rates, *Journal of Finance* 66, 1047–1108.

Coibion, Olivier, and Yuriy Gorodnichenko, 2015, Information rigidity and the expectations formation process: A simple framework and new facts, *American Economic Review* 105, 2644–78.

Cong, Lin William, Tengyuan Liang, and Xiao Zhang, 2020, Textual factors: A scalable, interpretable, and data-driven approach to analyzing unstructured information, *SSRN Working Paper* .

Ellingsen, Jon, Vegard Larsen, and Leif Anders Thorsrud, 2020, News Media vs. FRED-MD for Macroeconomic Forecasting, Working Paper ID 3720396, Social Science Research Network, Rochester, NY.

Engle, Robert, Takatoshi Ito, and Wen-Ling Lin, 1990, Meteor showers or heat waves? heteroskedastic intra-daily volatility in the foreign exchange market, *Econometrica* 58, 525–42.

Engle, Robert F, 2019, Financial volatility in an age of geopolitical risks, Technical report, NYU Working Paper.

Engle, Robert F, Stefano Giglio, Bryan T Kelly, Heebum Lee, and Johannes Stroebel, 2019, Hedging climate change news, *Review of Financial Studies* .

Fama, Eugene F., 1970, Efficient Capital Markets: A Review of Theory and Empirical Work on JSTOR, *Journal of Finance* 25, 383–417.

Fama, Eugene F, 1990, Stock returns, expected returns, and real activity, *The journal of finance* 45, 1089–1108.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, 2020, Dissecting Characteristics Nonparametrically, *Review of Financial Studies* 33, 2326–2377.

García, Diego, 2013, Sentiment during recessions, *Journal of Finance* 68, 1267–1300.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, 2019, Text as Data, *Journal of Economic Literature* 57, 535–574.

Gentzkow, Matthew, and Jesse M Shapiro, 2010, What drives media slant? evidence from us daily newspapers, *Econometrica* 78, 35–71.

Hansen, Stephen, Michael McMahon, and Andrea Prat, 2017, Transparency and deliberation within the fomc: a computational linguistics approach, *The Quarterly Journal of Economics* 133, 801–870.

Herskovic, Bernard, Bryan Kelly, Hanno Lustig, and Stijn Van Nieuwerburgh, 2016, The common factor in idiosyncratic volatility: Quantitative asset pricing implications, *Journal of Financial Economics* 119, 249–283.

Hoffman, Matthew, Francis Bach, and David Blei, 2010, Online Learning for Latent Dirichlet Allocation, *Advances in Neural Information Processing Systems* 23.

Ke, Zheng Tracy, Bryan T Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, Technical report, National Bureau of Economic Research.

Kelly, Bryan, Asaf Manela, and Alan Moreira, 2021, Text selection, *Journal of Business & Economic Statistics* 1–21.

Kelly, Bryan T., Semyon Malamud, and Kangying Zhou, forthcoming, The Virtue of Complexity in Return Prediction, *Journal of Finance* .

Kilian, Lutz, 1999, Finite-sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses, *Review of Economics and Statistics* 81, 652–660.

Larsen, Vegard H., and Leif A. Thorsrud, 2019, The value of news for economic developments, *Journal of Econometrics* 210, 203–218.

Lorenzoni, Guido, 2009, A theory of demand shocks, *American Economic Review* 99, 2050–84.

Maaten, Laurens van der, and Geoffrey Hinton, 2008, Visualizing data using t-sne, *Journal of machine learning research* 9, 2579–2605.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

McCracken, Michael W, and Serena Ng, 2016, Fred-md: A monthly database for macroeconomic research, *Journal of Business & Economic Statistics* 34, 574–589.

Mitchell, Mark L, and Jeffry M Netter, 1989, Triggering the 1987 stock market crash: Antitakeover provisions in the proposed house ways and means tax bill?, *Journal of Financial Economics* 24, 37–68.

Mullainathan, Sendhil, and Andrei Shleifer, 2005, The market for news, *American Economic Review* 95, 1031–1053.

Murtagh, Fionn, and Pierre Legendre, 2014, Ward?s hierarchical agglomerative clustering method: which algorithms implement ward?s criterion?, *Journal of classification* 31, 274–295.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev, 2010, How to analyze political attention with minimal assumptions and costs, *American Journal of Political Science* 54, 209–228.

Ramey, Valerie A, 2016, Macroeconomic shocks and their propagation, *Handbook of macroeconomics* 2, 71–162.

Romer, Christina D, and David H Romer, 1989, Does monetary policy matter? a new test in the spirit of friedman and schwartz, *NBER macroeconomics annual* 4, 121–170.

Shiller, Robert J., 2003, From Efficient Markets Theory to Behavioral Finance, *Journal of Economic Perspectives* 17, 83–104.

Steyvers, Mark, and Tom Griffiths, 2007, Probabilistic topic models, *Handbook of latent semantic analysis* 427, 424–440.

Tang, Jian, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang, 2014, Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis, in *International Conference on Machine Learning*, 190–198 (PMLR).

Thorsrud, Leif Anders, 2020, Words are the New Numbers: A Newsy Coincident Index of the Business Cycle, *Journal of Business & Economic Statistics* 38, 393–409.

Tibshirani, Ryan J, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani, 2016, Exact post-selection inference for sequential regression procedures, *Journal of the American Statistical Association* 111, 600–620.

Wachter, Jessica A., 2013, Can Time-Varying Risk of Rare Disasters Explain Aggregate Stock Market Volatility?, *Journal of Finance* 68, 987–1035.

Welch, Ivo, and Amit Goyal, 2008, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Rev. Financ. Stud.* 21, 1455–1508.

Xiao, Han, and Thomas Stibor, 2010, Efficient collapsed gibbs sampling for latent dirichlet allocation, in *Proceedings of 2nd asian conference on machine learning*, 63–78, JMLR Workshop and Conference Proceedings.

Yuan, Ming, and Yi Lin, 2006, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 49–67.

Internet Appendix

A Constructing the *WSJ* Document-Term Matrix

We conduct data processing steps in the following order:

1. Remove all articles prior to January 1984 and after June 2017 (data purchased at the beginning of July 2017).
2. Replace all non-alphabetical characters with an empty string and set the remaining characters to lower-case.
3. Parse article text into a white-space-separated word list retaining the article's word ordering. Exclude single-letter words.
4. Exclude articles with page-citation tags corresponding to any sections other than A, B, C, or missing.
5. Exclude articles corresponding to weekends.
6. Exclude articles with subject tags associated with obviously non-economic content such as sports. List of exclusions available from here:
   https://www.dropbox.com/sh/jvtvdpo56jvi5o3/AAAcAWVLtqrCnc-tAgdV0zaha?dl=0.
7. Exclude articles with the certain headline patterns (such as those associated with data tables or those corresponding to regular sports, leisure, or books columns). List of exclusions available from here:
   https://www.dropbox.com/sh/jvtvdpo56jvi5o3/AAAcAWVLtqrCnc-tAgdV0zaha?dl=0.
8. Concatenate articles with the same accession-number as these are chained articles.
9. Exclude articles with less than 100 words.
10. Remove common "stop" words and URL-based terms. List of exclusions is standard but available from here:
    https://www.dropbox.com/sh/jvtvdpo56jvi5o3/AAAcAWVLtqrCnc-tAgdV0zaha?dl=0.
11. Lastly, we conduct light lemmatizing of derivative words. The following rules are applied in the order given, where 'x' is a candidate term. In each case, the stemming is only applied if the multiple terms reduce to the same stem.
    (a) Replace trailing "sses" with "ss"
    (b) Replace trailing "ies" with "y"
    (c) Remove trailing "s"

(d) Remove trailing "ly"

(e) Remove trailing "ed." Replace remaining trailing "ed" with "e"

(f) Replace trailing "ing" with "e". For remaining trailing "ing" that follow a pair of identical consonants, remove "ing" and one consonant. Remove remaining trailing "ing"

(g) Remove words with less than 3 letters.

12. From the resulting uni-grams, generate the set of bi-grams as all pairs of (ordered) adjacent uni-grams.

13. Exclude terms (both uni-grams and bi-grams) appearing in less than 0.1% of articles. The unique set of terms is the corpus vocabulary. Each column of the DTM corresponds to an element of the vocabulary.

14. Convert an article's word list into a vector of counts for each term in the vocabulary. This vector is the row of the DTM corresponding to the article.

B Topic Model Estimation

As mentioned above, in principle, one could estimate ϕ and θ via maximum likelihood according to (1). However, in most text applications this is computationally unrealistic. Instead, we use Bayesian methods and approach estimation with the collapsed Gibbs sampler proposed by Xiao and Stibor (2010).[29]

The intuition behind the estimation procedure can be understood with minimal technical overhead. First, note that model (1) can be equivalently represented as a "generative" model; that is, as a set of sampling rules. Taking model parameters ϕ_k and θ_t as given,[30] we can conceptualize writing an article in the context of the model. Consider the t^{th} article with N_t total words. How does the model "write" the first term in the article? First, it draws a topic at random from the list of all topics, where the probabilities of drawing each topic are given by the $K \times 1$ vector θ_t. The topic assignment of that first term, call it $z_{t,1}$, is distributed as a unit multinomial (i.e., categorical) with parameter θ_t. Suppose we happened to draw topic k for the first term's topic. To then draw the first term itself, we randomly choose from the vocabulary of all terms, where the probability of drawing any given term is governed by the $V \times 1$ vector ϕ_k, which is the topic-specific term probability distribution. Call this first term $x_{t,1}$, which is distributed as a unit multinomial with parameter ϕ_k. This process is repeated for each subsequent term in the article, a total of N_t times, at which point

[29] We refer interested readers to Steyvers and Griffiths (2007) for an especially transparent description of LDA Gibbs samplers. Asuncion et al. (2009) review and compare alternative estimation approaches such as variational Bayes and maximum a posteriori estimation. They conclude that, with proper hyperparameter tuning, all estimators perform similarly in terms of model fit.

[30] These are drawn in a preliminary step from the Dirichlet priors, $\theta_t \sim \text{Dir}(\alpha)$ and $\phi_k \sim \text{Dir}(\beta)$.

the model has authored the complete article and stops. In distribution notation, each term i in the article is given by

$$x_{t,i} \sim \text{Mult}(\phi_{z_{t,i}}, 1), \quad z_{t,i} \sim \text{Mult}(\theta_t, 1).$$

The Gibbs sampling estimation procedure searches for parameter values ϕ_k and θ_t that simulate articles most similar to those found in our news data set.[31]

The estimated attention that article t allocates to topic k can be represented as the frequency with which its terms are assigned to k

$$\hat{\theta}_{t,k} = \frac{\sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = k)}{\sum_{q=1}^{K} \sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = q)}, \tag{4}$$

where $\hat{z}_{t,i}$ are topic assignments from the estimated model. Likewise, estimates of the topics themselves are determined by the number of times each term v in the vocabulary is assigned to topic k (aggregated over all articles):[32]

$$\hat{\phi}_{k,v} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(x_{t,i} = v)\mathbb{I}(\hat{z}_{t,i} = k)}{\sum_{m=1}^{V} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(x_{t,i} = m)\mathbb{I}(\hat{z}_{t,i} = k)}. \tag{5}$$

Finally, while $\hat{\theta}_t$ describes the allocation of news attention to each topic in article t, we are often interested in how the *WSJ* as a whole allocates attention over time. We measure the overall news attention at a point in time τ (e.g., a given day or month) by summing over all articles at that time. Thus, *WSJ* attention allocation to topic k in month τ is

$$\hat{\theta}_{\tau,k} = \frac{\sum_{t \in \tau} \sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = k)}{\sum_{t \in \tau} \sum_{q=1}^{K} \sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = q)}. \tag{6}$$

Our model reportage focuses on estimates of the topics, $\hat{\phi}_k$, and estimates of attention allocation across topics, $\hat{\theta}_t$.

C Topic Key Terms

[31] We take 300 iterations of the Gibbs sampler over the entire corpus, and use hyperparameters of $\beta = 1$ and $\alpha = 1$.

[32] As shown in (4) and (5), we choose to not incorporate priors in our θ and ϕ estimates. Griffiths and Steyvers (2004) show that priors can be easily incorporated as

$$\hat{\theta}_{t,k} = \frac{\sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = k) + \alpha}{\sum_{q=1}^{K} \sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = q) + K\alpha}, \quad \hat{\phi}_{k,v} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(x_{t,i} = v)\mathbb{I}(\hat{z}_{t,i} = k) + \beta}{\sum_{m=1}^{V} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(x_{t,i} = m)\mathbb{I}(\hat{z}_{t,i} = k) + V\beta}.$$

Table 7: Topic Key Terms

Metatopic label	Topic label	Key terms
Labor/Income	Executive pay	*bonus, base salary, total compensation, pay package, compensation package, compensation committee, restrict stock, tyco, executive compensation, bonuses, stockoption, executive pay, stock option, compensation, annual salary, tyco international, exercise option, severance package, severance payment, exercise price*
	Job cuts	*job cut, work force, fulltime employee, layoff, eliminate job, attrition, cut job, employment opportunity, employee work, unemployed, workforce, find job, parttime, lost job, temporary worker, retraining, job loss, skill worker, fulltime, humanresources*
	Unions	*uaw, teamsters, relation board, union member, national labor, labor relation, picket line, worker union, union leader, nonunion, union official, international union, walkout, collective bargain, strikers, teamsters union, picket, union worker, labor agreement, contract talk*
	Health insurance	*hmo, health plan, health coverage, health care, manage care, blue cross, health insurance, care act, blue shield, healthcare system, health maintenance, humana, affordable care, nurse home, medical cost, healthmaintenance, medicaid program, health insurer, maintenance organization, health system*
	Pensions	*pension plan, employee retirement, benefit guaranty, pension, state pension, retirement system, retirement save, pension liability, retirement benefit, pension system, pension benefit, retirement, underfunded, retirement plan, retirement fund, retiree, public pension, retirement program, pension fund, retirement age*
	Government budgets	*spend cut, deficit reduction, grammrudman, federal budget, balance budget, budget plan, cbo, house budget, entitlement program, budget, fiscal budget, federal spend, budget process, budget proposal, cut spend, budget office, domestic spend, congressional budget, budget cut, budget gap*
	Fees	*fee paid, pay fee, fee, charge fee, fee charge, higher fee, receive payment, license fee, company pay, upfront, annual fee, reimburse, total cost, paid, extra cost, cash payment, company paid, additional cost, administrative cost, cover cost*
	Taxes	*irs, tax rate, deduction, income tax, tax bill, irss, capitalgains, capitalgains tax, tax court, taxable income, corporate tax, tax law, tax system, marginal tax, tax expert, tax deduction, minimum tax, incometax rate, top tax, tax burden*
Management	Connecticut	*stamford conn, stamford, greenwich conn, conn, haven conn, clark, greenwich, norwalk, norwalk conn, westport, wallace, moore, kelley, westport conn, allen, murphy, evan, robinson, cunningham, johnson*
	C-suite	*executive officer, chief operate, operate officer, financial officer, chief financial, chairman president, officer john, officer robert, company chief, administrative officer, named chief, chief administrative, chairman chief, officer told, officer david, cfo, officer, remain chairman, act chief, finance chief*
	Mid-level executives	*vice, vice president, executive vice, senior vice, named senior, named vice, named executive, president finance, president market, group vice, company vice, vice chairman, named president, group executive, newly create, business development, assistant vice, general manager, named, group president*
	Management changes	*succession plan, top job, departure, management shakeup, management style, shakeup, succession, heir apparent, top post, daytoday operation, nonexecutive chairman, successor, top management, management structure, daytoday, resignation, nonexecutive, management committee, management team, executive suite*
Trans/Defense/Local	Natural disasters	*katrina, quake, hurricane katrina, fishermen, coast guard, storm, emergency management, tsunami, earthquake, boat, hurricane, natural disaster, fisheries, baton rouge, salmon, beaches, mississippi river, debris, rita, dam*
	Police/crime	*inmates, police officer, violent crime, police department, gang, police, homicide, firearms, robbery, gunman, police chief, murder, traffickers, offenders, police arrest, death penalty, riot police, gun, heroin, sheriff*
	Mid-size cities	*kansas city, fla, lauderdale, palm beach, salt lake, lauderdale fla, lake city, fort lauderdale, beach fla, west palm, jacksonville, tampa, tampa fla, raton fla, raton, jacksonville fla, boca, fort, boca raton, tucson ariz*
	NY politics	*blasio, cuomo, gov andrew, mayor bill, mayor michael, city council, andrew cuomo, city mayor, mayor, city hall, mayoral, borough, councilman, york city, bronx, york gov, staten island, staten, city official, long island*
	Rail/trucking/shipping	*railroad, norfolk southern, conrail, csx, rail, union pacific, amtrak, burlington northern, norfolk, freight, ship company, truckers, shippers, truck company, transit, truck driver, port, railway, locomotive, uber*
	California	*los angeles, angeles, los, francisco, san francisco, diego, san diego, orange county, santa monica, angeles county, san, oakland, bay area, santa barbara, santa ana, los angelesbased, monica calif, angelesbased, northern california, san antonio*
	Rental properties	*square feet, square foot, vacancy rate, tenant, office space, landlord, office tower, squarefoot, retail space, highrise, apartment build, rent, redevelopment, tower, condo, midtown, square, time square, downtown, apartment*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Disease	*disease control, aid virus, cdc, breast cancer, breast, infectious disease, heart disease, virus, flu, diagnosed, implants, immune system, infected, immune deficiency, deficiency syndrome, acquire immune, infectious, symptom, infection, world health*
	US defense	*rumsfeld, missile defense, marine corps, joint chief, donald rumsfeld, strategic defense, defense secretary, naval, senate arm, defense official, military service, cruise missile, defense initiative, nuclear missile, maj, pentagon official, arm service, military personnel, aircraft carrier, armor*
	Pharma	*pfizer, generic, bristolmyers, lilly, generic version, drug company, generic drug, eli lilly, novartis, drug maker, scheringplough, pharmaceutical industry, genentech, beecham plc, generic competition, roche, bristolmyers squibb, astrazeneca, squibb, drug price*
	Aerospace/defense	*force contract, lockheed, mcdonnell, mcdonnell douglas, northrop, pratt whitney, grumman, lockheed martin, airbus, general dynamic, european aeronautic, aeronautic defence, northrop grumman, martin marietta, defence space, raytheon, plane maker, airbus industrie, defense contractor, jet engine*
	Automotive	*auto maker, domestic car, import car, ford, sportutility, chevrolet, chrysler, mazda, toyota, eestimated, car maker, small car, sportutility vehicle, japanese auto, car market, honda, sedan, vehicle sale, minivan, ford motor*
	Airlines	*airline, texas air, twa, american airline, airways, pilot association, pilot union, flight attendant, load factor, northwest airline, usair, ual corps, unit airline, delta air, air line, line pilot, america west, delta, amr corps, continental airline*
	Retail	*walmart, department store, store open, walmart store, penney, kmart, retail analyst, comparablestore, bentonville, bentonville ark, largest retailer, neiman marcus, bloomingdales, specialty store, costco, discount store, macy, comparablestore sale, departmentstore, neiman*
Government	Political contributions	*donation, political contribution, philanthropy, campaign contribution, charity, nonprofit organization, nonprofit group, philanthropic, donor, donate, action committee, pac, nonprofit, political action, foundation, charitable, federal election, campaignfinance, advocacy, fundraising*
	Regulation	*propose rule, rule govern, rule require, final rule, federal regulation, consumer protection, regulation, rule change, current rule, protect consumer, federal rule, rulemaking, report requirement, regulate, government regulation, ftc, stricter, propose legislation, requirement, state regulation*
	Environment	*epa, emission, protection agency, environmental protection, carbon dioxide, dioxide, greenhouse gases, clean air, environmental, environmental group, greenhousegas, greenhousegas emission, air act, superfund, global warm, pollutants, hazardous waste, environmentalists, greenhouse, carbon*
	Private/public sector	*private sector, private enterprise, private business, public sector, private firm, private company, privatization, government aid, privatize, taxpayer money, private capital, privatesector, federal government, government money, government subsidy, government intervention, local government, private owned, state enterprise, government plan*
	State politics	*rhode island, state legislature, state include, west virginia, state legislator, schwarzenegger, north carolina, rhode, washington state, state lawmaker, legislature, state official, carolina, state house, republican gov, state level, virginia, outofstate, state capital, state budget*
	Watchdogs	*inspector general, gao, account office, general account, accountability office, government accountability, page report, report conclude, inspector, review process, house subcommittee, release report, correspondence, issue report, oversight, internal review, documentation, receive letter, wrote letter, examination*
	Utilities	*nrc, megawatt, nuclear regulatory, pge, nuclear plant, electricity, electric utility, public utility, utility commission, power company, power plant, energy regulatory, power generation, grid, edison, utility hold, nuclear power, utility company, power light, nuclearpower plant*
	Safety administrations	*safety board, traffic safety, highway traffic, safety administration, transportation safety, osha, national highway, occupational safety, health administration, accident, safety standard, national transportation, air bag, safety, crashes, safety problem, safety issue, occupational, fatalities, firestone*
	National security	*cybersecurity, security agency, hacker, intelligence committee, fbi, cia, espionage, homeland security, security expert, senate intelligence, hacking, intelligence agency, spy, fbi agent, encryption, surveillance, security number, central intelligence, intelligence, breaches*
Courts	Justice Department	*spitzer, antitrust division, general eliot, receive subpoena, eliot spitzer, department antitrust, department investigation, justice department, internal investigation, general office, investigate allegation, subpoena, cooperate fully, deputy attorney, probe, york attorney, attorney general, eliot, assistant attorney, criminal probe*
	Indictments	*indictment, plead guilty, plead, plea agreement, mail fraud, face maximum, guilty plea, wire fraud, prosecutor allege, guilty, federal grand, indicted, grand jury, conspiracy, felony count, allege scheme, defense attorney, criminal trial, acquitted, security fraud*
	Courts	*supreme court, appeal court, scalia, federal appeal, chief justice, high court, justice william, william rehnquist, antonin, court rule, court decision, justice john, supreme, lowercourt, antonin scalia, anthony kennedy, justice antonin, threejudge, oral argument, threejudge panel*
	Lawsuits	*punitive damage, suit file, classaction, plaintiff lawyer, class action, suit allege, plaintiff attorney, lawsuit file, compensatory damage, classaction suit, suit seek, file suit, classaction status, seek classaction, damage award, suit, unspecified damage, plaintiff, file lawsuit, seek unspecified*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
Political Leaders	Clintons	*lewinsky, whitewater, rodham clinton, hillary rodham, rodham, clinton administration, president clinton, independent counsel, impeachment, janet reno, clinton, starr, clinton plan, house counsel, rubin, watergate, robert rubin, bill clinton, paula, reno*
	US Senate	*majority leader, senate bill, daschle, house bill, hastert, senate majority, appropriations committee, senate floor, senate version, gop leadership, lott, minority leader, senate appropriations, leader trent, senate democrat, house floor, committee chairman, baucus, leader harry, house gop*
	Reagan	*contra, reagan administration, president reagan, shultz, george shultz, state george, regan, irancontra, james baker, secretary james, nicholas brady, reagan, secretary nicholas, assistant secretary, nicaraguan, meese, baker, carter administration, deputy assistant, sandinista*
	Bush/Obama/Trump	*obama administration, trump administration, president donald, president barack, bush administration, president dick, president bush, president joe, house press, barack, barack obama, bush plan, president obama, obama, cheney, dick cheney, president george, senior administration, bush, presidentelect*
	Elections	*sen obama, romney, sen mccain, dukakis, mitt romney, democratic primary, republican primary, mitt, gop primary, run mate, republican presidential, gop candidate, caucuses, republican nominee, senate race, republican candidate, democratic candidate, rubio, marco rubio, democratic nominee*
	European politics	*opposition party, socialist party, ldp, parliamentary election, centerleft, parliamentary, rule party, centerright, rule coalition, main opposition, opposition politician, liberal party, conservative party, opinion poll, national election, party member, upper house, coalition government, national assembly, referendum*
Terrorism/Mideast	Middle east	*palestinian, gaza, hamas, serb, gaza strip, israeli, plo, israel, hezbollah, palestinian authority, arafat, mubarak, assad, milosevic, west bank, palestinian leader, netanyahu, israeli official, damascus, lebanese*
	Nuclear/North Korea	*pyongyang, north korea, north korean, kim jong, nuclear program, daewoo, korean, seoul south, jong, kim, iran nuclear, nuclearweapons program, nonproliferation, south korean, international atomic, korea, nuclearweapons, korean government, enrichment, korean company*
	Terrorism	*taliban, karzai, afghan, qaeda, kabul, pakistani, islamabad, musharraf, hamid karzai, afghan government, kashmir, president hamid, islamic state, osama, laden, pakistan, pervez, extremist group, hamid, suicide bomber*
	Iraq	*iraqi, saddam, saddam hussein, kurds, baghdad, iraqi leader, iraqi government, erdogan, iraqi official, ankara, northern iraq, tayyip, abu dhabi, dhabi, recep, recep tayyip, bahrain, turkish, hussein, persian*
International Affairs	Russia	*yeltsin, moscow, putin, russian, kremlin, gorbachev, kiev, ukraine, russian president, boris yeltsin, president vladimir, medvedev, vladimir putin, ruble, ukrainian, mikhail gorbachev, vladimir, sergei, mikhail, gazprom*
	Trade agreements	*gatt, wto, tariff, trade representative, antidumping, trade pact, protectionist, trade official, general agreement, freetrade agreement, trade talk, freetrade, trade policy, bilateral trade, trade barrier, unfair trade, trade dispute, protectionism, trade war, trade deal*
	Latin America	*argentina, chavez, latin, latin america, latin american, hugo chavez, cuban, colombia, president hugo, paulo brazil, castro, silva, argentine, puerto, puerto rico, fidel castro, caracas, havana, fidel, luiz*
	Japan	*japanese company, tokyo japan, japan largest, japan biggest, shinzo, minister shinzo, abe, japanese business, strong yen, japan ministry, japanese consumer, japanese bank, japan bank, japan economy, japanese firm, osaka, japanese market, japan financial, big japanese, major japanese*
	Canada/South Africa	*south africa, quebec, canadian government, toronto stock, british columbia, calgary, canadian cent, mandela, ottawa, calgary alberta, johannesburg, canada largest, africa, vancouver, apartheid, vancouver british, nova scotia, canadian, scotia, alberta*
	China	*yuan, beijing, xinhua, chinese official, chinese government, chinese company, chinese, zhu, beijing china, china sea, chinese authority, chinese leader, jiang, wen, xinhua news, zhang, chinese bank, deng, guangzhou, china central*
	Southeast Asia	*jakarta, thailand, southeast asia, rupiah, southeast asian, suharto, indonesian, baht, indonesia, bhd, jakarta indonesia, thai, malaysia, malaysian, hutchison whampoa, hutchison, philippine, pte, asia pacific, singapore dollar*
	Germany	*east german, helmut kohl, bonn, chancellor helmut, helmut, german company, east germany, german official, west germany, chancellor gerhard, germany largest, gerhard, german chancellor, german bank, gmbh, berlin, supervisory board, merkel, german economy, german government*
	France/Italy	*generale, chirac, societe generale, french government, sarkozy, societe, french bank, belgium, francois, french company, cie, italian, french, belgian, jacques chirac, italian bank, president francois, french president, president jacques, italy*
	UK	*gbp, scotland, hold plc, scotland group, plc british, british bank, british company, scottish, group plc, unit kingdom, minister margaret, london british, lloyd bank, plc, westminster, bank plc, westminster bank, edinburgh, gordon brown, britons*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
Entertainment	Music industry	*album, record label, song, record company, music industry, universal music, music group, music company, emi, musicians, music, concert, warner music, jazz, rap, band, singer, musical, singing, promoter*
	Broadcasting	*clear channel, broadcaster, broadcast, cable channel, broadcast network, television station, cos nbc, nbc, viewership, liberty media, news network, television network, cable news, primetime, national broadcast, cbss, news channel, espn, channel, corps fox*
	Publishing	*barnes noble, circulation, time mirror, random house, magazine publisher, daily news, gannett, newspaper publisher, daily newspaper, chicago tribune, newsweek, publisher, publishes, barnes, journalism, newspaper, simon schuster, tribune, publish company, manage editor*
	Marketing	*saatchi, interpublic, omnicom, ddb, rubicam, saatchi saatchi, bbdo, wpp, young rubicam, publicis, omnicom group, wpp group, walter thompson, interpublic group, leo burnett, ogilvy mather, foote cone, mccannerickson, ogilvy, mediabuying*
	Movie industry	*boxoffice, hollywood, paramount picture, motion picture, dreamworks, studio, sony picture, eisner, hollywood studio, mattel, theme park, movie, disney, hasbro, movie studio, columbia picture, animation, marvel, sequel, theatrical*
Social/Cultural	Economic ideology	*letter letter, relate letter, hoover institution, moral, liberalism, enterprise institute, american enterprise, morality, hoover, franklin roosevelt, capitalism, roosevelt, civilization, rightly, truman, freedom, myth, selfinterest, welfare state, courage*
	Schools	*public school, private school, high school, school, curriculum, test score, teacher union, school board, student, school district, school system, educators, classroom, teacher, elementary, elementary school, graduation, faculty, education department, graduate*
	Sales call	*conference room, salesman, salesmen, boss, phone call, hotline, colleague, receive call, desk, stunned, huddled, caller, wanted, summoned, knew, began work, replied, salespeople, staffer, facetoface*
	Cultural life	*dog, cowboy, deer, jacket, shirt, sweat, tshirt, cat, wear, shower, horse, purple, dirt, dusty, wooden, baseball, bike, tennis, stadium, shiny*
	Arts	*sothebys, museum, auction house, art, collector, gallery, portrait, exhibit, thcentury, artistic, paint, statue, exhibition, monument, opera, auctioneer, collection, historian, archive, tribute*
	Immigration	*immigrant, illegal immigrant, migrant, census, census bureau, immigration law, undocumented, immigration, citizenship, illegal immigration, migration, legal status, deportation, poverty, inequality, population growth, residency, latino, hispanic, food stamp*
	Positive sentiment	*awful lot, pretty good, spend lot, good time, awful, good idea, making money, people feel, good deal, big money, feel good, pretty, good reason, lot, spend money, lose money, luck, feel comfortable, hard time, bad idea*
	Humor/language	*phrase, fool, word, humor, polite, translation, stupid, funny, emotion, tongue, honest, dumb, jargon, metaphor, honesty, silly, joke, truth, answer, fascinating*
	Gender issues	*female, young women, mother, daughter, divorce, father, young children, sexual harassment, family member, wives, male, husband, women, dad, child, married, young adult, mom, marry, sex*
Challenges	Changes	*major shift, shift, fundamental change, big change, major change, significant change, rethink, rapid change, direction, change, dramatic, emphasis, turn point, chang, reversal, urgency, move closer, move quick, reshape, focus*
	Key role	*key role, play role, play key, play major, major role, important role, big role, play big, lead role, active role, central role, role, bigger role, play, close relationship, potential conflict, play part, relationship, play field, influence*
	Problems	*problem face, biggest problem, big problem, matter worse, major problem, real problem, bigger problem, similar problem, solve, fundamental problem, problem, predicament, mess, plague, debacle, fiasco, disastrous, financial problem, problem include, inability*
	Challenges	*uphill battle, uphill, face tough, face pressure, battle, intense pressure, challenge face, status quo, quo, face challenge, increase pressure, tactic, fight, bruising, tough decision, backfire, backlash, bitter, face grow, persuade*
	Small possibility	*door open, time frame, frame, door, move quick, left open, wait, leave, ready, takes place, open, exit, left, stay, lock, quick, bringing, hand, final, rush*
	Spring/summer	*early march, late summer, close watch, late march, watch close, early sign, early stage, summer, watch, show sign, dash, early, spring, stage, warn sign, midmarch, turn point, slow start, began, fade*
	Long/short term	*long term, short term, long run, longer term, long time, term limit, short run, long period, long haul, long, term, long history, long overdue, long list, longer, short time, short period, shorter, short, sooner*
Science/Language	Research	*survey conduct, respondents, survey show, survey, survey release, methodology, data compile, statistical, gallup, study release, data show, percent, study show, compile, satisfaction, measurement, statistics show, study suggest, snapshot, sample*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Scenario analysis	*financial analysis, commentary, worstcase scenario, worstcase, analysis, foreseeable, foreseeable future, conventional wisdom, imply, scenario, probability, adverse effect, plausible, theoretical, certainty, wisdom, consequence, assumption, negative effect, theory*
	Programs/initiatives	*program aim, goal, broader effort, top priority, pilot program, initiative, ambitious, achieve, program design, high priority, priority, revamp, aim, important step, remain committed, ambitious plan, objective, making progress, progress, development program*
	Biology/chemistry/physics	*gene, dna, mice, molecules, biology, scientist, physics, molecular, national laboratory, genetic, fusion, lab, massachusetts institute, physicist, mit, cell, research team, chemistry, nobel prize, experiment*
	Space program	*nasa, space administration, national aeronautic, space station, space shuttle, spacecraft, astronauts, orbit, booster, mann, shuttle, rocket, unmanned, launches, moon, satellite, robot, space, drone, aeronautic*
	Systems	*task force, task, current system, centralized, decentralized, system work, complexity, efficient, inefficiency, workable, cumbersome, innovation, antiquated, inefficient, function, rigid, flexible, solution, mechanism, find ways*
	Size	*big chunk, chunk, large chunk, smaller, bigger, smaller company, rough double, rely heavily, size, rapid grow, larger company, larger, bigger share, big reason, grow rapid, huge amount, larger share, grow number, grow fast, big part*
	Wide range	*wide range, vary wide, vast majority, broad range, wide variety, large number, limit number, large portion, vary, wide, small portion, unusual large, large amount, significant number, significant portion, small number, significant amount, wide considered, vast, range*
Activism/Language	Activists	*icahn, carl icahn, investor carl, getty, pennzoil, carl, texaco, white plain, source close, holmes, interfere, forestall, attempt, source, phillip, plain, nybased, control, intention, apparent*
	Announce plan	*announce plan, plan announce, unveil plan, announce earlier, previous announce, yesterdays announcement, announce restructure, company announce, outline plan, latest move, announce, plan include, restructure plan, move aim, current plan, announcement, company plan, plan call, expansion plan, unveil*
	Major concerns	*raise concern, major concern, express concern, concern, increase concern, voice concern, concern raise, cit, held, addition, grow concern, significant, continue, remain, considered, represent significant, add, fully, represent major, lead*
	Futures/indices	*spot index, dollar yen, brother treasury, commodity oil, oil barrel, treasury index, future index, bond lehman, stock volume, jonesaig future, vol share, nasdaq vol, vol, jonesaig, bond shearson, market stock, share nasdaq, volume share, yld, industrial rose*
	Corrections/amplifications	*amplifications, correction amplifications, article correction, incorrect, news article, incorrect state, relate article, incorrect report, article, correction, misstated, york contribute, wsj, wsj sept, contribute, wsj march, source wsj, wsj research, money invest, washington contribute*
	Buffett	*buffett, berkshire, hathaway, berkshire hathaway, warren buffett, warren, omaha neb, omaha, neb, billionaire investor, brandon, cit, billionaire, significant, large part, continue, part due, charlie, held, substantial*
	Mexico	*mexican, salinas, mexican government, grupo, mexico, mexico city, ernesto, lopez, mexican peso, telefonos, president carlo, peso, enrique, nafta, raul, miguel, slim, gonzalez, luis, american free*
Negotiations	Restraint	*temporary restrain, restrain order, postponement, postpone, delay, extension, indefinite, temporarily, expiration date, restrain, deadline, halt, cancel, temporary, extend, cancellation, schedule, resume, suspension, shutdown*
	News conference	*news conference, told news, press conference, told reporter, conference, press release, news release, remark, attendees, public appearance, reporter, conference call, public statement, statement release, issue statement, apology, transcript, industry conference, apologize, forum*
	Company spokesperson	*call seek, seek comment, return call, spokesman confirm, return phone, spokesman decline, spokesman add, return telephone, company spokesman, comment spokesman, official decline, company spokeswoman, immediately respond, decline comment, spokeswoman decline, company decline, elaborate, unavailable, comment, spokeswoman*
	People familiar	*people familiar, familiar, person familiar, person close, people close, people involve, person, matter, individual familiar, executive familiar, potential deal, situation, briefed, earlier report, official familiar, potential sale, knowledgeable, discussion, unclear, remain unclear*
	Agreement reached	*definitive agreement, sign letter, sign definitive, definitive, reach definitive, agreement call, lone star, preliminary agreement, agreement include, agreement reach, disclose term, agreement sign, sign agreement, tentative agree, agreement, financial term, final agreement, license agreement, reach agreement, lone*
	Committees	*panel member, advisory committee, advisory panel, recommendation, proposal include, proposal call, advisory group, final decision, panel, vote unanimous, advisory, commission report, advisory board, proposal, committee member, public hear, recommend, commission approve, unanimous, final report*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Negotiations	*stick point, negotiator, negotiate table, side agree, negotiation, impasse, reach deal, reach compromise, resolve dispute, unresolved, talk aim, stumble block, negotiate, haggling, talk, contentious issue, start talk, hold talk, deadlock, resolve*
Industry	Soft drinks	*coke, bottler, pepsi, softdrink, cola, cocacola, soft drink, pepsico, north america, beverage company, schweppes, soda, north american, pepper, beverage, fountain, cadbury schweppes, american operation, volume growth, cadbury*
	Small business	*smallbusiness, small business, smallbusiness owner, business owner, national federation, independent business, small firm, small company, business administration, mediumsize business, large corporation, mediumsize, midsize company, big business, entrepreneurship, big corporation, entrepreneur, businesss, business community, large company*
	Cable	*warner cable, time warner, aol time, warner, parsons, warner communication, levin, adelphia, aol, warner incs, bronfman, buy time, cox, turner, hbo, comcast, cable system, america online, armstrong, media giant*
	Fast food	*franchisee, taco bell, burger king, fastfood, rite aid, starbucks, pizza hut, mcdonald, taco, fastfood chain, pizza, rite, restaurant, restaurant chain, burger, wendy, fastfood restaurant, hut, sandwiches, hamburger*
	Competition	*goodyear, stiff competition, tire rubber, increase competition, fierce competition, goodyear tire, competitive advantage, competitive market, increase competitive, high competitive, intense competition, tire maker, competitive environment, competition, face stiff, competitive pressure, akron ohio, tough competition, akron, grow competition*
	Chemicals/paper	*union carbide, carbide, specialty chemical, georgiapacific, chemical company, pont, weyerhaeuser, chemical concern, chemical business, dupont, chemical maker, paper product, hercules, international paper, paper company, coatings, danbury, chemical industry, wilmington, forestproducts*
	Venture capital	*venture, joint venture, venture capitalist, venture capital, venturecapital firm, venturecapital, form joint, venture partner, venture fund, venturecapital fund, jointventure partner, joint, capitalist, startup company, jointventure, byers, startup, perkins, joint owned, form partnership*
	Tobacco	*tobacco, smokers, tobacco company, cigarette, philip morris, nicotine, tobacco industry, morris, cigarette maker, reynolds tobacco, philip morriss, altria, morriss, antismoking, marlboro, morris cos, bradstreet, dun, dun bradstreet, nabisco hold*
	Subsidiaries	*owned subsidiary, wholly owned, spinoff, itt, parent company, unit include, separate company, wholly, noncore, business unit, noncore asset, operate unit, spin, operation include, service unit, service division, spun, divestiture, subsidiary, core business*
	Credit cards	*mastercard, credit card, debit card, debit, creditcard company, card, creditcard issuer, american express, creditcard, expresss, creditcard business, check account, visa, teller machine, teller, payment system, hallmark, prepaid, express, merchant*
	Couriers	*postal service, fedex, postal, unit parcel, ups, parcel service, post office, federal express, parcel, mail, envelope, firstclass, memphis tenn, direct mail, memphis, downs, delivery system, airborne, mailorder, stamp*
	Foods/consumer goods	*sara lee, general mill, nestle, hershey, avon, campbell soup, diapers, detergent, cereal, toothpaste, heinz, quaker oats, ice cream, sara, information resource, unilever, quaker, yogurt, shampoo, chocolate*
	Insurance	*policyholders, life insurance, reinsurance, insurance department, insurance regulator, allstate, insurance commissioner, mutual life, life insurer, insurance business, insurance broker, propertycasualty, metlife, insurance unit, metropolitan life, insurance operation, life casualty, aig, lifeinsurance, aetna life*
	Luxury/beverages	*lvmh, vuitton, hennessy louis, louis vuitton, moet, moet hennessy, lvmh moet, beer, brewer, wine, coors, brewery, anheuserbusch cos, liquor, brew, anheuserbusch, scotch, whiskey, luxurygoods, budweiser*
	Casinos	*casino, atlantic city, marriott, hilton, carnival, hilton hotel, hotel chain, hotel room, inn, caesars, resort, mirage, hotel, hyatt, poker, casino operator, vegas, circus, hospitality, lodge*
	Revenue growth	*growth strategy, growth opportunity, annual revenue, boost revenue, growth area, company revenue, total revenue, company growth, revenue growth, growth potential, core business, generate revenue, sanford bernstein, post revenue, future growth, growth plan, company core, fastergrowing, revenue stream, increase revenue*
Technology	Internet	*facebook, web site, search engine, alibaba, yahoo, ebay, youtube, home page, internet user, online advertise, portal, socialnetworking, social network, internet company, web, twitter, web page, online, internet site, amazon*
	Mobile devices	*iphone, smartphone, android, handset, ipad, blackberry, mobile device, nokia, apple incs, mobile phone, tablet, rim, cellphone, qualcomm, ipod, steve job, mobile, galaxy, telefon, text message*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Electronics	*playstation, nintendo, console, kodak, digital camera, lcd, flatpanel, xbox, cdrom, electronic art, videogame, photographic, video game, lenses, highdefinition, panasonic, compact disk, camera, photography, stereo*
	Phone companies	*att, sprint, longdistance, local phone, bell atlantic, phone company, baby bell, bell company, mci, longdistance company, regional bell, mccaw, phone service, qwest, longdistance service, bellsouth, nynex, telephone company, ameritech, nextel*
	Computers	*ibm, compaq, mainframe, dell, mainframe computer, machine corps, armonk, computer giant, packard, computer business, workstation, big blue, supercomputer, cray, compaq computer, computer company, computer associate, emc, computer market, business machine*
	Software	*linux, window operate, ballmer, microsoft window, redmond wash, redmond, microsoft, oracle, software giant, lotus, software industry, ellison, bill gate, spreadsheet, netscape, novell, sap, adobe, window, java*
	Microchips	*amd, advance micro, memory chip, santa clara, semiconductor industry, clara, intel chip, intel, clara calif, chip maker, chip, memorychip, computer chip, micro device, micron, pentium, cisco, micro, national semiconductor, seagate*
Oil & Mining	Steel	*steel price, steel industry, steelmaker, alcan, steel product, nucor, steel producer, steel, steel mill, steel company, aluminum, bethlehem steel, bethlehem, steelmaking, alcoa, usx, newsprint, smelter, raw material, metric*
	Mining	*bhp, rio tinto, tinto, newmont, bhp billiton, mining company, miner, mining, billiton, newmont mining, mining concern, sydney australia, anglo american, gold producer, goldmining, coal miner, mines, melbourne, anglo, australian*
	Machinery	*machine tool, manufacture plant, whirlpool, manufacture facility, construction equipment, manufacture operation, manufacture capacity, appliance maker, production facility, cummins, production line, factory, manufacture company, supply chain, deere, caterpillar, machinery, manufacture, wash machine, assembly line*
	Oil drilling	*exxon, drill, cubic feet, mobil, exxon mobil, drill rig, offshore oil, chevron, amoco, royal dutchshell, occidental, oil spill, dutchshell, cubic, arco, dutchshell group, deepwater, offshore drill, liquefied natural, conocophillips*
	Agriculture	*genetically modified, farm, growers, poultry, madcow, farmer, ranchers, tyson food, agriculture secretary, tyson, farmland, cow, usda, agriculture, agribusiness, cargill, tomatoes, agricultural, beef, fertilizer*
	Oil market	*opec, nonopec, opec member, oil minister, oil demand, world oil, brent crude, oil market, oil exporter, export country, petroleum export, refiner, oil supply, international energy, crude price, oil price, higher oil, high oil, oil export, oil refiner*
Economic Growth	Record high	*record level, highest level, reach record, hit record, previous record, highest, pentup demand, high level, pentup, alltime, alltime high, strong demand, record high, set record, higher level, helped fuel, remain strong, helped drive, helped push, surpass*
	Economic growth	*gdp growth, productivity growth, gdp, domestic product, gross domestic, economy grew, national product, economy expand, gross national, growth forecast, economic growth, business cycle, economic output, gnp, economic performance, growth rate, annual growth, slow economic, economy grow, economic expansion*
	Federal Reserve	*greenspan, alan greenspan, yellen, federalfunds rate, federalfunds, fomc, chairman alan, raise rate, janet yellen, inflation expectation, reserve chairman, quarterpoint, policy meet, market committee, target rate, federal open, bernanke, cut rate, rate hike, raise interest*
	European sovereign debt	*ecb, imf, eurozone, euro zone, monetary fund, international monetary, draghi, eurozone country, trichet, ecb president, jeanclaude trichet, greece, eurozone economy, greek, athens, austerity measure, bailout fund, european central, debt crisis, sovereigndebt*
	Recession	*economic downturn, steep drop, steep decline, downturn, bright spot, hardest hit, hit hard, weak economy, hit bottom, hard hit, economic slowdown, weak demand, sharp drop, economic slump, slump, weaken economy, steep, lowest level, gloomy, sharp decline*
	Product prices	*raise price, increase price, price reduction, price increase, price cut, wholesale price, higher cost, cut price, retail price, reduce price, price power, rise cost, higher price, cost increase, slash price, list price, discount price, price war, increase cost, price tag*
	Optimism	*cautious optimistic, remain optimistic, remain cautious, remain uncertain, uncertainty surrounding, positive sign, optimistic, overly optimistic, cautious, remain confident, remain skeptical, express confidence, positive impact, positive, optimism, uncertain, remain concern, uncertainty, remain wary, economic uncertainty*
	Macroeconomic data	*seasonal variation, durable good, construction spend, seasonal adjust, nonresidential, business inventory, nondurable, durable, conference board, jobless claim, manager index, factory order, purchase manager, supply management, manufacture activity, seasonal, initial claim, adjust annual, lead indicator, adjust basis*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
Corporate Earnings	Profits	*net profit, profit rose, profit fell, firsthalf net, profit increase, profit jump, operate profit, profit climb, profit surge, pretax profit, profit dropped, profit rise, firsthalf, profit decline, group sale, group net, firsthalf profit, group profit, fullyear net, post rise*
	Revised estimate	*previous estimate, original estimate, preliminary estimate, revision, earlier estimate, revise downward, downward revision, initial estimate, revise, substantial increase, significant increase, estimate, official estimate, significant impact, substantial reduce, previous thought, government estimate, figure, figure show, revise wsj*
	Earnings losses	*loss include, net loss, discontinue operation, operate loss, loss widen, pretax charge, discontinue, aftertax charge, post loss, restructure charge, writedown, include loss, loss narrow, restructure cost, pretax loss, expect loss, loss total, report loss, impairment, special charge*
	Small changes	*typical require, average yield, decline slight, decrease, slight increase, remain unchanged, remain flat, sharp increase, part reflect, slight decline, increase reflect, continue decline, fell slight, slight, jumbo, slight higher, period, total number, rose slight, attribute*
	Financial reports	*ended sept, forma, pro forma, ended march, latest fiscal, march compare, current fiscal, full fiscal, ended, company fiscal, period ended, fiscal, sept, fiscalyear, pro, fiscal earn, report revenue, report sale, company expect, march*
	Earnings forecasts	*adjust earn, analyst earn, analyst poll, forecast earn, earn forecast, earningspershare, earn expectation, beat expectation, earn guidance, earn estimate, earn projection, earn growth, earn release, company forecast, analyst forecast, analyst expect, report earn, project earn, earn season, revenue forecast*
	Earnings	*cent dilute, dilute share, share revenue, share relate, callthomson financial, cent share, callthomson, call consensus, pershare figure, thomson financialfirst, share exclude, financialfirst call, financialfirst, earlier exclude, period revenue, income total, net income, dilute, share compare, exclude special*
Financial Markets	IPOs	*ipo, ipo market, initial public, public offer, ipo price, firstday, public stock, stock offer, roadshow, offer price, followon, price range, secondary offer, share offer, public market, initial offer, lockup, prospectus, company initial, lead underwriter*
	Bond yields	*yield rise, highyield bond, sell bond, yield fall, corporatebond, bond issue, treasury yield, issue bond, municipalbond, bond sale, higher yield, treasury debt, highyield, high yield, bond portfolio, bond deal, risk premium, muni, buy bond, governmentbond*
	Short sales	*short seller, short interest, short sell, borrow stock, borrow share, shortselling, shortsellers, short sale, sell borrow, seller, short position, prospective buyer, find buyer, potential buyer, buyer, short, investor sell, share sold, purchaser, foreign buyer*
	Small caps	*york smallcapitalization, poor smallcap, smallcapitalization stock, russell index, nyse rose, smallcap index, smallcap market, nyse fell, nasdaq gain, smallcapitalization, nasdaq fell, nasdaq rose, nasdaq advance, smallcap stock, initiate coverage, nasdaq decline, smallcap, add cent, market perform, lost cent*
	Treasury bonds	*accept noncompetly, cusip, bid widen, price accept, noncompetitive tender, coupon equivalent, application accept, price rate, cusip number, bill mature, noncompetly, bid narrow, accept noncompetitive, rate coupon, bid accept, fourweek bill, minimum denominations, marketclearing, competitive tender, equivalent yield*
	International exchanges	*perspective geneva, dax stock, australia world, switzerland australia, germany hong, firstsection, firstsection volume, international perspective, kong switzerland, index calculate, world index, index eafe, mscip, stanley capital, xetra dax, timesstock exchange, firstsection issue, msci indexes, share index, japan france*
	Exchanges/composites	*composite trade, exchange composite, exchange trade, market composite, york stock, stock exchange, adrs, share close, board trade, composite, close unchanged, depositary, share rose, depositary receipt, board composite, trade share, american depositary, share fell, depositary share, stock close*
	Currencies/metals	*dollar british, frankfurt gold, home mkt, dollar french, london toronto, comex base, tokyo zurich, comex london, yen swiss, ounce comex, toronto paris, yen sterling, pound canadian, yen late, franc japanese, franc west, mark base, dollar stood, mark late, paris tokyo*
	Share payouts	*nyse symbol, share buyback, stock buyback, dividend increase, buyback, buyback program, share repurchase, share common, pay dividend, time estimate, stock repurchase, dividend yield, daily trade, dividend payment, dividend, special dividend, cash dividend, annual dividend, average daily, share share*
	Bear/bull market	*bear market, bull market, stock pick, stockpicking, priceearnings ratio, growth stock, pricetoearnings ratio, bull, priceearnings, pricetoearnings, contrarian, stock portfolio, wilshire, foreign stock, ned, stockmarket, overvalued, investment strategist, big stock, beatendown*
	Commodities	*cocoa exchange, comex division, sugar cocoa, coffee sugar, soybean future, coffee future, lifeofcontract, pork belly, wheat future, york cotton, copper future, march contract, commodity market, william oneill, commodity fund, corn future, contract rose, contract fell, cotton exchange, future price*
	Trading activity	*action stock, fell bond, market action, otc trade, volume total, investor report, issue advance, board volume, index exclude, transportation average, otc market, otc, share chang, stock advance, board issue, bond gain, professional investor, decline stock, euro rose, active issue*

Continued on next page

Table 7 – *Continued from previous page*

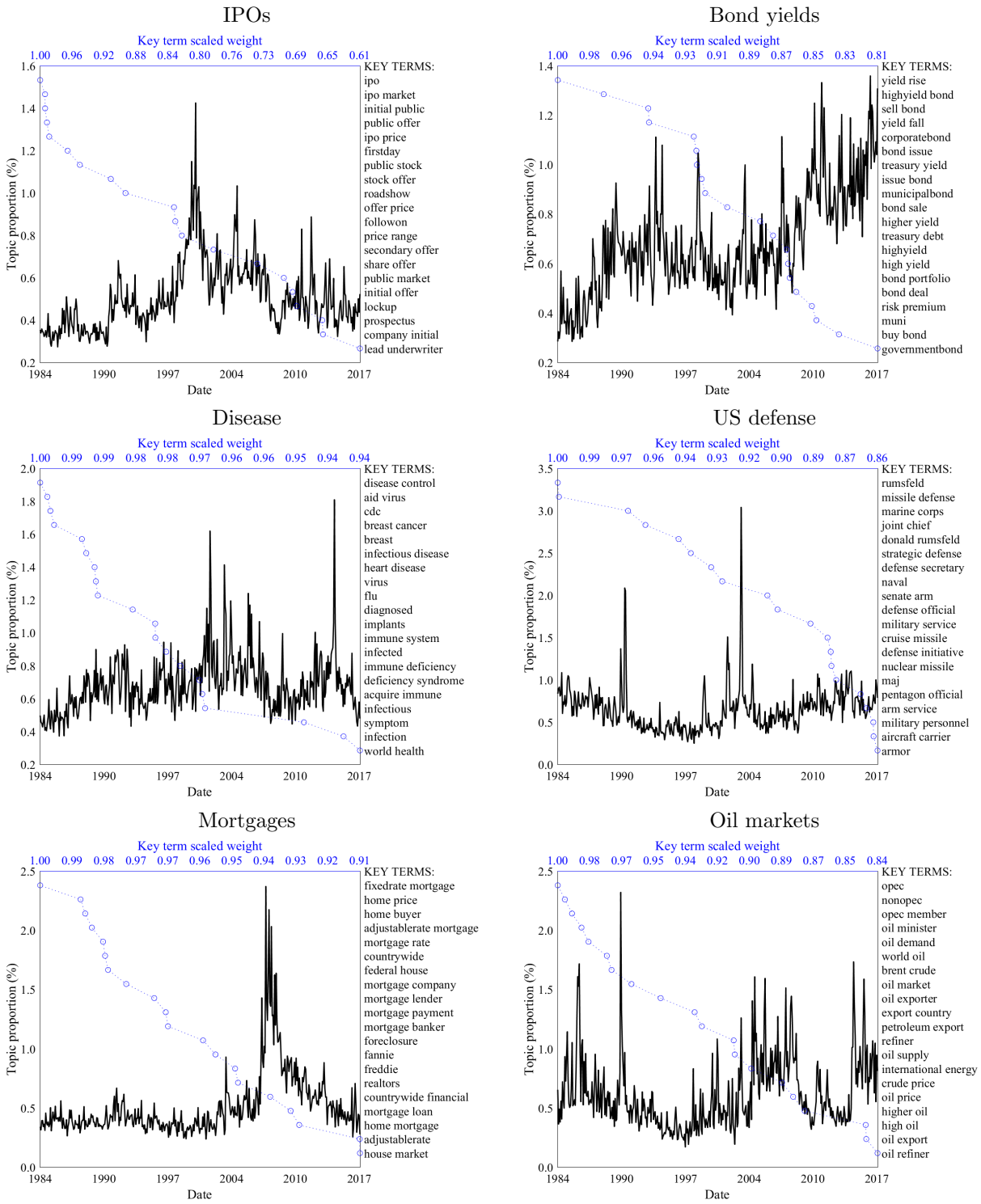
Metatopic label	Topic label	Key terms
	Options/VIX	*option exchange, volatility index, board option, contract compare, put trade, option strategist, option market, vix, cboe, option trader, call trade, cftc, option trade, euronext, call option, trade contract, put option, option contract, future exchange, exchange official*
Buyouts & Bankruptcy	M&A	*combine company, merge company, propose merger, combine entity, merger plan, merger talk, planned merger, merger, planned acquisition, antitrust concern, merge, deal close, propose acquisition, synergy, deal announce, deal valued, merger partner, antitrust regulator, acquire close, consolidation*
	Control stakes	*control stake, minority stake, majority stake, buy stake, abn, sell stake, remain stake, bought stake, minority shareholder, owns stake, acquire stake, hold stake, abn amro, amro, ownership stake, stake, company controlled, large stake, control interest, equity stake*
	Drexel	*drexel, leverage buyout, kohlberg kravis, drexel burnham, transaction finance, kohlberg, kravis robert, kravis, boesky, pickens, burnham lambert, burnham, buyout group, lambert, investor acquire, fund generate, ivan boesky, finance large, boone pickens, hicks*
	SEC	*sec file, sec, sec allege, sec official, commission file, sec enforcement, sec chairman, exchange commission, sec staff, sec investigation, sec spokesman, sec charge, arthur levitt, nonpublic information, chairman arthur, nonpublic, file show, violate security, enforcement division, sec rule*
	Bankruptcy	*bankruptcy, unsecured creditor, bankruptcy court, federal bankruptcy, bankruptcy code, chapter company, creditor lawsuit, bankruptcy judge, reorganization plan, bankruptcy protection, bankruptcylaw, chapter bankruptcy, chapter, creditor committee, court protection, bankruptcy file, bankruptcy proceed, bankruptcylaw protection, bankruptcy law, chapter protection*
	Corporate governance	*annual meet, shareholder meet, proxy contest, corporate governance, independent director, dissident shareholder, board seat, proxy battle, annual shareholder, proxy fight, shareholder vote, member board, shareholder group, institutional shareholder, board vote, bylaws, proxy material, governance, board meet, board member*
	Takeovers	*poison pill, hostile takeover, takeover bid, hostile bid, ashare offer, hostile tender, ashare tender, higher offer, takeover battle, takeover attempt, takeover offer, ashare bid, takeover defense, poisonpill, unsolicited bid, tender offer, unsolicited offer, takeover proposal, bid war, ashare*
	Real estate	*real estate, reit, estate investment, realestate investment, estate market, estate, estate developer, estate development, commercial real, realty trust, realty, investment trust, limit partnership, realestate company, commercial property, realestate, commercial realestate, realestate market, real, property*
	Convertible/preferred	*common share, preferred share, convertible preferred, common stock, class common, convertible debenture, company common, national overthecounter, convertible subordinate, preferred, preferred stock, overthecounter trade, class stock, stock outstanding, principal amount, corporate purpose, redeemable, share outstanding, convertible, newly issue*
Asset Managers/I-Banks	Mutual funds	*compound yield, simple yield, money fund, taxfree fund, assume reinvestment, taxable fund, stock fund, taxable money, fund company, taxexempt fund, moneymarket fund, closedend, company institute, noload, morningstar, fund rose, mutual fund, vanguard, fund industry, fund fell*
	Accounting	*andersen, account firm, kpmg, arthur andersen, fasb, deloitte, ernst young, audit firm, ernst, skadden, peat marwick, deloitte touche, skadden arps, meagher flom, arps, standard board, marwick, arps slate, slate meagher, meagher*
	Investment banking	*csfb, sachs group, goldman, suisse group, sachs, goldman sachs, credit suisse, suisse, ubs, morgan, dimon, group credit, stanley, morgan stanley, morgan chase, ubs warburg, swiss bank, investmentbanking, stanley dean, mack*
	Acquired investment banks	*salomon smith, smith barney, barney, merrill, bear stearns, stearns, merrill lynch, lynch, stearns cos, painewebber group, salomon, lufkin, donaldson lufkin, jenrette, donaldson, shearson, lufkin jenrette, brokerage unit, brokerage business, kidder*
	Private equity/hedge funds	*hedge fund, privateequity, privateequity firm, hedgefund, blackstone, carlyle, privateequity fund, hedgefund firm, hedgefund manager, blackstone group, tpg, management llc, bain capital, carlyle group, private equity, sovereignwealth fund, sovereignwealth, capital llc, soros, partner llc*
	NASD	*nasd, customer account, security transaction, member firm, barred, disciplinary action, admit, disciplinary, misappropriated, deny, fine, security dealer, restitution, misrepresentations, authorization, consent, suspend, quotation, supervisory, unauthorized*
Banks	Savings & loans	*fdic, thrift, loan association, federal deposit, federal save, deposit insurance, bank board, save loan, thrift industry, loan insurance, thrift supervision, rtc, save association, insurance fund, fail bank, federal insure, bank failure, save bank, loan bank, depositors*
	Nonperforming loans	*wells fargo, fargo, loanloss provision, bruyette, bankamerica, keefe bruyette, bruyette wood, nonperforming asset, loanloss, noninterest, interest margin, nonperforming, nonperforming loan, nationsbank, interstate bancorp, loanloss reserve, suntrust, pnc, loan growth, security pacific*
	Credit ratings	*debt rate, fitch rate, rate firm, rate service, creditwatch, rate agency, creditrating, taxes depreciation, junk status, notches, creditrating agency, senior debt, interest taxes, rate downgrade, operate cash, ebitda, negative implication, poor rate, rate company, fitch*

Continued on next page

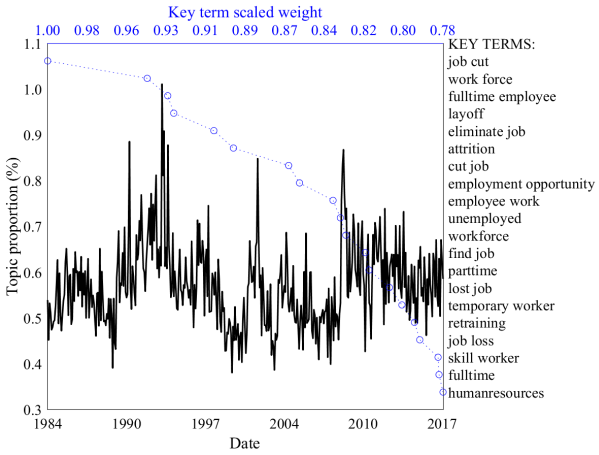
Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Financial crisis	*stress test, relief program, large financial, trouble asset, financial system, creditdefault swap, financial regulator, creditdefault, financial firm, financial crisis, credit crisis, secretary henry, financial regulation, timothy geithner, geithner, government bailout, henry paulson, financial instrument, bank balance, basel*
	Bank loans	*syndicate loan, revolving credit, credit line, libor, credit agreement, loan agreement, revolving, bank loan, creditor bank, rescheduling, credit facility, principal payment, shortterm loan, repayment, loan program, corporate loan, bank debt, interbank offer, offer rate, london interbank*
	Mortgages	*fixedrate mortgage, home price, home buyer, adjustablerate mortgage, mortgage rate, countrywide, federal house, mortgage company, mortgage lender, mortgage payment, mortgage banker, foreclosure, fannie, freddie, realtors, countrywide financial, mortgage loan, home mortgage, adjustablerate, house market*

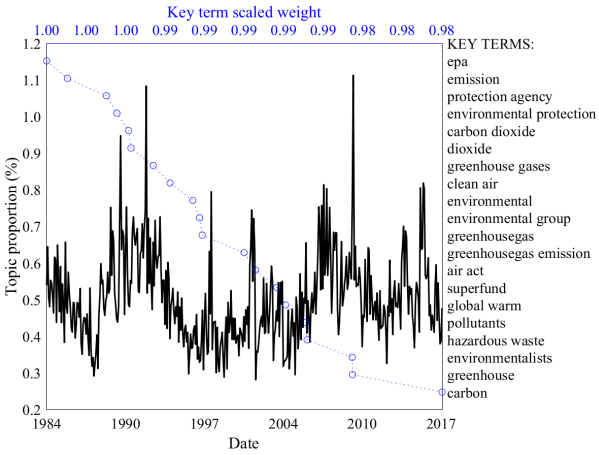
Figure 12: Topic Attention Proportions



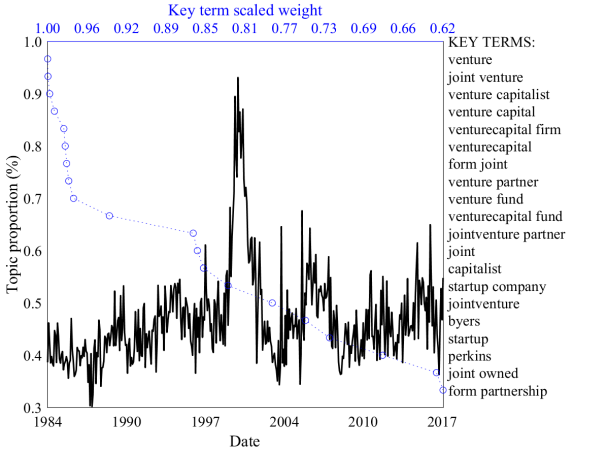
Job cuts



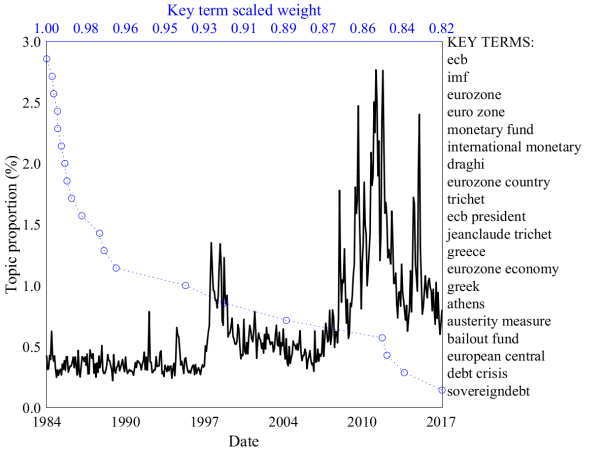
Environment



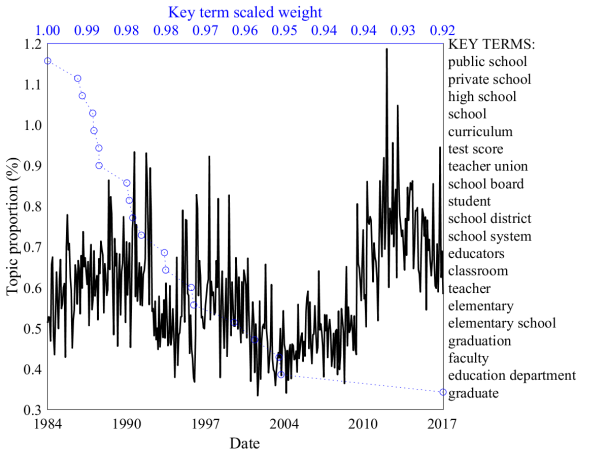
Venture capital



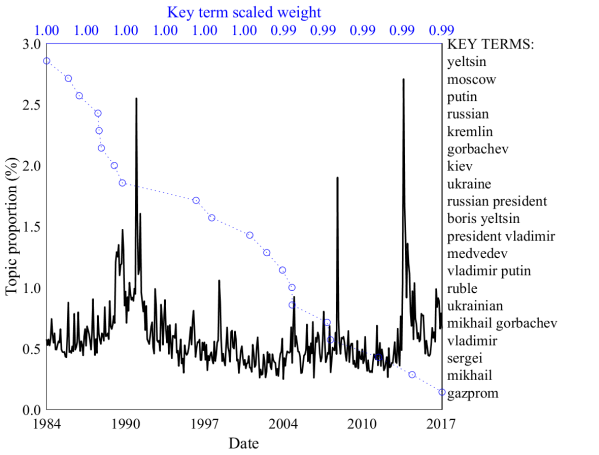
European sovereign debt



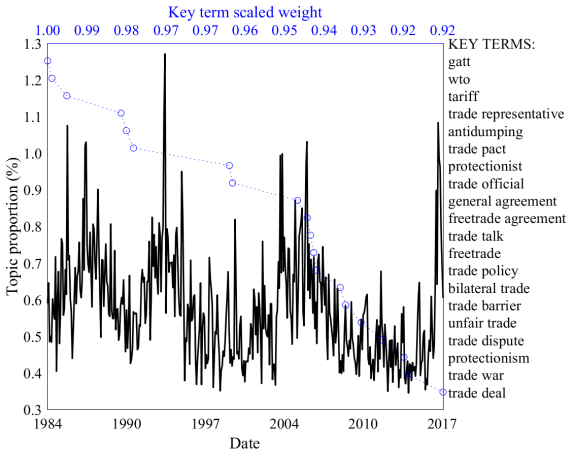
Schools



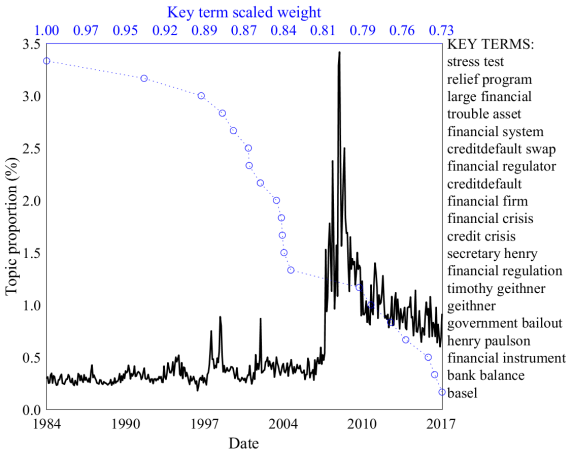
Russia



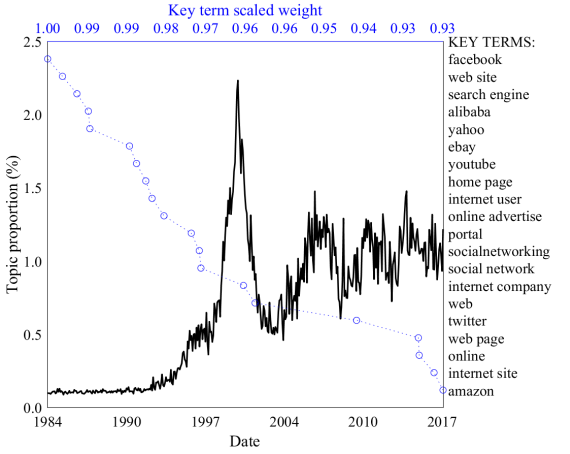
Trade agreements



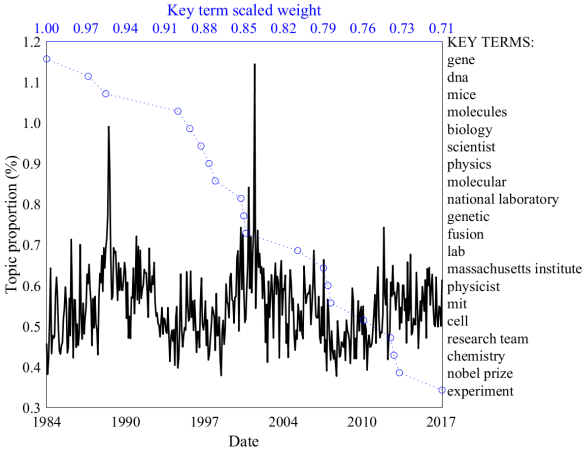
Financial crisis



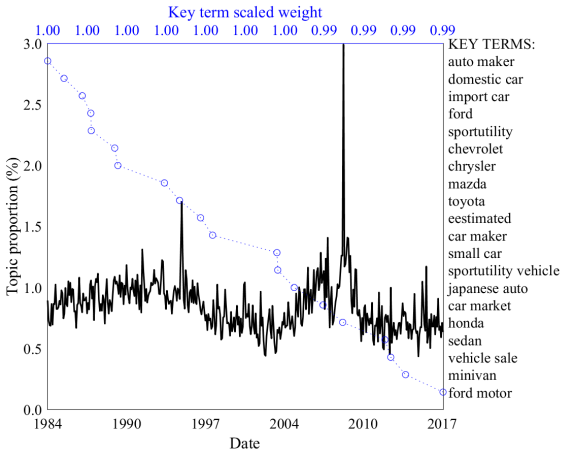
Internet



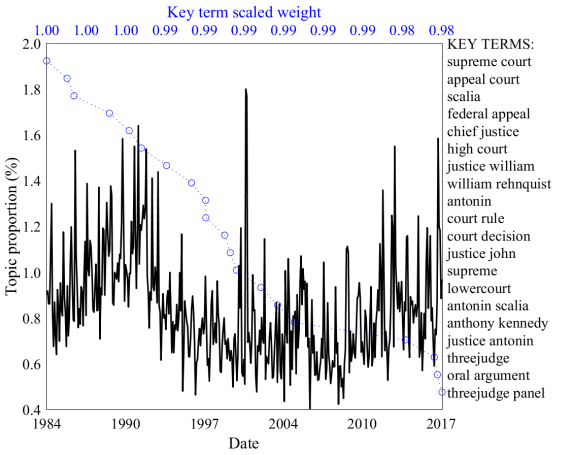
Biology/chemistry/physics

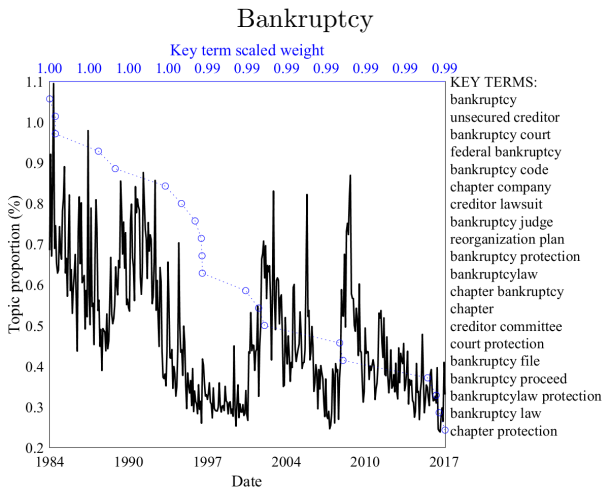


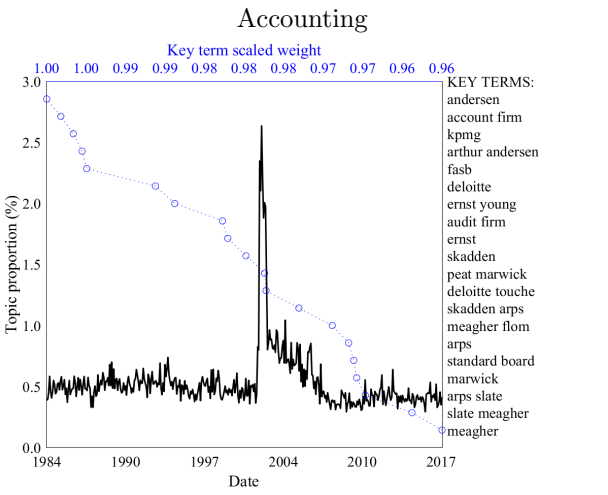
Automotive

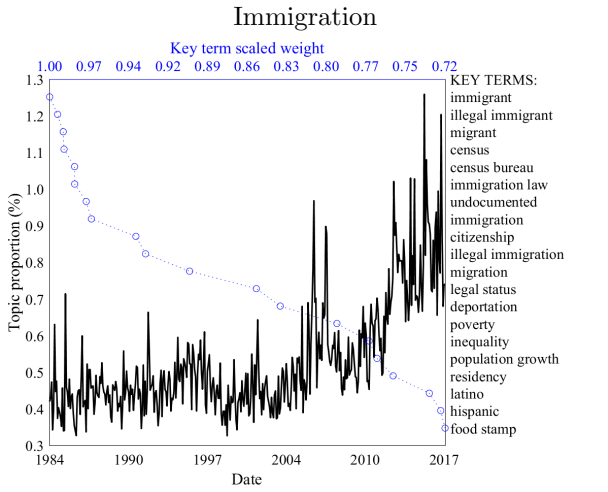


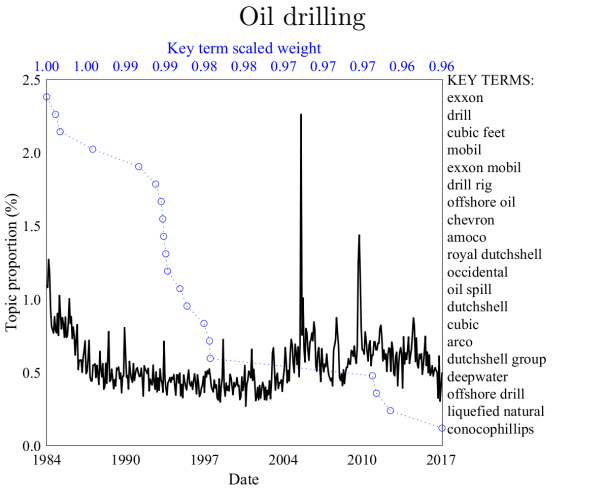
Courts

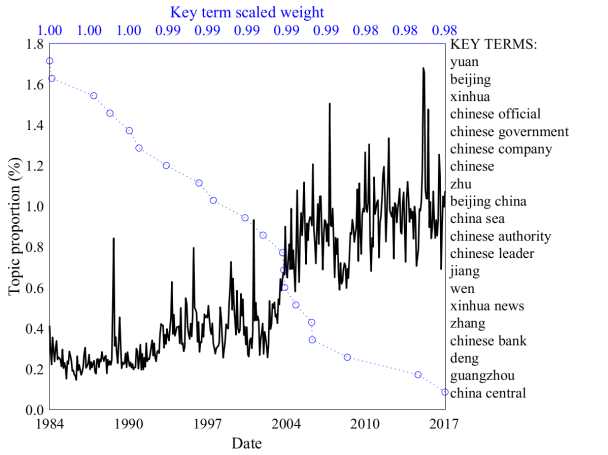


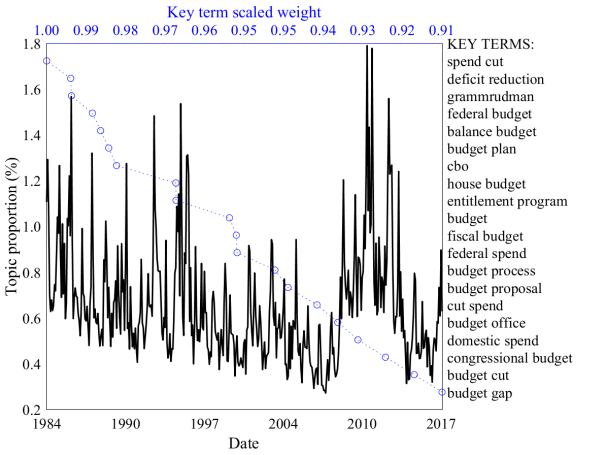

Bankruptcy
Key term scaled weight
1.00 1.00 1.00 1.00 0.99 0.99 0.99 0.99 0.99 0.99 0.99
KEY TERMS:
bankruptcy
unsecured creditor
bankruptcy court
federal bankruptcy
bankruptcy code
chapter company
creditor lawsuit
bankruptcy judge
reorganization plan
bankruptcy protection
bankruptcylaw
chapter bankruptcy
chapter
creditor committee
court protection
bankruptcy file
bankruptcy proceed
bankruptcylaw protection
bankruptcy law
chapter protection
Topic proportion (%)
Date


Accounting
Key term scaled weight
1.00 1.00 0.99 0.99 0.98 0.98 0.98 0.97 0.97 0.96 0.96
KEY TERMS:
andersen
account firm
kpmg
arthur andersen
fasb
deloitte
ernst young
audit firm
ernst
skadden
peat marwick
deloitte touche
skadden arps
meagher flom
arps
standard board
marwick
arps slate
slate meagher
meagher
Topic proportion (%)
Date


Immigration
Key term scaled weight
1.00 0.97 0.94 0.92 0.89 0.86 0.83 0.80 0.77 0.75 0.72
KEY TERMS:
immigrant
illegal immigrant
migrant
census
census bureau
immigration law
undocumented
immigration
citizenship
illegal immigration
migration
legal status
deportation
poverty
inequality
population growth
residency
latino
hispanic
food stamp
Topic proportion (%)
Date


Oil drilling
Key term scaled weight
1.00 1.00 0.99 0.99 0.98 0.98 0.97 0.97 0.97 0.96 0.96
KEY TERMS:
exxon
drill
cubic feet
mobil
exxon mobil
drill rig
offshore oil
chevron
amoco
royal dutchshell
occidental
oil spill
dutchshell
cubic
arco
dutchshell group
deepwater
offshore drill
liquefied natural
conocophillips
Topic proportion (%)
Date


China
Key term scaled weight
1.00 1.00 1.00 0.99 0.99 0.99 0.99 0.99 0.98 0.98 0.98
KEY TERMS:
yuan
beijing
xinhua
chinese official
chinese government
chinese company
chinese
zhu
beijing china
china sea
chinese authority
chinese leader
jiang
wen
xinhua news
zhang
chinese bank
deng
guangzhou
china central
Topic proportion (%)
Date


Government budgets
Key term scaled weight
1.00 0.99 0.98 0.97 0.96 0.95 0.95 0.94 0.93 0.92 0.91
KEY TERMS:
spend cut
deficit reduction
grammrudman
federal budget
balance budget
budget plan
cbo
house budget
entitlement program
budget
fiscal budget
federal spend
budget process
budget proposal
cut spend
budget office
domestic spend
congressional budget
budget cut
budget gap
Topic proportion (%)
Date

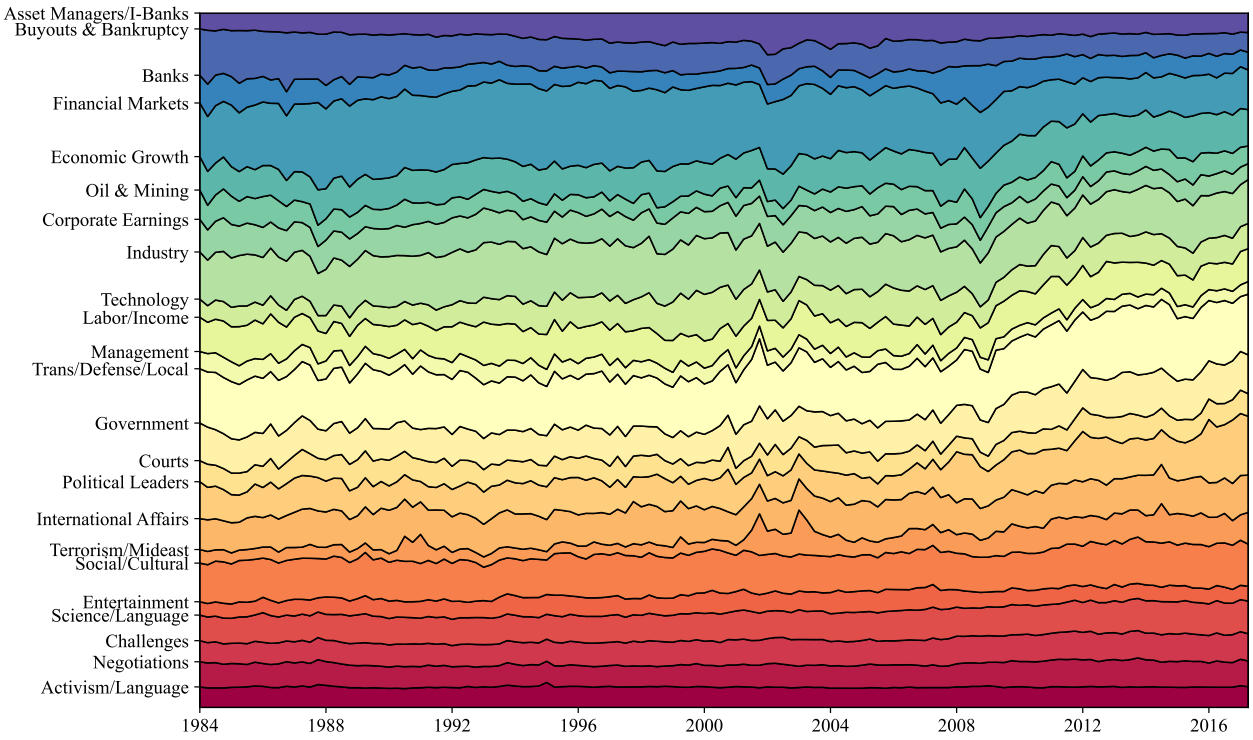
D Dominant Topics Over Time

With 180 topics, it is difficult to isolate the focus of attention at each point in time in a single line plot overlaying many time series. To overcome this, Figure 13 shows the top two topics receiving the most news attention each month. Dominant topics vary depending on exogenous factors (such as election cycles or natural disasters) and endogenous states of the economy (such as recessions and sovereign debt crises), with primary attention rotating between political, macroeconomic, industrial, and financial topics.

Figure 13: Dominant Topics By Month

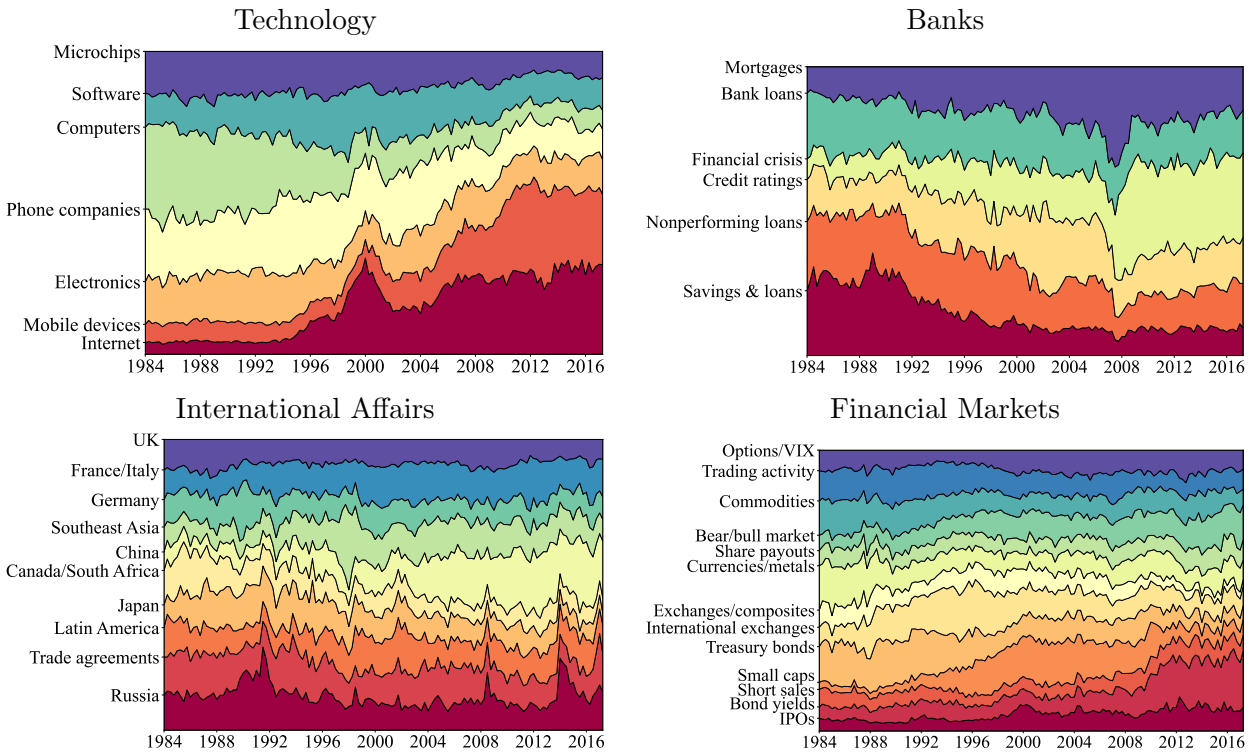
Note. The figure reports the topics ranked 1 and 2 each month in terms of highest *WSJ* news attention proportion.

Figure 14: News Attention Across Metatopics



Note. Stacked area plot of attention allocation to 23 metatopics based on the first-level clustering of baseline topics.

Figure 15: News Attention Within Metatopics



Note. Stacked area plot of attention allocation to baseline topics within four of the first-level metatopics.

E Meta Topics

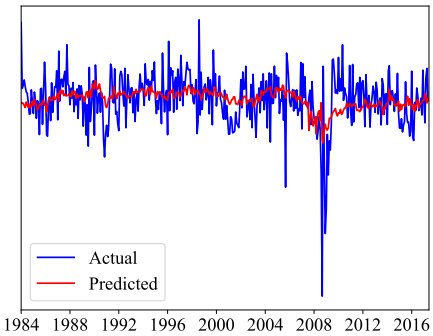
Figure 14 zooms out to show how attention behaves at the 23-metatopic level (the first level of aggregation on the right side of Figure 2). The fraction of news attention allocated to metatopics is fairly stable over time. Yet there is a high degree of attention churn within metatopic clusters. For example, while attention to "technology" is stable over time in Figure 14, its contents transition rather dramatically over time (upper left panel of Figure 15). It is dominated by mainframe and desktop computers at the start of the sample, is displaced by attention to software mid-sample, which is in turn displaced by mobile devices and the internet in the later part of the sample. Likewise, while attention to the "banks" metatopic is stable over time, the focus shifts from savings and loans in the 1980's to mortgages in the mid 2000's, and remains dominated by attention to financial crises even through the recent economic expansion. The "international affairs" metatopic includes steady attention to western Europe, but also shows episodic spikes in attention to Russia, a brief focus on Southeast Asia during its 1998 crisis, and the gradual emergence of China as a focal point by the end of the sample. The "financial markets" metatopic shifts focus from Treasuries and international markets in the first half of the sample, to small caps during the dotcom era and leading up to the financial crisis, after which bond yields become a focal point.

Metatopics provide additional dimension reduction, beyond that provided by LDA. It is therefore possible that metatopics are better able to explain variation in macroeconomic fundamentals or stock market prices. To investigate this possibility, in Table 8 we repeat the analysis of Table 1 but using the 23 metatopics. We find that R-squares are roughly halved when moving from the 180 topics to the 23 metatopics. This comparison would not be interesting in OLS regressions. But because in both cases we select a Lasso model with exactly five regressors, the comparison is valid and suggests that the more granular 180 topic level of aggregation provides a superior summary of the state of the economy.

Table 8: Reconstructing Macroeconomic Time Series Using Metatopics

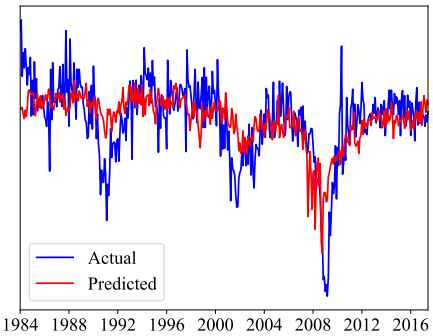
Industrial Production Growth

Topic	Coeff.	p-val.
International/Commodities	-0.23	0.04
Culture	0.14	0.55
Law/Courts	-0.11	0.95
International Affairs	-0.09	0.02
Entrepreneurship	0.07	0.77
R^2	0.12	



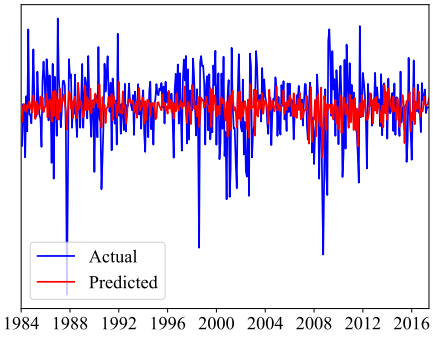
Employment Growth

Topic	Coeff.	p-val.
International/Commodities	-0.39	0.00
Technology	0.30	0.20
Banks	-0.25	0.00
Entrepreneurship	0.11	0.50
National Events	0.08	0.93
R^2	0.37	



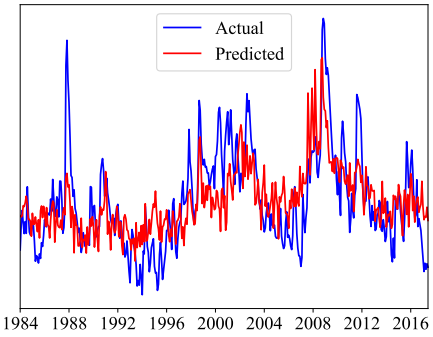
Market Returns

Topic	Coeff.	p-val.
Economic Growth	-0.20	0.00
International Affairs	-0.20	0.14
International/Commodities	-0.10	0.22
Law/Courts	0.10	0.35
Culture	0.09	0.85
R^2	0.11	



Market Volatility

Topic	Coeff.	p-val.
Banks	0.41	0.00
Oil/Industry	-0.30	0.00
Entrepreneurship	-0.28	0.00
International/Commodities	0.25	0.00
Economic Growth	0.08	0.06
R^2	0.44	

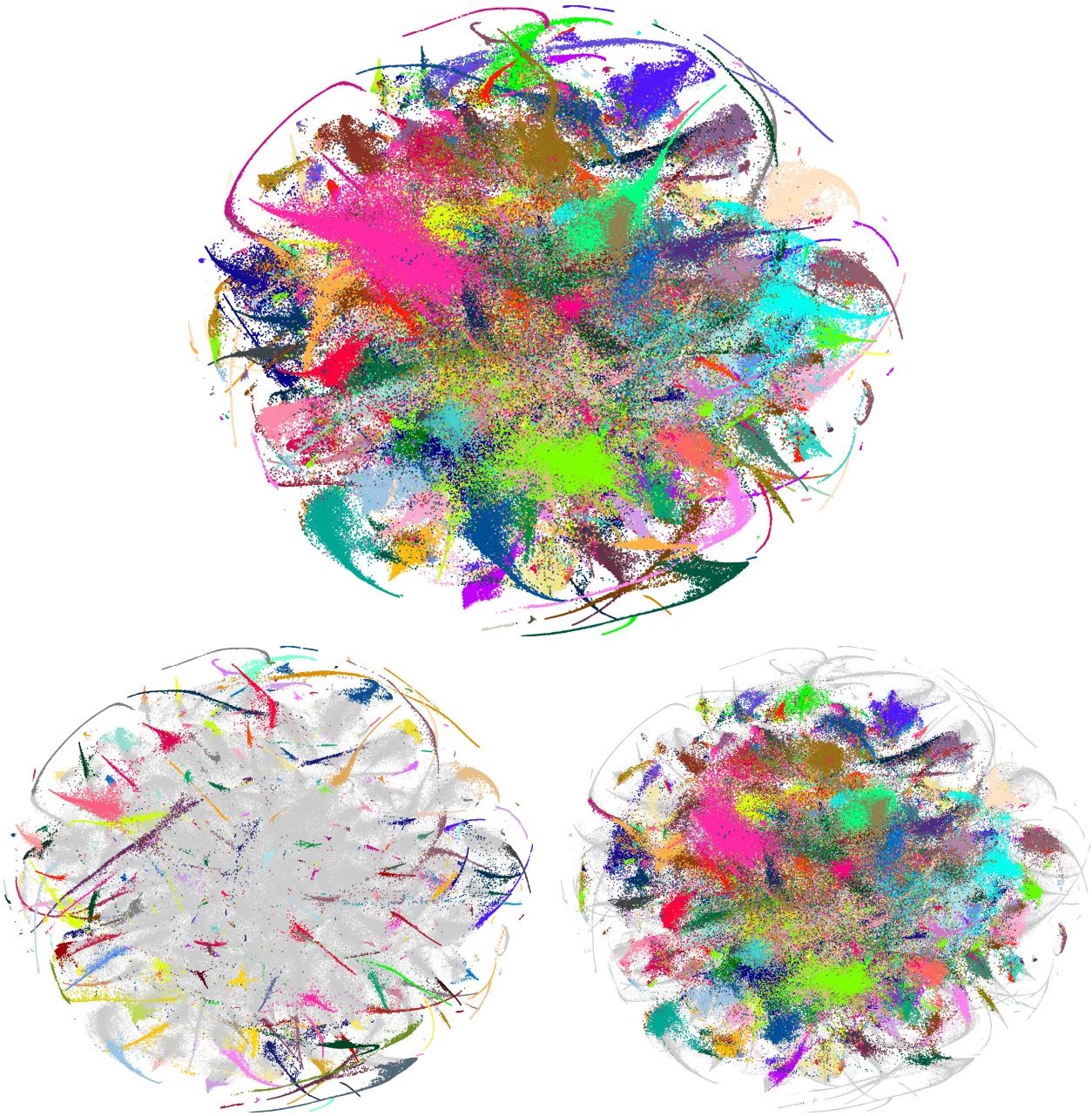


Note. Five-regressor active-set regression estimates based on lasso selection with p-values adjusted for post-selection inference. Here, instead of topic attentions, we use the 23 higher-level metatopics as explanatory variables.

F Article-level News Attention

Time series plots describe the attention paid to each topic by the *WSJ* as a whole. In this section, we explore patterns in topic attention at the article level. Article-level attention is a 180-dimension

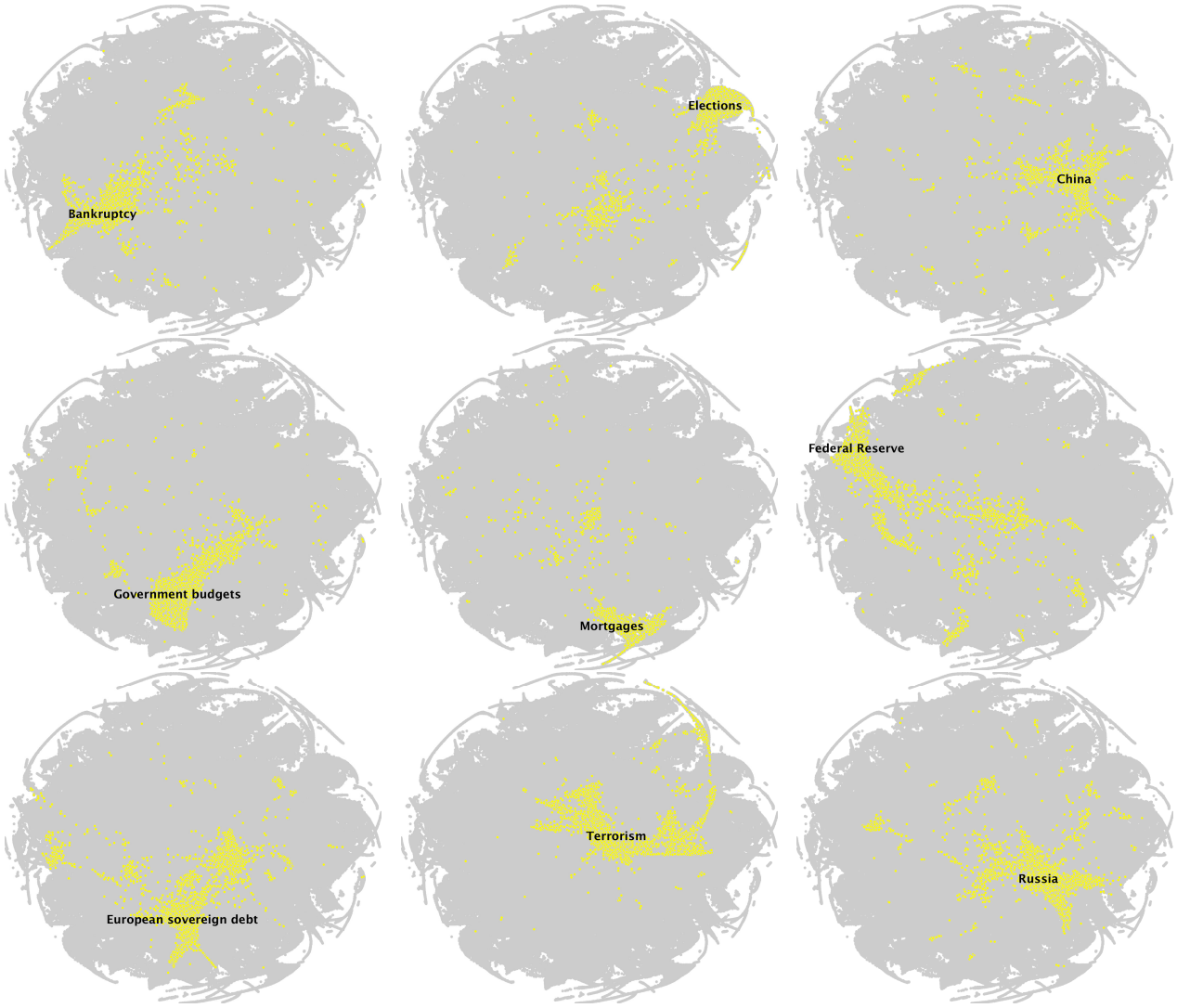
Figure 16: Article-level Nearest Neighbor Embedding



Note. The top shows the two-dimensional embedding of *WSJ* articles, where each topic is assigned a color and articles are colored in correspondence with their dominant topic. The bottom left panel highlights only "pure" (those with maximum topic proportion exceeding 33%), while the bottom right shows "mixed" articles (maximum topic proportion below 25%).

vector listing the proportion attention that the article pays to each topic. While this is an enormous reduction in dimensionality relative to the term count representation, the fits amount to a large dataset in their own right: a matrix of roughly 800,000 articles by 180 topics. In order to visualize and

Figure 17: Article-level Nearest Neighbor Embedding (Topic Breakout)

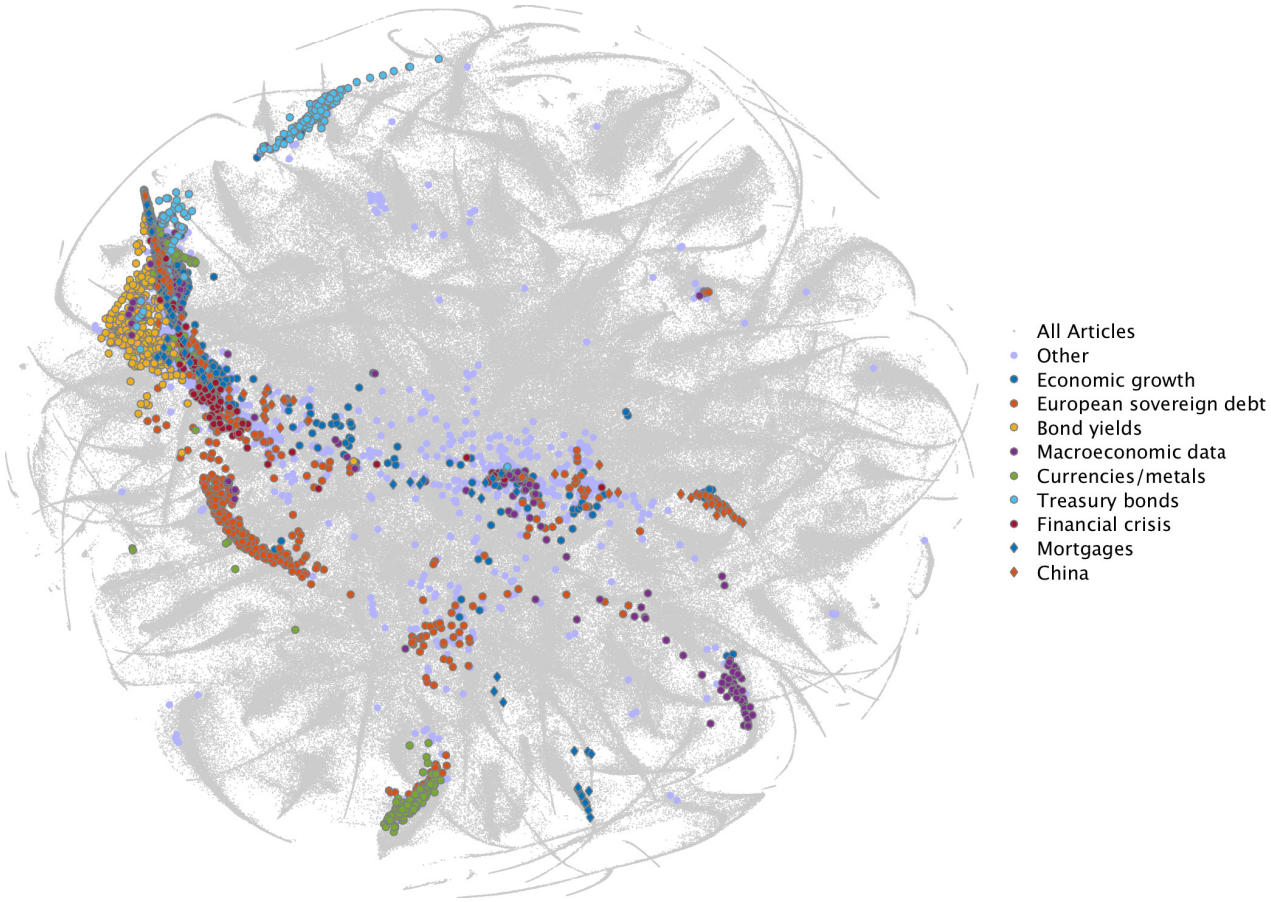


Note. Two-dimensional embedding of *WSJ* articles broken out by individual topics.

interpret article-level fits, we project the fitted proportions into two dimensions using t-distributed stochastic nearest neighbor embedding (t-SNE, Maaten and Hinton, 2008). Intuitively, the t-SNE plot assigns a two-dimensional coordinate to each article so that the nearness of articles in the plane best represents the nearness of the article fits in the full 180-dimensional space.

The top panel of Figure 16 shows the two-dimensional embedding of model fits for every *WSJ* article in our data set, and Figure 17 breaks down the complexity of Figure 16 by separately plotting articles based on their leading topic for an illustrative subset of nine topics. In Figure 16, each topic is assigned a color, and each article (a single point) is colored in correspondence with the topic having the highest weight in that article. Naturally, articles cluster spatially based on their dominant topics. Graphically, an article (which the model views as a mixtures of topics) is an intersection of topic

Figure 18: Articles Featuring the Federal Reserve

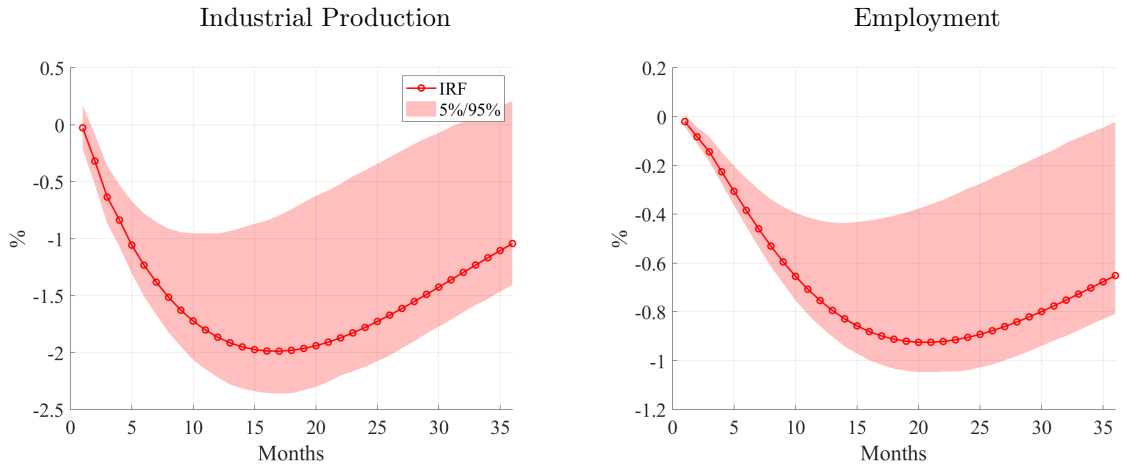


Note. Detail of two-dimensional embedding for article whose dominant topic is "Federal Reserve." Articles within this set are colored according to their second largest topic proportion.

clouds. Articles that mix many topics must intersect with many clouds, so they are centrally located in the plot. Similarly, clouds of topics that tend to co-occur with other topics (such as "Federal Reserve") need to be dispersed throughout the t-SNE, while those that mix with only a few other topics are more concentrated (such as "mortgages" and "terrorism"). This can also be seen at the article level via the two bottom panels of Figure 16, which separates articles by their purity. "Pure" articles are shown in color on the left, and are defined as those with a maximum topic proportions exceeding 33%. These articles (18% by count) are dominated by a single topic. The right shows the set of "mixed" articles (maximum topic proportion below 25%). Pure articles, which have little need for proximity to other topics, and tend to live on the periphery of the embedding. Mixed articles require topic crossing and are thus centrally located.

Figure 18 provides more detail for one specific example. It documents the other topical content of articles whose leading topic is "Federal Reserve." It splits articles within the Fed topic based on their second largest topic allocation. Figure 18 shows that articles about the Fed also tend to relate to the

Figure 19: Impulse Responses With Confidence Intervals Adjusted for Lasso Selection



Note. Responses of output and employment to a "recession" attention impulse in various VAR specifications, 1985–2017. Confidence intervals are adjusted to reflect lasso model selection.

"economic growth," "European sovereign debt," "bond yields," "macroeconomic data," "Treasury bonds," "financial crisis," "mortgages," and "China" topics, with each of these secondary topics tightly spatially clustered within the Fed topic.

Like the hierarchical metatopic analysis, the t-SNE plots as a whole demonstrate coherence of the estimated topic model. They show that an economic event—captured verbally as a news article—is summarized via our model as the constellation of topics that constitute that event. Likewise, the point-in-time distribution of aggregate topic proportions summarizes the overall state of the economy. These model estimates illustrate that the economy is fundamentally complex. An attempt at parsimoniously representing the verbal state of the economy, while a drastic simplification of the full *WSJ* narrative, nonetheless requires an elaborate specification to capture its common themes. But our model estimation results, and our further empirical analysis below, also demonstrate that this complexity can be effectively harnessed for economic analysis.

G Post-selection Inference For Impulse Responses

When developing our text-augmented VAR, we use group-lasso to select news topics for model inclusion. This selection step affects inferences of statistical significance for our impulse response estimates. In this section we describe a post-selection inference method for constructing impulse response confidence intervals that account for group-lasso selection.

Our method to construct appropriate confidence intervals builds on the bootstrap procedure of Kilian (1999) that we use in Figure 4. First, we construct VAR residuals from our baseline VAR specification (which includes "recession" attention as the only news-based variable). Next, we

resample residuals to produce bootstrap data samples. In each bootstrap sample, we select topics for inclusion in the VAR based on the cross-validated value of $\lambda = 0.054$ from our analysis in Figure 9. All topics that receive non-zero group-lasso estimates based on this value of λ are included in the VAR model, which we re-estimate with OLS. From this we produce the "recession" attention impulse response function (if "recession" is not selected in the VAR for a given bootstrap sample, the impulse responses are set to zero). Confidence intervals are then defined as the point-wise 5% to 95% range of bootstrapped impulse response functions.

Figure 19 reports the post-selection adjusted confidence intervals for our baseline VAR specification. We note that the confidence intervals are somewhat wider than those in Figure 4, but that the quantitative magnitudes are similar and our conclusions regarding statistical significance of the impulse responses are unaffected by the post-selection adjustment.

H Macroeconomic VAR Impulse Responses Robustness

For brevity, in the paper we reported impulse responses visually in Section 5. Here, we supplement Figure 4 which only reports confidence intervals for a single alternative specification with confidence intervals for all specifications every five years.

Table 9 reports point estimates and confidence intervals for industrial production and employment. The table shows for all alternative specifications we can reject a null effect for a considerable period of time following a "recession" attention impulse. Baseline reports our main specification results. oLDA reports the comparable results for an instance of our oLDA specification and EPU reports the comparable results using EPU in place of our recession topic. News last places our recession topic last in the VAR specification, while news 2nd places it second. Incl. EPU, Mich, and VIX, each represent a specification that includes the corresponding variable. Incl. Trend includes a trend term and 6 Lags uses 6 lags instead of 3.

Table 9: Recession Impulse Response

Specification	1	5	10	15	20	25	30	35
Panel A: Industrial Production								
Baseline	-0.03 [-0.16 0.12]	-1.03 [-1.27 -0.69]	-1.69 [-2.02 -1.03]	-1.93 [-2.30 -0.99]	-1.90 [-2.29 -0.77]	-1.69 [-2.09 -0.45]	-1.40 [-1.77 -0.09]	-1.08 [-1.50 0.21]
oLDA	-0.01 [-0.17 0.15]	-0.84 [-1.16 -0.45]	-1.47 [-1.91 -0.74]	-1.77 [-2.23 -0.77]	-1.82 [-2.27 -0.61]	-1.71 [-2.11 -0.39]	-1.51 [-1.84 -0.14]	-1.28 [-1.56 0.09]
EPU	-0.11 [-0.25 0.03]	-0.73 [-1.04 -0.33]	-0.96 [-1.37 -0.28]	-0.88 [-1.33 -0.06]	-0.68 [-1.15 0.19]	-0.45 [-0.95 0.41]	-0.22 [-0.76 0.60]	-0.03 [-0.59 0.72]
News Last	0.00 [0.00 0.00]	-0.67 [-0.86 -0.42]	-1.08 [-1.38 -0.57]	-1.22 [-1.60 -0.50]	-1.18 [-1.61 -0.34]	-1.01 [-1.46 -0.11]	-0.78 [-1.23 0.13]	-0.53 [-0.98 0.32]
News 2nd	-0.10 [-0.23 0.03]	-0.84 [-1.06 -0.53]	-1.26 [-1.59 -0.69]	-1.37 [-1.79 -0.57]	-1.27 [-1.75 -0.32]	-1.04 [-1.54 -0.00]	-0.73 [-1.25 0.30]	-0.42 [-0.96 0.54]
Incl. EPU	-0.06 [-0.19 0.09]	-1.08 [-1.32 -0.73]	-1.68 [-2.00 -1.03]	-1.90 [-2.26 -0.96]	-1.87 [-2.24 -0.72]	-1.68 [-2.06 -0.44]	-1.41 [-1.79 -0.14]	-1.11 [-1.49 0.14]
Incl. Mich	-0.02 [-0.16 0.11]	-0.99 [-1.23 -0.64]	-1.55 [-1.85 -0.91]	-1.70 [-2.03 -0.83]	-1.61 [-1.96 -0.59]	-1.40 [-1.75 -0.31]	-1.15 [-1.49 -0.05]	-0.91 [-1.27 0.16]
Incl. VIX	0.01 [-0.13 0.16]	-1.08 [-1.32 -0.67]	-1.75 [-2.10 -0.97]	-1.99 [-2.38 -0.88]	-1.92 [-2.36 -0.58]	-1.67 [-2.10 -0.24]	-1.32 [-1.73 0.10]	-0.97 [-1.39 0.38]
Incl. Trend	-0.02 [-0.15 0.11]	-1.02 [-1.25 -0.68]	-1.66 [-1.93 -0.98]	-1.87 [-2.16 -0.87]	-1.81 [-2.10 -0.59]	-1.59 [-1.87 -0.26]	-1.31 [-1.62 0.04]	-1.04 [-1.38 0.29]
6 Lags	0.01 [-0.12 0.13]	-0.94 [-1.23 -0.55]	-1.51 [-1.92 -0.79]	-1.61 [-2.07 -0.63]	-1.47 [-1.98 -0.39]	-1.25 [-1.78 -0.16]	-0.99 [-1.51 0.06]	-0.74 [-1.26 0.27]
Panel B: Employment								
	1	5	10	15	20	25	30	35
Baseline	-0.02 [-0.04 0.00]	-0.30 [-0.36 -0.21]	-0.64 [-0.74 -0.42]	-0.84 [-0.97 -0.48]	-0.91 [-1.05 -0.43]	-0.87 [-1.03 -0.33]	-0.78 [-0.94 -0.19]	-0.66 [-0.83 -0.06]
oLDA	-0.02 [-0.05 0.00]	-0.27 [-0.34 -0.16]	-0.58 [-0.71 -0.32]	-0.79 [-0.94 -0.37]	-0.89 [-1.05 -0.36]	-0.89 [-1.04 -0.28]	-0.84 [-0.96 -0.18]	-0.75 [-0.85 -0.09]
EPU	-0.03 [-0.05 -0.00]	-0.17 [-0.26 -0.07]	-0.31 [-0.45 -0.08]	-0.33 [-0.52 -0.03]	-0.29 [-0.49 0.06]	-0.21 [-0.44 0.16]	-0.12 [-0.37 0.25]	-0.03 [-0.30 0.31]
News Last	0.00 [0.00 0.00]	-0.15 [-0.19 -0.09]	-0.35 [-0.45 -0.19]	-0.47 [-0.61 -0.21]	-0.50 [-0.68 -0.17]	-0.47 [-0.67 -0.09]	-0.40 [-0.60 -0.00]	-0.30 [-0.51 0.08]
News 2nd	-0.02 [-0.04 0.00]	-0.22 [-0.28 -0.13]	-0.44 [-0.55 -0.25]	-0.56 [-0.72 -0.25]	-0.58 [-0.77 -0.19]	-0.52 [-0.74 -0.08]	-0.42 [-0.65 0.05]	-0.29 [-0.55 0.16]
Incl. EPU	-0.02 [-0.04 0.00]	-0.30 [-0.35 -0.20]	-0.63 [-0.72 -0.41]	-0.82 [-0.94 -0.47]	-0.89 [-1.02 -0.42]	-0.87 [-1.00 -0.31]	-0.79 [-0.93 -0.19]	-0.67 [-0.82 -0.07]
Incl. Mich	-0.02 [-0.04 0.01]	-0.27 [-0.33 -0.18]	-0.57 [-0.66 -0.36]	-0.71 [-0.83 -0.39]	-0.74 [-0.87 -0.33]	-0.69 [-0.84 -0.22]	-0.60 [-0.76 -0.10]	-0.51 [-0.66 -0.01]
Incl. VIX	-0.03 [-0.05 -0.00]	-0.31 [-0.37 -0.20]	-0.68 [-0.76 -0.41]	-0.88 [-0.99 -0.46]	-0.94 [-1.07 -0.38]	-0.89 [-1.02 -0.25]	-0.77 [-0.92 -0.10]	-0.64 [-0.78 0.03]
Incl. Trend	-0.02 [-0.04 0.00]	-0.30 [-0.35 -0.20]	-0.64 [-0.72 -0.40]	-0.84 [-0.94 -0.45]	-0.91 [-1.00 -0.39]	-0.88 [-0.97 -0.28]	-0.79 [-0.88 -0.15]	-0.68 [-0.77 -0.02]
6 Lags	-0.01 [-0.03 0.01]	-0.25 [-0.30 -0.16]	-0.53 [-0.63 -0.33]	-0.70 [-0.84 -0.38]	-0.77 [-0.94 -0.34]	-0.75 [-0.95 -0.26]	-0.67 [-0.89 -0.14]	-0.56 [-0.78 -0.03]

Note. Responses of output to a "recession" attention impulse in various VAR specifications, 1985–2017. Each column reports a given horizon for the impulse response function. The first number reports the point estimate, while the numbers in brackets report the lower and upper bounds of the 90% confidence interval.