The Structure of Economic News*

Leland Bybee
Yale University

Bryan Kelly
Yale University, AQR Capital Management, and NBER

Asaf Manela
Washington University in St. Louis

Dacheng Xiu
Booth School of Business
University of Chicago

September 22, 2019

Abstract

We propose an approach to measuring the state of the economy via textual analysis of business news. From the full text content of 800,000 *Wall Street Journal* articles for 1984–2017, we estimate a topic model that summarizes business news as easily interpretable topical themes and quantifies the proportion of news attention allocated to each theme at each point in time. We then use our news attention estimates as inputs into statistical models of numerical economic time series. We demonstrate that these text-based inputs accurately track a wide range of economic activity measures and that they have incremental forecasting power for macroeconomic outcomes, above and beyond standard numerical predictors. Finally, we use our model to retrieve the news-based narratives that underly "shocks" in numerical economic data.

Key words: Textual analysis, macroeconomic news, attention, Wall Street Journal, volatility, VAR, machine learning

*We benefitted from discussions with Rob Engle, Will Goetzmann, Toby Moskowitz, Lasse Pedersen, Bob Shiller, and seminar and conference participants at AQR, Chicago, NYU, and Yale.

Disclaimer: The views and opinions expressed are those of the authors and do not necessarily reflect the views of AQR Capital Management, its affiliates, or its employees; do not constitute an offer, solicitation of an offer, or any advice or recommendation, to purchase any securities or other financial instruments, and may not be construed as such.

1 Introduction

One of the primary functions 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. For example, the BEA maintains detailed national accounts for income and spending at the aggregate and industry level, while the BLS specializes in measuring employment, wages, and general price levels. 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. Together, these comprise an overlapping ecosystem of numerical business cycle indicators. Ascertaining the nature and evolution of the state of the economy from these data is a notoriously difficult task.

In this paper, we offer a new 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 and impact, and inferring their causes. The information disseminated by news media is an equilibrium outcome determined by the confluence of consumer preferences, news production technologies, and the competitive landscape of the news industry (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 text is a one-stop shop for simultaneously understanding many facets of the state of the economy and how they interact with each other. It uses the richness of narrative, derived from human understanding and interpretation of complex contexts.

Despite the fact that news media reflects information that consumers rely on to make allocation and consumption choices 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.[1] They have two essential elements. Just as principal components 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 that are most likely to co-occur in the same article. Those clusters are optimized so that relatively few clusters (many fewer than the number of distinct terms in the data set) preserve as much of the meaning in original

[1] Topic models have only recently begun to be explored in empirical economics research. The earliest example we are aware of is Hansen et al. (2017).

corpus as possible, by best explaining the variation in term usage across articles.

A topic model's second main element estimates for each article, the proportion of text dedicated to each topic. 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 estimation tasks of 1) summarizing dominant themes throughout the history of the *WSJ*, and 2) tracking how media attention to news topics evolves. From this, we provide 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 more topics use more parameters than are warranted by the gain in 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. A leading hypothesis for this fact (e.g., Engle et al., 1990) is that volatility is driven by news arrivals, and that news itself arrives in clusters. The persistence in attention paid to a particular news topic supports at least part of this hypothesis. For example, the prevalence of news associated with the "oil drilling" {key terms: *exxon mobil, cubic feet, drill rig, offshore oil*} and "oil market" {key terms: *opec, nonopec, oil minister, oil demand*} topics in the *WSJ* closely coincide with the volatility of crude oil prices.

We find that news is a combination of recurrent, seasonal, and episodic topics. Recurrent topics are those garnering media attention consistently throughout the sample. For example, the "Federal Reserve" topic, whose key terms include {*greenspan, yellen, federalfunds rate, fomc*}, is regularly active throughout the sample. As is the "health insurance" topic {key terms: *hmo, health plan, health coverage, blue cross*}, which becomes highly prevalent during debate of the Clinton Health Plan proposal (peaking around President Clinton's speech to Congress in September 1993), the proposal and passage of Obamacare in 2008–2010, and the Obamacare repeal debate around the 2016 presidential election. Seasonal topics include "presidential elections" {key terms: *obama, romney, dukakis, campaign finance*}, whose prevalence has a clockwork seasonal pattern peaking every four years, and "earnings forecasts" {key terms: *analyst poll, earn forecast, earningspershare, earn expectation*} which spikes ahead of each earnings announcement season. Episodic topics are dormant for much of the sample, then spike around particular events. For example, the "terrorism" topic {key terms: *taliban, queda, suicide bomber, osama*} is a prototypical regime shift: It has a prevalence near zero for the first half of our sample, spikes dramatically at 9/11, and remains high henceforth. Similarly, the "natural disasters" topic {key terms: *katrina, quake, tsunami, hurricane*} has a very low proportion 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 the effects of climate change (Engle et al., 2019). These examples 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 with which the co-occur in a given article.

We draw on a number of additional tools from the machine learning literature that help us further understand the rich structure of text. One is a machine learning graphing tool known as stochastic nearest neighbor embedding (t-SNE, Maaten and Hinton, 2008) which helps us visualize *WSJ* news in two-dimensional space. In a topic model, a news article is a mixture of themes, which can be interpreted as a news-based snapshot of the state of the economy. The t-SNE plots help digest the high-dimensionality and complexity news events, illustrating their intuitive underlying structure as recurring constellations of topics. Another machine learning tool that we use is hierarchical agglomerative clustering (Murtagh and Legendre, 2014), which we use to group topics together based on their thematic similarities. We recursively combine our 180 base topics into a tree structure of gradually broadening metatopics. In doing so, we establish a complete taxonomy of business news. We find that the proportion of attention paid to high-level metatopics (like "markets," "government," and "technology") is roughly constant over time. Instead, it is at finer resolutions where we find often dramatic time variation in attention. That is, news churn occurs as rapid shifts in attention to specific sub-topics within broad and stable metatopics.

Our main contribution is to show that the text of business news summarizes wide ranging facets of the state of the economy. Similar to the oil volatility example, we report a variety of validation checks showing that economic topics identified from *WSJ* text coincide with closely 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 thematically related news topics provide a close match to the paths of the numerical macro series.

One example of economic activity that we study is financing volume. We find that the volume of leveraged buyout (LBO) transactions is most significantly associated with the "takeovers" topic {key terms: *poison pill, hostile takeover, share tender, higher offer*} and "control stakes" 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. But they outline how discussion of prevailing economic topics helps understand realized macroeconomic fluctuations.

Contemporaneous correlation between news narratives and macroeconomic fluctuations is a useful validation that text can be productively converted into a summary of the state of the economy. This, however, has little to say about whether news text contains distinct or incremental information versus standard numerical indicators. To investigate this directly, we study news attention within an

otherwise standard macroeconomic vector autoregression (VAR). We find that attention to recession news has economically large and highly significant predictive power for future output and employment. For example, one conservative variant of our analysis shows that a standard deviation impulse in recession attention associates with a decline in industrial production of 1.7% after twelve months, after controlling for a variety of common VAR components such as stock market returns, interest rates, and measures of economic uncertainty. That is, news coverage of recession risk captures useful information about future economic outcomes above and beyond commonly used indicators. Taken together, our results highlight fascinating possibilities for using textual data in problems such as macroeconomic forecasting, which tend to rely exclusively on numerical data sources, and modeling the role of information transmission and media in the macroeconomy.[2]

Modeling economic activity in terms of news-based topics offers economists a new set of tools for understanding *shocks*. For example, in a standard macroeconomic VAR, shock interpretation often boils down to a choice of rotation for the error covariance matrix. But with news-based models, fluctuations can be directly mapped to textual narratives of economic conditions. We describe an approach to *narrative retrieval* in a macroeconomic forecasting application. In a first step, we forecast industrial production growth (after controlling for past growth) using news attention and identify the most predictive topics. To understand a specific shock to expected growth within the model, we examine which of the influential topics experienced a coincident shift in news attention. Next, the estimated topic model traces these specific shifts in topic attention back to the original news articles that most closely associate with topic coverage at the time of the shift. For example, we find that the largest shifts in industrial production growth expectations are associated with news article headlines, including "Survey Finds Sharp Drop in Consumer Confidence," "Slower Growth in Manufacturing Raises Concerns," and "Gasoline Prices Fall as Imports Rise and Demand Drops." Using the topic model as a narrative retrieval device, the researcher can map any shock of interest to a nuanced verbal interpretation of economic events (i.e., news articles) that are statistically closest to the shock. In essence, the estimated model flags articles that the researcher should read throughly, alleviating the need to manually read every article in the history of the *WSJ* corpus. By combining macroeconomic analysis with topic modeling, the researcher achieves the hitherto unmanageable task of drawing on vast written texts to better understand quantitative economic phenomena.

We end the paper with brief explorations of extensions that highlight the fascinating possibilities for text-based approaches to economic modeling. First, because news data arrives daily, we can estimate the state of economy at a higher frequency than most macroeconomic series allow. We show how the time-varying correlation of high frequency financial frictions measures with our news topics provides a narrative understanding of episodes such as financial crises in which events impacting the macroeconomy unfold rapidly—inbetween monthly or quarterly macro announcements.

We also show that the slew of economic policy uncertainty indices developed by (Baker et al., 2016, BBD) are closely related to our estimated topics. For example, their government spending and fiscal policy uncertainty index has a 66% correlation with our "government budget" topic {key

[2]See Kelly et al. (2019) for an alternative approach to incorporating textual data into macroeconomic forecasting models.

terms: *spend cut, deficit reduction, grammrudman, federal budget, balance budget*}. And the BBD sovereign debt and currency crises uncertainty index looks like a combination of our "European debt" and "Southeast Asia" {key terms: *jakarta, thailand, southeast asia, rupiah*} topics, which combine to explain more than 40% of the variation in their BBD counterpart. What is impressive about this comparison is that our topics are estimated in an entirely unsupervised manner.

Section 2 briefly discusses our model structure and estimation. In Section 3 describes the estimated structure of economic news underlying *WSJ* text. Finally, Section 4 analyzes the connection between news text and numerical economic data, and Section 5 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 Model

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 1) imposes explicit parametric assumptions on the distribution of term counts and 2) imposes a factor structure on expected term counts. In particular, 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 thus governs the scale (or number of trials) 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, describes 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, LDA views articles as a distribution over topics. It determines how article t allocates its attention to themes associated with each topic. A brief inspection of (1) discerns a factor structure embedded in LDA, in which topics (ϕ_k) serve common factors and θ_t captures articles-specific exposure to those factors. At an intuitive level, LDA achieves its dimension reduction by trading off two goals. First, for a given topic k, it assigns high probabilities to as few terms as possible. Second, for any article, LDA tries to allocate words in the article to as to as few topics as possible. This tradeoff is optimized by identifying a relatively small number of word clusters that tend to frequently co-occur in news articles.[3]

2.2 Estimation

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 Gibbs sampler proposed by Griffiths and Steyvers (2004).[4]

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,[5] 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 given by the $K \times 1$ vector θ_t. That is, 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 given 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 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.[6]

The estimated attention that article t allocates to topic k can be represented as the frequency

[3] Paraphrased from Blei (2013).

[4] We refer interested readers to Steyvers and Griffiths (2007) for an especially transparent description of their LDA Gibbs sampler. 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.

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

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

with which its terms are assigned to k

$$\hat{\theta}_{t,k} = \frac{\sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = k)}{\sum_{q=1}^{K} \sum_{i=1}^{N_t} \mathbb{I}(\hat{z}_{t,i} = q)}, \tag{2}$$

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):[7]

$$\hat{\phi}_{k,v} = \frac{\sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(\hat{x}_{t,i} = v)\mathbb{I}(\hat{z}_{t,i} = k)}{\sum_{q=1}^{K} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \mathbb{I}(\hat{x}_{t,i} = v)\mathbb{I}(\hat{z}_{t,i} = q)}. \tag{3}$$

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

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

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

2.3 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

[7] As shown in (2) and (3), we choose not incorporate priors in our θ and ϕ estimates. Griffiths and Steyvers (2004) show that priors can be easily incorporated as

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

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 with predominantly non-economic content such as sports, leisure, books, and arts, and exclude articles that contain regular 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. We arrange term counts in an article-term matrix, $\mathbf{w}$. As described in the model section, columns of $\mathbf{w}$ correspond to unique terms in our dataset and rows correspond to individual articles. Our vocabulary 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 describes in step-by-step detail all data processing and filtering steps we take to transform raw article text into numerical counts that we use in our empirical analysis. Our final data set consists of 763,887 articles with a vocabulary of 18,432 unique terms. The left panel of Figure 1 conveys how the size of the data set varies over time, reporting the number of articles published each month and total term counts aggregated over those articles. The structure of the topic model absorbs some of the secular changes in news production by essentially modeling news term proportions rather than term count levels. But it is important to bear in mind that, our homogenization efforts notwithstanding, the *WSJ* is an evolving product that undergoes structural shifts over the course of our sample.

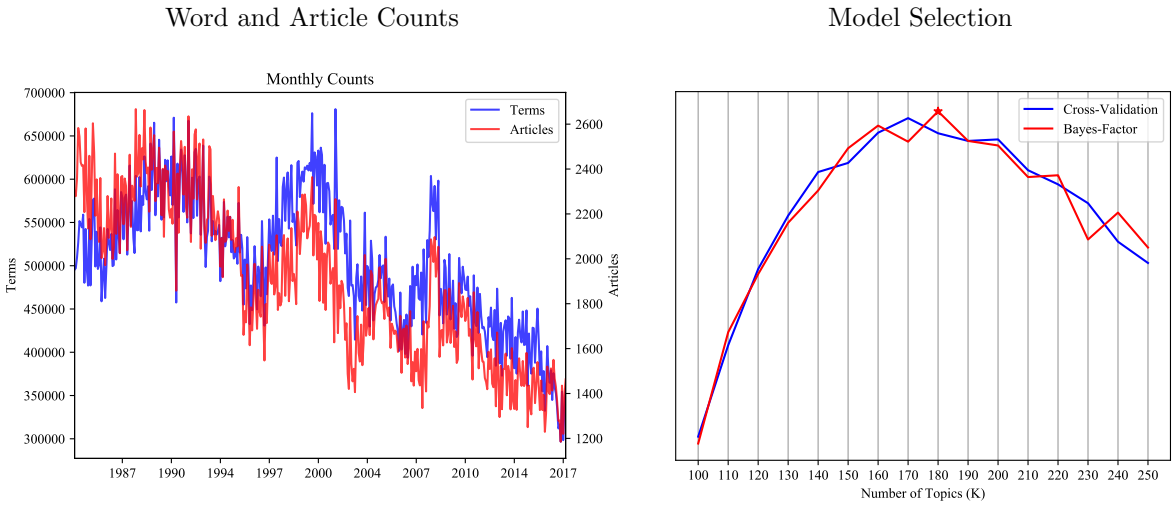
2.4 Topic Model Specification Choice

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.[8] As a complementary analysis, we also examine model selection based on ten-fold cross-validation. In particular, we partition the *WSJ* sample into ten validation samples, each having an equal number of articles. For the i^{th} model specification, we re-estimate the K_i-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 calculate an 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

[8]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 final filtered article count each month (red line) and the total term counts aggregated over all articles in each month (blue line). The right panel plots the cross-validated model fit (average log likelihood over the validation samples) and Bayes factor for topic models with K ranging from 50 to 250 in increments of 10.

our specification criteria. Manual inspection of estimates for various choices of K also indicates that $K = 180$ is a sensible choice. A model with many fewer topics (e.g., 50) produces mixed topics that contain multiple separate themes. A very large 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 decomposition.

3 The Structure of Economic News

In this section we dissect the model representation of economic and financial news from the *WSJ*. The topic model estimates reported below focus on the 180-topic specification and are based on a single estimation using the full data sample. Note that, unlike other dimension reduction techniques such as principal components 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 in detail, we have built an interactive website, `www.structureofnews.com`, to allow users to visualize and inspect a wide variety of features from our estimated model.

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 define 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 calculate scaled topic-term weights:

$$\tilde{\phi}_{k,v} = \frac{\hat{\phi}_{k,v}}{\sum_{q=1}^{K} \hat{\phi}_{q,v}}.$$

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 many topics (terms like "price" and "company"). Sorting the elements of $\tilde{\phi}_k$ identifies the terms that are most diagnostic of thematic content of topic k.

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, 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 very broadly. There are no topics that appear to mix topics. To draw a contrast, when we estimate a smaller model with $K = 50$ topics, we find articles 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.

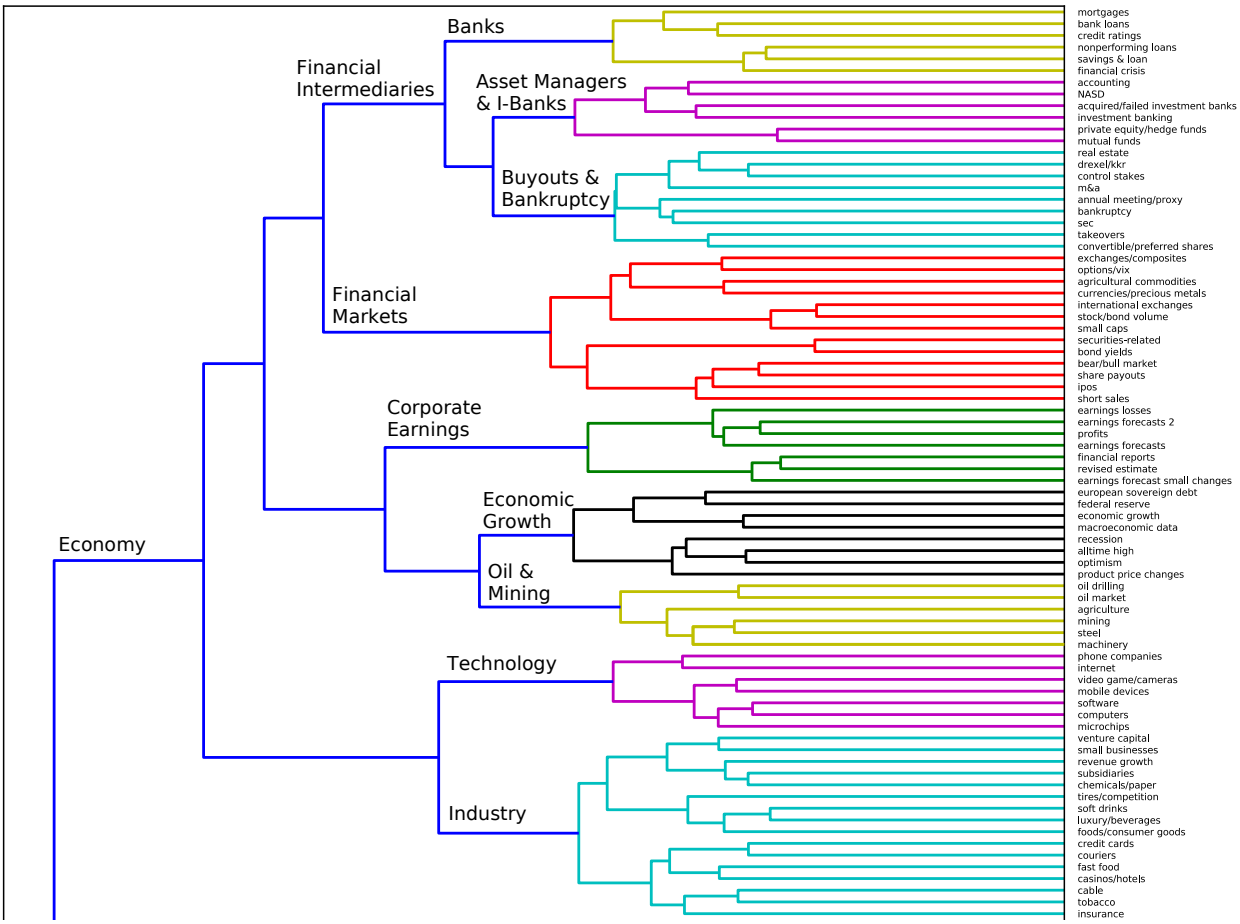
Second, we see that most topics represent news subjects, as opposed to appraisals of good versus bad news for a given subject. Examples of appraisal-free subject topics include "airlines," "Federal Reserve," and "China." Some subject topics carry an implicit appraisal, usually of bad news, such as "terrorism," "natural disasters," and "recession." A small minority of sentimental topics are interpretable as good versus bad news and includes topics labeled "problems," "concerns," and

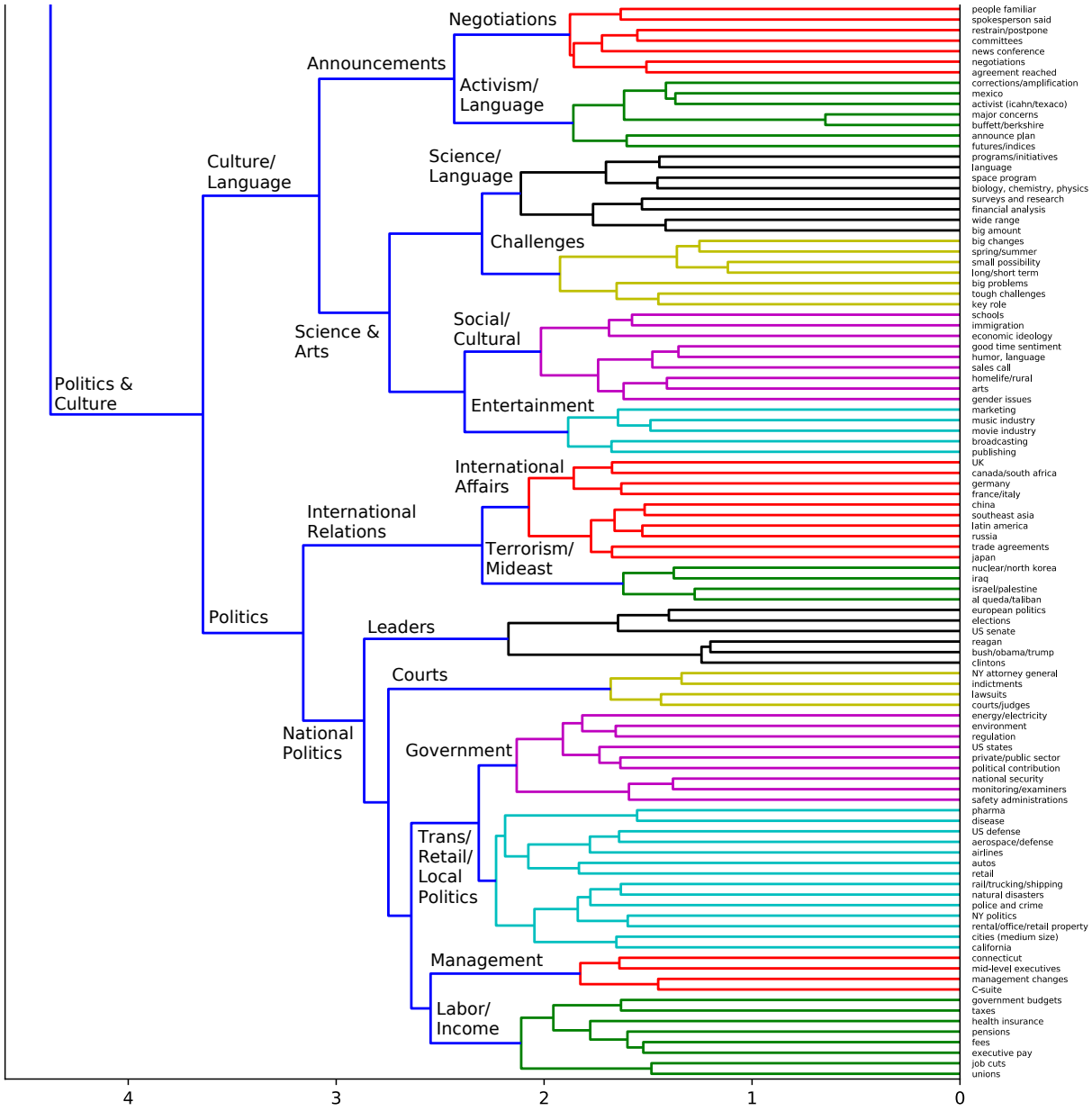
[9] Another example in the $K = 50$ specification is a topic with key terms

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.

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.

"positive sentiment." We interpret these as "modifier" topics that, when combined in a news article with a given subject topic, convey a signed/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

[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.

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 dendrogram is entirely data driven, and we only supply metatopic labels based on our manual read of the estimated clusters.

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 the 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—its 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.

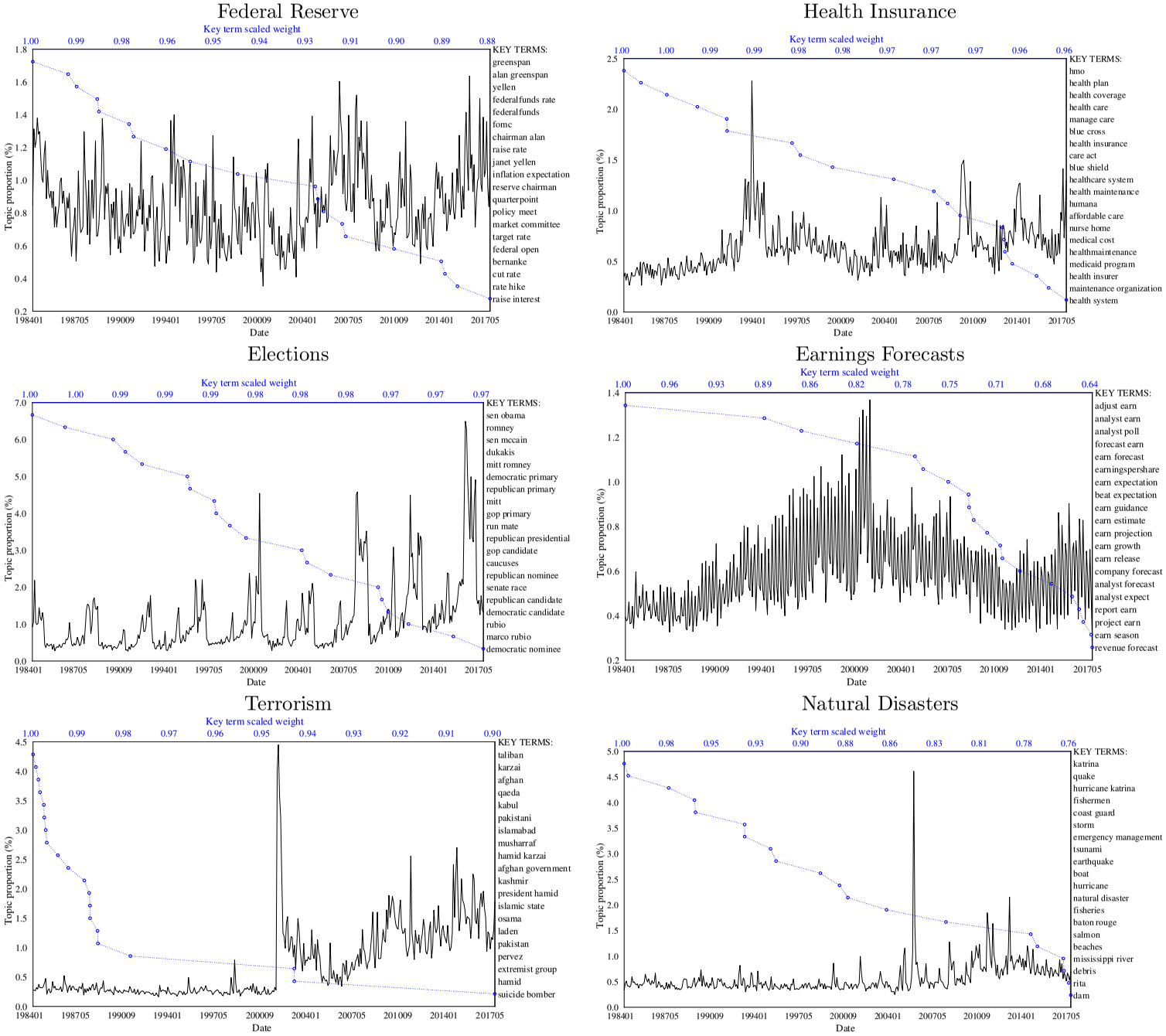
3.3 News Attention by Month

In the model, topic attention proportions are estimated at the article level but can be summed to any level of aggregation according to equation (4). Throughout this section we focus on aggregate *WSJ* topic attention at the monthly frequency by collating article-level attention estimates for all articles published in the same calendar month. Figure 3 shows the time series variation in attention for a subset of a 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. More specifically, it reports the fraction of total terms published in a given month that the model infers are drawn from that topic. 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.

News consists of a combination of recurrent, seasonal, and episodic topics. Recurrent topics reappear consistently throughout the sample. For example, the "Federal Reserve" topic is regularly active throughout the sample. As is the "health insurance" topic, which becomes highly prevalent

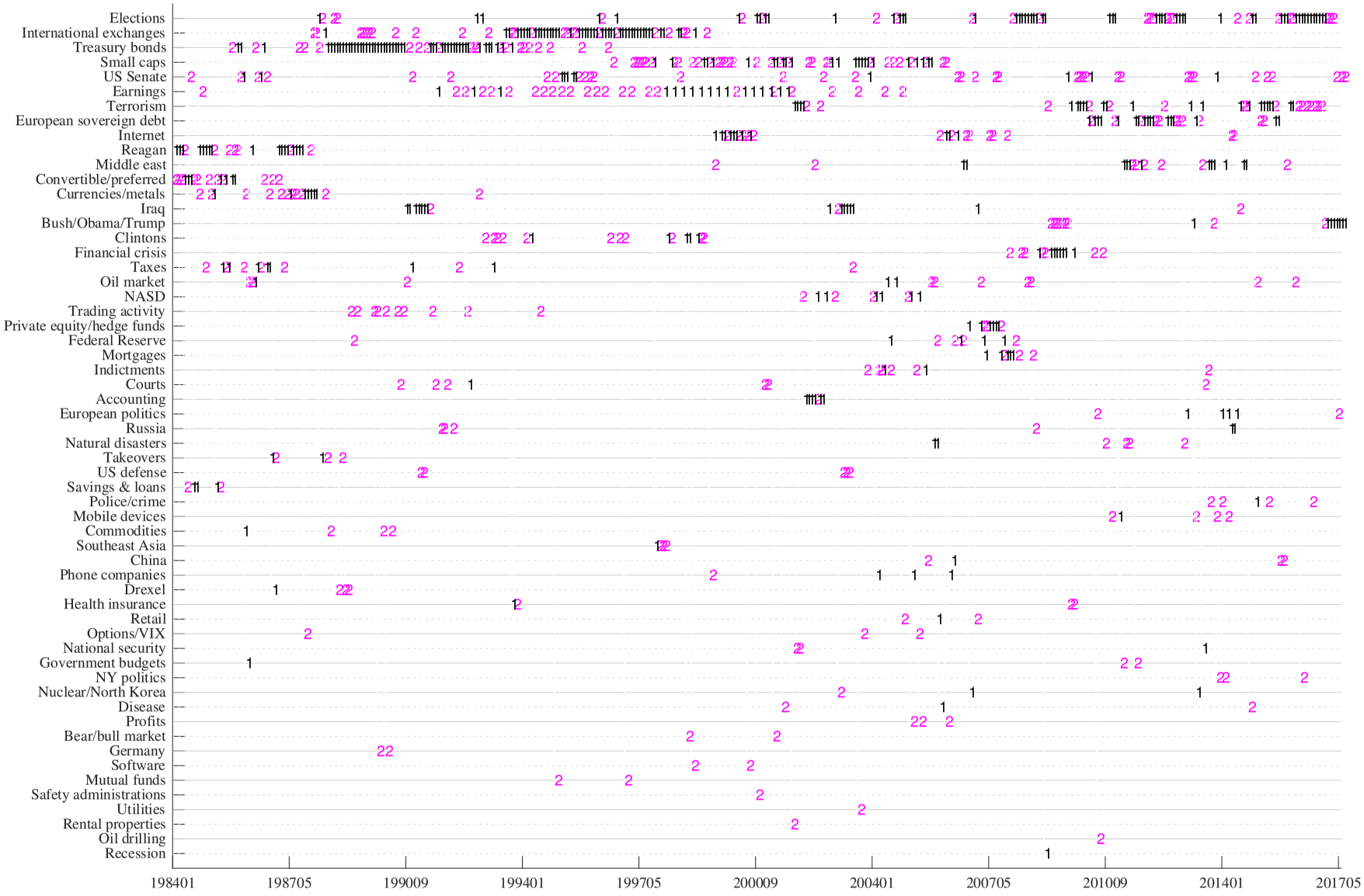
[11] Appendix Figure 15 plots an additional 24 illustrative topic attention time series, and the website for this research project, `www.structureofnews.com`, includes time series plots for all 180 topics.

Figure 3: Topic Attention Proportions



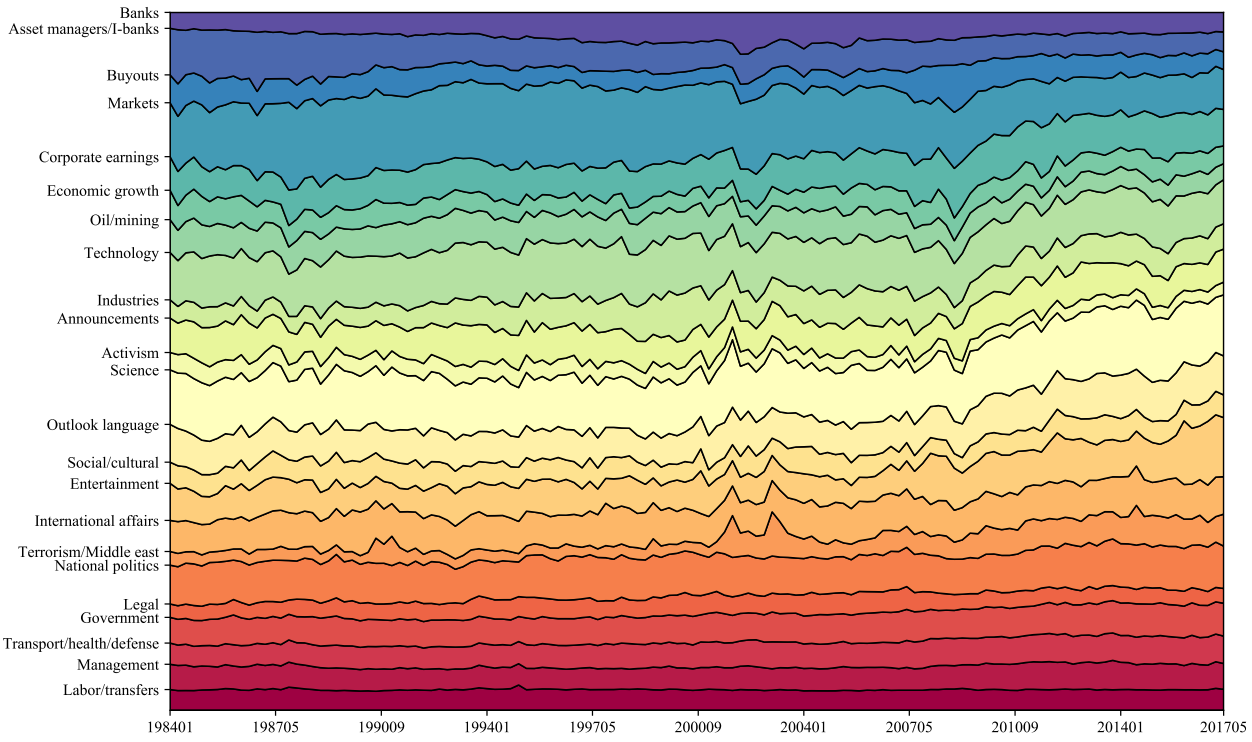
Note. The black line shows the topic attention as a percent of total monthly *WSJ* news production. The right vertical axis lists the topic's key terms, and the blue line shows the weight of each term in the topic (scaled by the maximum term weight within that topic).

Figure 4: 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 5: News Attention Across Metatopics

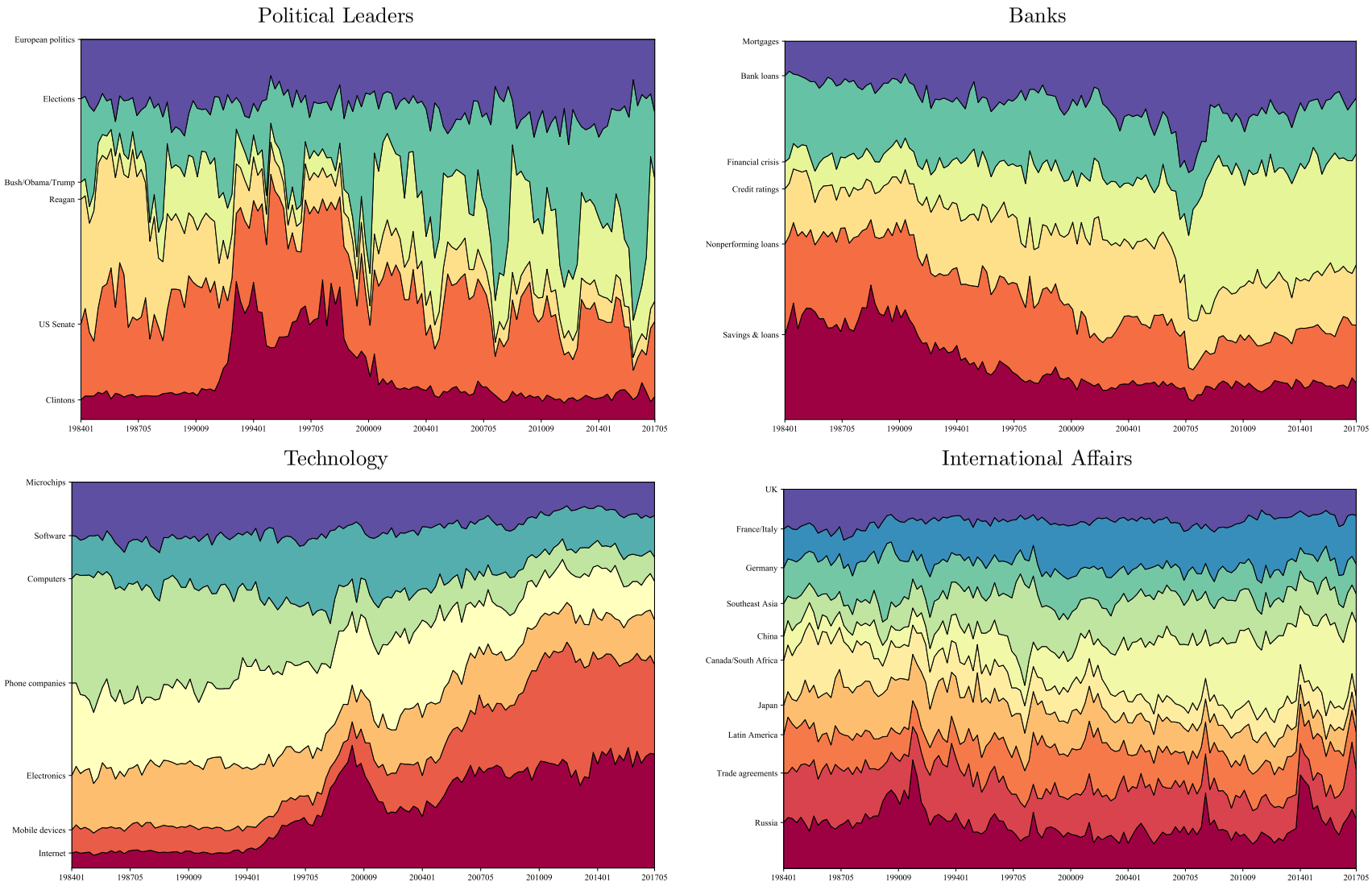


Note. Stacked area plot of attention allocation to 23 metatopics based on the first-level clustering of baseline topics.

during debate of the Clinton Health Plan proposal (peaking around President Clinton's speech to Congress in September 1993), the proposal and passage of Obamacare in 2008–2010, and around the debated repealing of Obamacare around the 2016 presidential election. Seasonal topics include "presidential elections," whose attention 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. Episodic topics are dormant for much of the sample, then spike around particular events. For example, 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 the effects of 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 overcome this, Figure 4 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 6: News Attention Within Metatopics



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

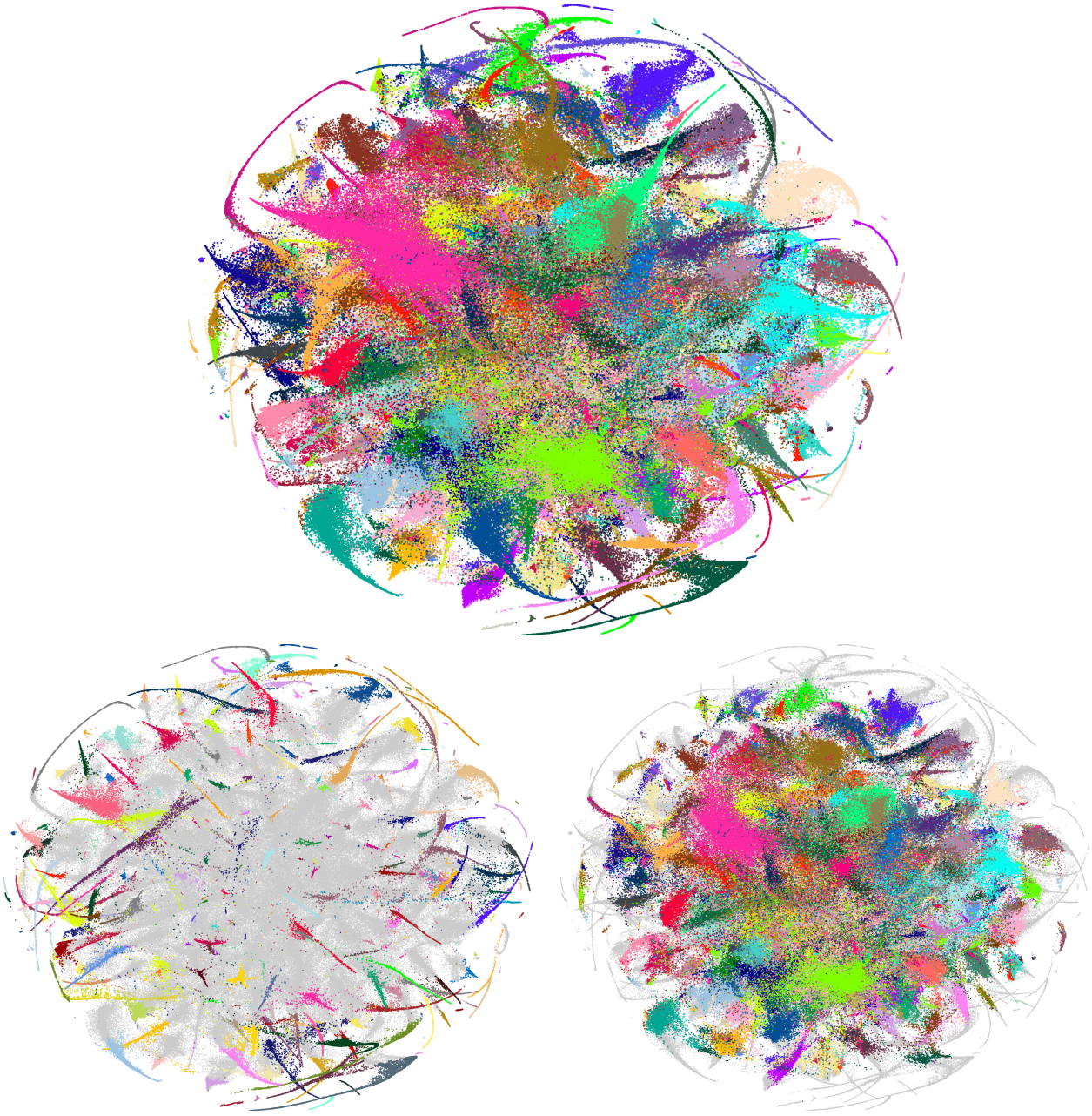
Figure 5 zooms out to show how attention behaves at the 23-metatopic level. The fraction of news attention allocated to metatopics is remarkably stable over time. This metatopic stability belies a high degree of attention churn within metatopic clusters. The upper left panel of Figure 6 shows stable attention to the "political leaders" metatopic as a whole, but that attention within this metatopic transitions dramatically as power changes hands, from Reagan at the beginning of the sample to Trump at the end. 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 stable attention to "technology" is dominated by mainframe and desktop computers at the start of the sample, which 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. 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.

3.4 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 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 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 7 shows the two-dimensional embedding of model fits for every *WSJ* article in our data set, and Figure 8 breaks down the complexity of Figure 7 by separately plotting articles based on their leading topic for an illustrative subset of nine topics. In Figure 7, 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 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 7, 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

Figure 7: Article-level Nearest Neighbor Embedding

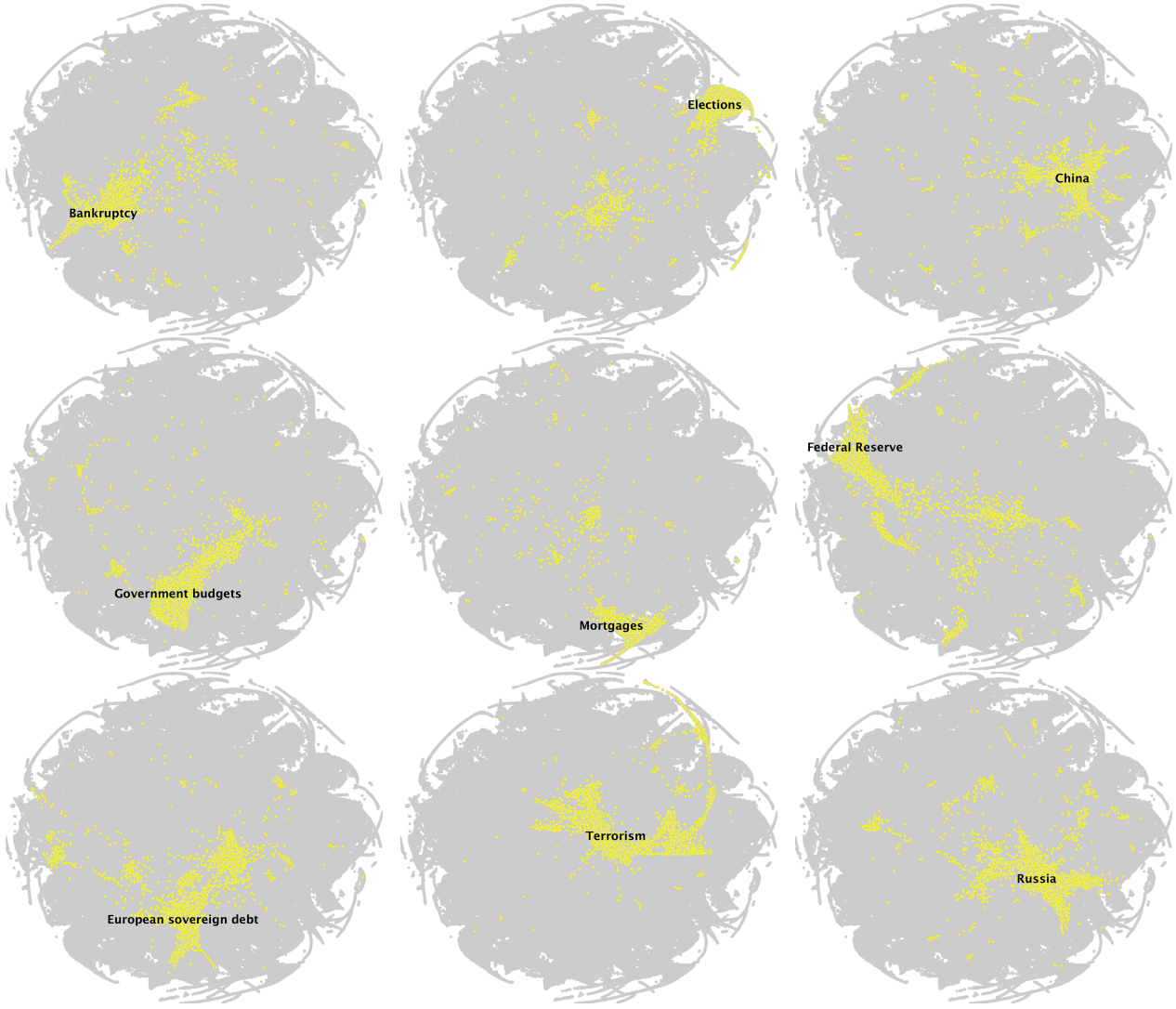


Note. The top shows the two-dimensional embedding of *WSJ* articles, where each topic is assigned a color and articles are colored in correspondence with their dominant topic. The bottom left panel highlights only "pure" (those with maximum topic proportion exceeding 33%), while the bottom right shows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

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 9 provides more detail for one specific example. It documents the other topical content of

Figure 8: Article-level Nearest Neighbor Embedding (Topic Breakout)

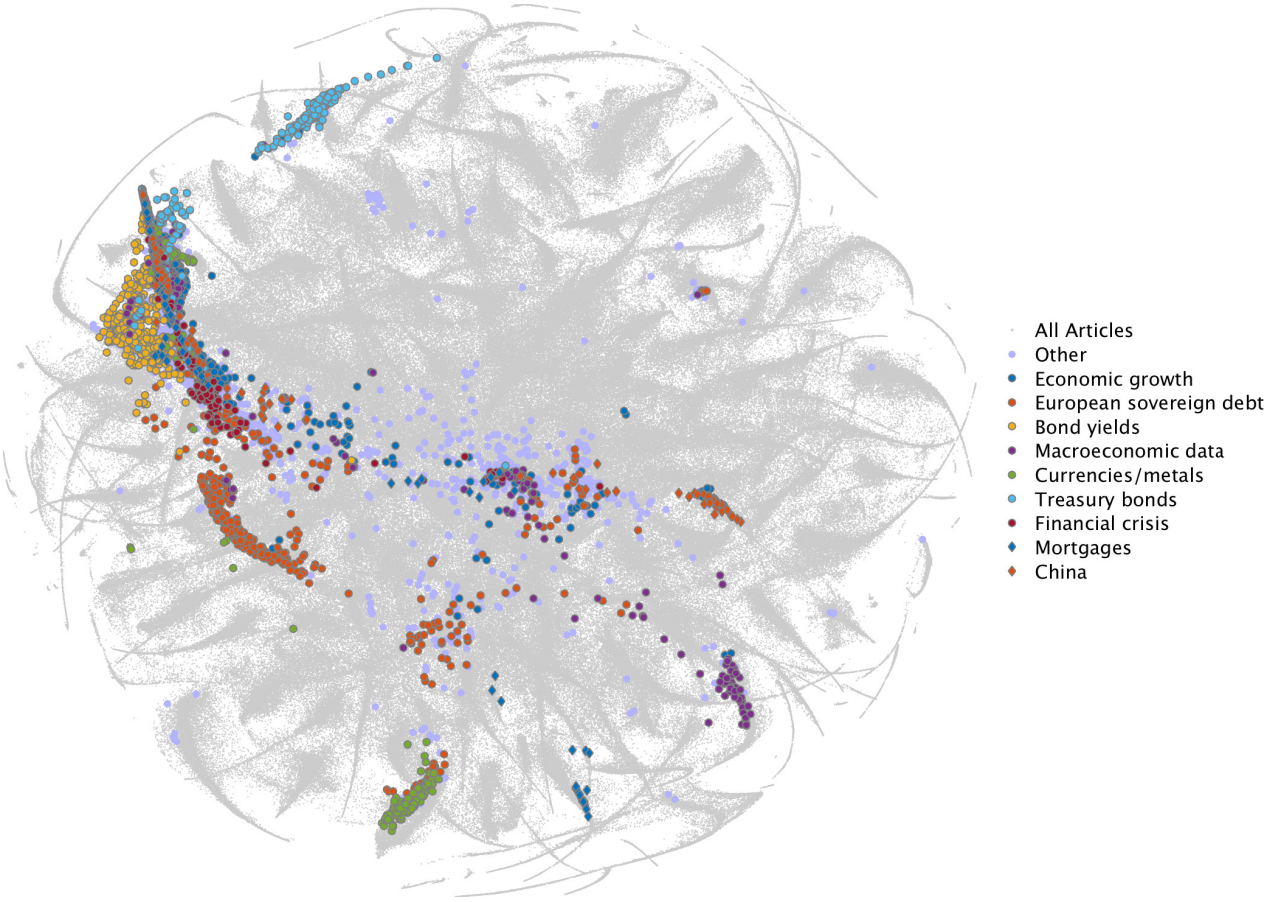


Note. Two-dimensional embedding of *WSJ* articles broken out by individual topics.

articles whose leading topic is "Federal Reserve." It splits articles within the Fed topic based on their second largest topic allocation. Figure 9 shows that articles about the Fed also tend to relate to the "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

Figure 9: 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.

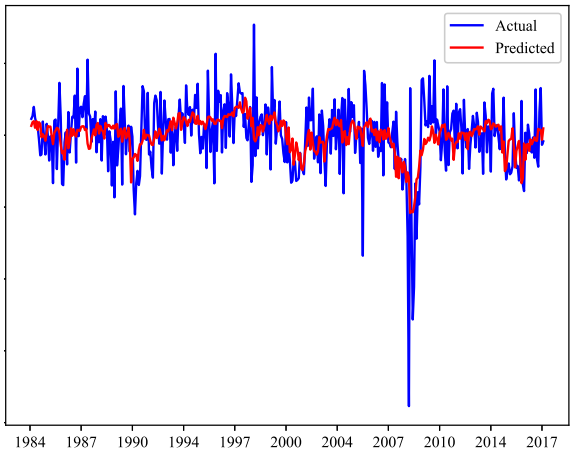
at parsimoniously representing the verbal state of the economy, while a drastic simplification of the full *WSJ* narrative, nonetheless requires an elaborate specification to capture its common themes. But our model estimation results, and our further empirical analysis below, also demonstrate that this complexity can be effectively harnessed for economic analysis.

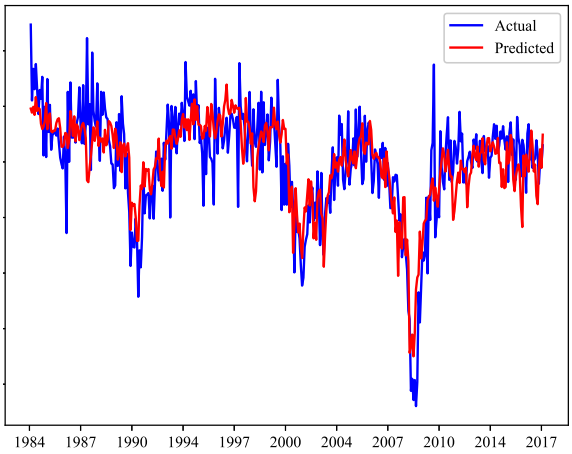
4 News Attention and the State of the Economy

In this section we empirically investigate how faithfully the topical content of economic news captures the state of the economy. Our analysis proceeds by linking the output of our estimated topic model, both in terms of topical content and attention allocation across topics, to commonly studied numerical macroeconomic data series.

Table 1: Reconstructing Macroeconomic Time Series

Industrial Production Growth			Employment Growth		
Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
Recession	-0.38	0.03	Recession	-0.61	0.00
Oil market	-0.18	0.00	Rail/trucking/shipping	0.21	0.00
Southeast Asia	0.11	0.07	Bush/Obama/Trump	-0.15	0.08
Health insurance	0.06	0.96	Iraq	-0.14	0.00
Clintons	0.03	0.23	Clintons	0.13	0.07
R^2	0.22		R^2	0.59	





Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference.

4.1 Reconstructing Economic Activity With News Text

We begin by investigating the extent to which our estimates of news topic attention align with a range of numerical measures of economic and financial activity. We then move to predictive analysis via vector autoregression.

4.1.1 Macroeconomic Activity

Ideally, an accurately estimated vector of economic state variables would be able to statistically reconstruct other measures of macroeconomic activity with a high degree of accuracy.

We attempt to reconstruct a candidate numerical measure of economic activity, x_t (e.g., US industrial production growth), through a time series regression of x_t on set of estimated topic attention proportions, $a_t = (a_{1,t}, ..., a_{K,t})$. 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 180 topics, the attention vector is high-dimensional relative to the number of monthly time series observations, so OLS regression of x_t on a_t is subject to severe overfit. To limit the undue influence of data mining on our conclusions, we apply heavy lasso penalization in the estimation. We make a stark choice on the extent of penalization—we set the lasso penalty parameter such that exactly five out of the 180 coefficients take non-zero values. We choose a fixed five-variable regression

to ensure clear interpretability of results, as well as to have uniformity and comparability across the many x variables that we study.[12]

Because lasso penalization behaves like 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 approach of Tibshirani et al. (2016). We scale regression coefficients to reflect the dependent variable's response in standard deviations to a one standard deviation change in the regressor.

To begin, we study two high profile macroeconomic indicators that are available at the monthly frequency, aggregate output (industrial production) and employment (non-farm payrolls). The left side of Table 1 reports results for the lasso regression of (seasonally adjusted) monthly log industrial production growth on the 180 news attention time series. With a lasso penalty chosen to include exactly five non-zero regression coefficients, we find that attention to "recession," "oil market," "Southeast Asia," "health insurance," and "Clintons" news topics most closely associate with contemporaneous industrial production growth. After accounting for the effects of model selection, only "recession" and "oil market" significantly correlate with output growth at the 5% level. A one standard deviation rise in *WSJ* attention to these two topics associates with 0.38 and 0.18 standard deviation drops in industrial production growth, respectively. With just five non-zero topics, we find that fluctuations in news topic attention explain 22% of output growth variance.

Log employment growth most closely associates with attention to "recession," "rail/trucking/shipping," and "Iraq" topics, which are significant at the 5% level. The fitted regression explains 59% of the variance in realized employment growth. A one standard deviation rise in "recession" attention associates with a 0.61 standard deviation decrease in employment growth and the "Iraq" topic associates with a 0.14 standard deviation drop. In contrast, a standard deviation rise in attention to "rail/trucking/shipping" aligns with a 0.21 standard deviation rise in employment growth. Finally, the lasso regression juxtaposes the employment experience under various presidential administrations, though these are insignificant effects at the 5% level.[13] As a whole, Table 1 illustrates that news attention closely tracks macroeconomic indicators, and the most correlated topics have intuitive thematic associations with the numerical data.

4.1.2 Financial Activity

We next analyze the ability of *WSJ* to track fluctuations in financial markets, including various forms of financing activity, measures of market risk, and values of financial assets.

In terms of financing activity, we study IPO volume and leveraged buyout (LBO) volume. Table 2 shows that news attention closely tracks fluctuations in each financial time series with R^2 values of 43% 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

[12] Separately tuning cross-validation parameters for each x generally produces even higher R^2 values than what we present here, but do not provide much additional economic insight.

[13] 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.

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.

The upper right panel of Table 2 reports the monthly count of bankruptcies among US public companies. News attention explains the variance in bankruptcy intensity with an R^2 of 42%, with the three significantly associated topics "recession," "accounting," and "small caps" all being themes that closely relate to debt default. The lower left panel reports regressions of average CDS spreads among Eurozone countries on news attention. A one standard deviation rise in the lone significant news topic, "European sovereign debt," associates with a 0.59 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 and reveals that this data series only amounts to a handful of distinct observations. Despite this, news text data is informative enough that lasso regression hones in on the correct narrative account of these events.[14]

The remaining two panels study risk and return in the US stock market. The lower center panel shows that the five-topic news attention regression captures 63% of the variance in stock market volatility.[15] By far the most important topic is "recession," which carries a coefficient of 0.54, and reflects the strongly countercyclical nature of market volatility. It also shows that there is more news attention to options and the VIX index, and tends to describe adverse conditions more generally (captured by the "problems" topic) when market volatility is elevated. It also shows that discussion of "small business" and "electronics" topics tends to associate with the low volatility regimes in our sample.

The lower right panel regresses returns on the value-weighted US stock market index on AR(1) innovations in topic attention. Realized returns significantly negatively associate with increases in "recession" attention. Collectively, the most informative five topics explain 25% of the variation in stock market fluctuations. This is a rather extraordinary result. One of the central puzzles in financial economics is the notorious inability to explain stock market fluctuations even ex post with anything other than other asset prices (Roll, 1988).[16] As a benchmark for comparison, AR(1) innovations in the four components of the Chicago Fed National Activity Index[17] jointly explain less than 5% of the variation in contemporaneous monthly stock returns. As another benchmark, we consider a contemporaneous regression of market returns on AR(1) innovations in the 108 macroeconomic variables in the FRED-MD database.[18] To remain comparable with news attention regressions, we

[14] While other 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.

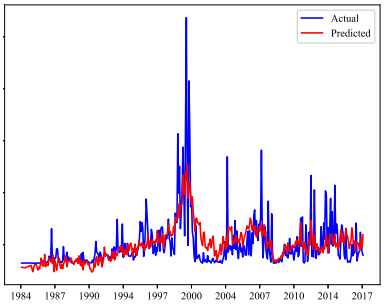
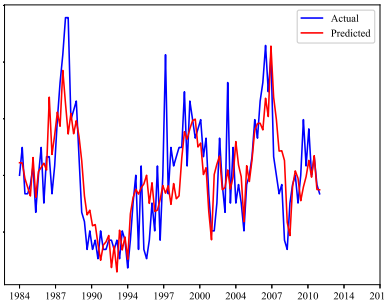
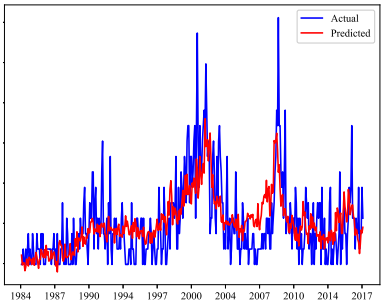
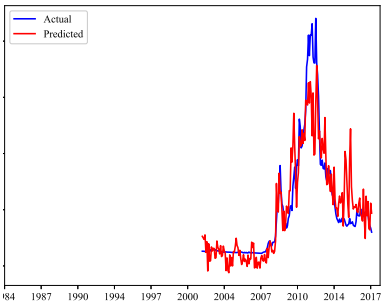
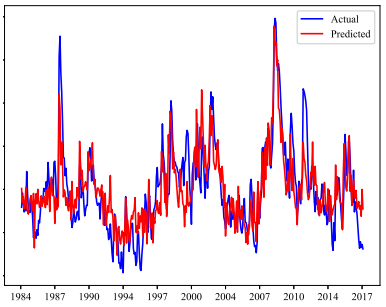
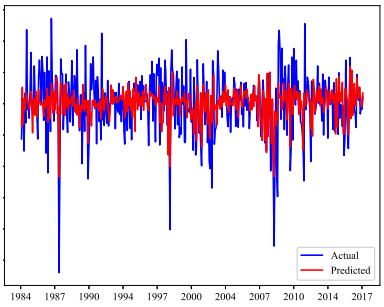
[15] We measure monthly stock market volatility as the standard deviation of daily market returns within the month.

[16] Our analysis focuses on explaining aggregate stock market fluctuations. Boudoukh et al. (2018) examine the complementary question of explaining idiosyncratic returns for individual stocks using news text.

[17] The four component indices measure production/income, employment/unemployment/hours, personal consumption/housing, and sales/orders/inventory.

[18] Available at https://research.stlouisfed.org/econ/mccracken/fred-databases/. We exclude FRED-MD in-

Table 2: Reconstructing Financial Activity

Topic	Coeff.	p-val.	Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
IPO Volume			LBO Volume			Bankrupty Filings		
IPOs	0.27	0.00	Takeovers	0.31	0.28	Recession	0.48	0.00
Venture cap.	0.22	0.00	Insurance	-0.28	0.02	Venture cap.	0.25	0.06
Bankruptcy	-0.14	0.20	Control stakes	0.27	0.03	Accounting	0.24	0.03
Internet	0.12	0.05	Key role	0.23	0.70	Small caps	0.12	0.00
M&A	0.09	0.61	Job cuts	-0.18	0.13	Machinery	-0.06	0.34
R^2	0.43		R^2	0.58		R^2	0.42	
								
European CDS Spreads			Stock Market Volatility			Stock Market Return		
Euro. sov. debt	0.59	0.00	Recession	0.54	0.00	Recession	-0.33	0.00
Middle east	0.16	0.26	Options/VIX	0.28	0.00	Problems	-0.21	0.06
Govt. budgets	0.13	0.35	Electronics	-0.24	0.02	Convert./pref.	-0.17	0.46
Company spokes.	0.13	0.91	Problems	0.16	0.00	Options/VIX	-0.11	0.95
Mobile devices	0.11	0.60	Small business	-0.15	0.00	Record high	0.10	0.08
R^2	0.79		R^2	0.63		R^2	0.25	
								

Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference.

use a lasso penalty to select exactly five non-zero coefficients among the macro variables. We find an R^2 of 10%, which exceeds that from the CFNAI regression, but is less than half of the text-based result.[19] The ability of non-price variables, and interpretable text-based regressors in particular, to produce such a large gain in explained return variation, is refreshing progress in resolving the Roll (1988) puzzle. It indicates that the lack of explanatory power in past analyses is likely to be driven by inadequate methods for introducing narratives into the statistical model, a problem that our approach helps solve.

terest rate and exchange rate variables. These are asset prices and have mechanically links to stock market returns.

[19]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%.

4.1.3 Industry Volatility

Our third set of contemporaneous lasso regressions investigates how well *WSJ* 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. Essentially 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 patterns that are distinct from the market volatility analysis above, 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.[20]

Table 3 reports results from five-regressor lasso regressions of industry volatility innovations on news topic attention. For the sake of brevity, the table reports results for nine of the 49 industries. 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 "health insurance" topic.

4.2 News Attention and Macroeconomic Dynamics

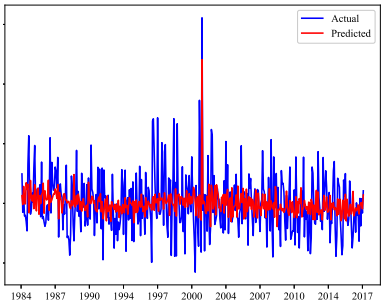
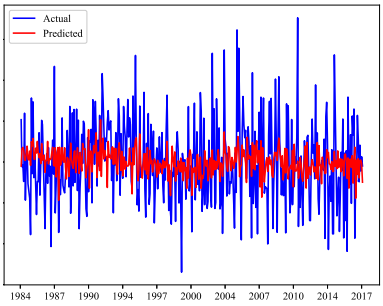
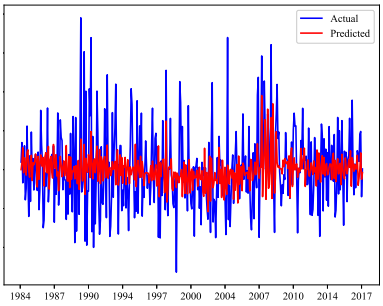
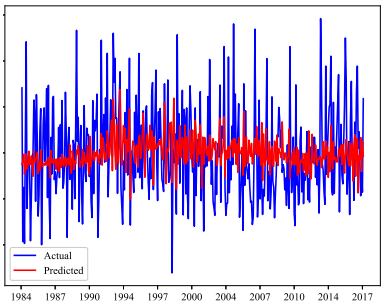
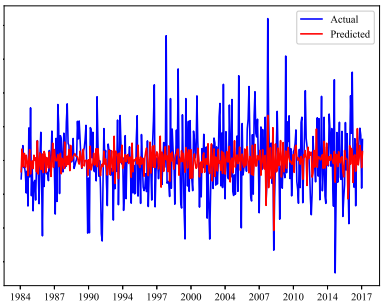
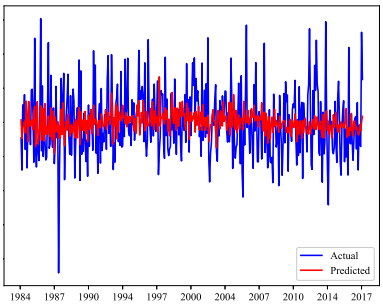
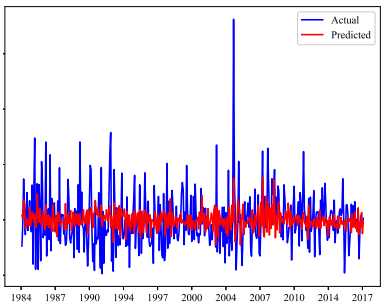
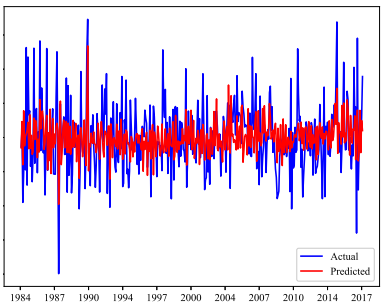
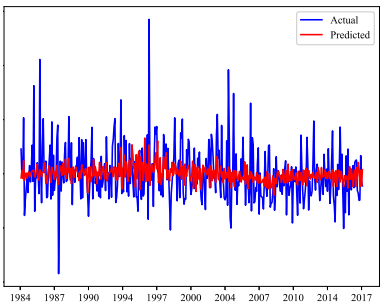
Analyzing contemporaneous correlation between news narratives of macroeconomic fluctuations helps validate the extent to which news can be converted to a numerical summary of the state of the economy—one with inherent interpretability. This, however, has little to say about whether news text conveys new information that is distinct from information in standard numerical macroeconomic indicators, and the extent to which it helps predict longer term macroeconomic trajectories.

We investigate whether news attention has a distinct impact on macroeconomic outcomes by studying its incremental forecasting power in a macroeconomic VAR. We build on the five-variable monthly VAR specification studied by Baker et al. (2016), which includes (in order) the BBD news-based measure of economic policy uncertainty (EPU), log value of the S&P 500 index, Federal Reserve funds rate, log employment, and log industrial production. We augment this VAR to include a sixth variable—media attention to the "recession" topic—based on its central role in the lasso regressions of Table 1. We define orthogonal shocks based on the Choleski decomposition for this variable ordering, and include three lags of all variables.

We estimate output and employment impulse response functions for a shock to recession news

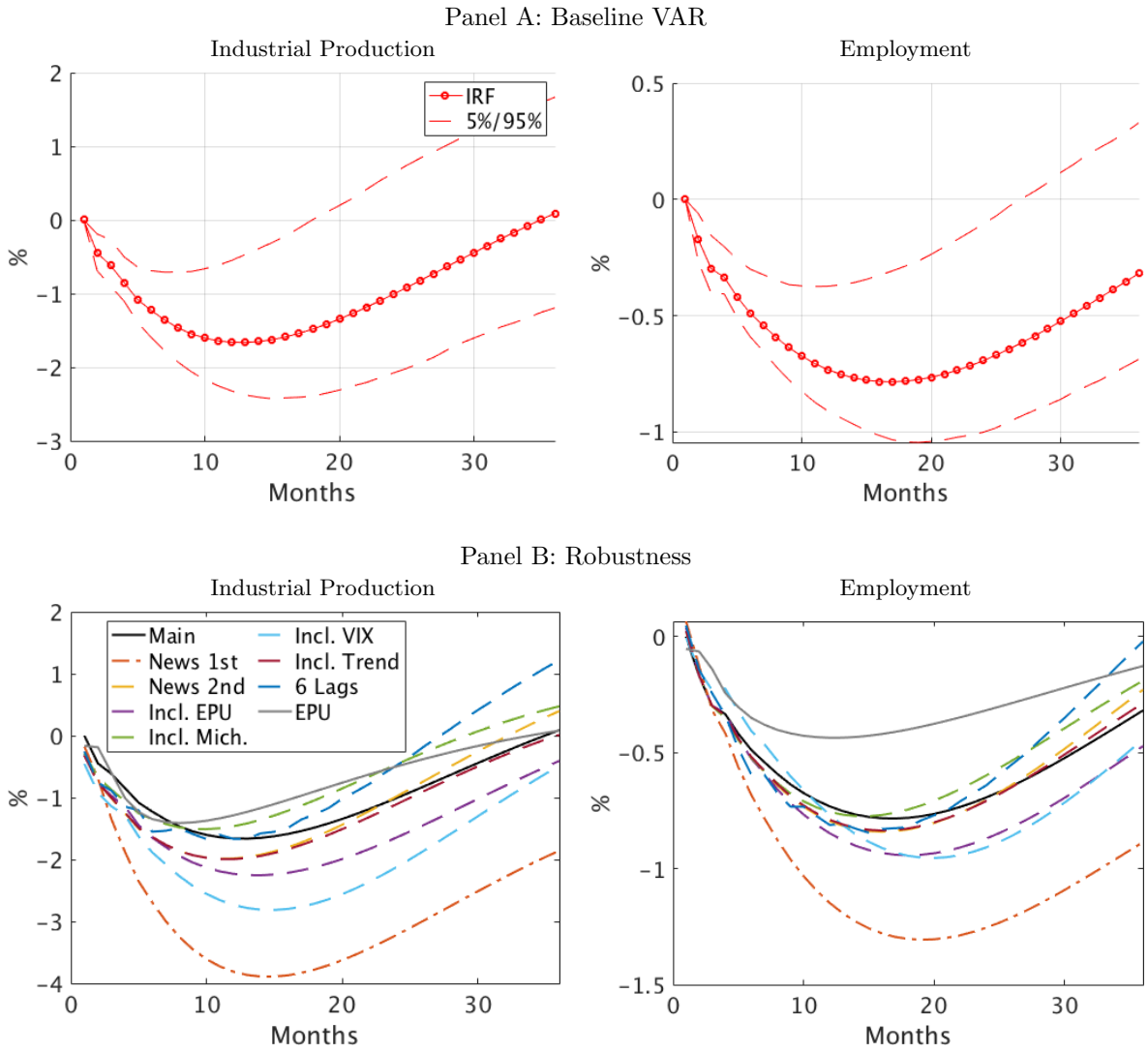
[20] Without the principal component adjustment, we essentially recover the results seen for market volatility in the lower center panel of Table 2. While the principal component adjustment is necessary for our conclusions, the results from our analysis are largely unchanged if we use levels of orthogonalized industry volatility rather than AR(1) innovations.

Table 3: Reconstructing Industry Volatility (Innovations)

Topic	Coeff.	p-val.	Topic	Coeff.	p-val.	Topic	Coeff.	p-val.
Aircraft and Parts			Automotive			Banking		
Sales call	0.19	0.00	Automotive	0.18	0.01	Nonperf. loans	0.23	0.00
Airlines	0.15	0.00	Mutual funds	-0.14	0.26	Mortgages	0.18	0.00
Research	-0.15	0.03	Fees	0.11	0.12	Options/VIX	-0.13	0.31
Terrorism	0.09	0.46	Corporate gov.	0.11	0.79	NASD	-0.10	0.89
Restraint	0.08	0.96	Credit cards	0.10	0.94	Financial crisis	0.08	0.77
R^2	0.15		R^2	0.09		R^2	0.14	
								
Pharmaceuticals			Investments and Trading			Computer Hardware		
Clintons	0.19	0.05	Investment banks	0.18	0.23	Retail	0.19	0.01
Pharma	0.15	0.01	Elections	-0.13	0.40	Computers	0.17	0.01
Health insurance	0.12	0.08	Earnings losses	-0.13	0.49	Software	0.12	0.09
News conference	-0.12	0.30	Recession	-0.13	0.09	Couriers	0.09	0.96
Earnings forecasts	0.09	0.96	European politics	0.10	0.91	Competition	0.05	0.68
R^2	0.11		R^2	0.11		R^2	0.09	
								
Insurance			Oil and Gas			Tobacco		
Insurance	0.21	0.00	Oil market	0.35	0.00	Tobacco	0.18	0.01
Revised estimate	-0.16	0.09	Cultural life	0.18	0.01	Mexico	0.12	0.79
Mortgages	0.12	0.43	Casinos	-0.17	0.36	Lawsuits	0.09	0.13
Mgmt. changes	-0.10	0.91	Options/VIX	-0.16	0.10	Product prices	-0.09	0.53
Bank loans	0.09	0.54	Earnings losses	-0.15	0.09	Recession	-0.08	0.94
R^2	0.12		R^2	0.23		R^2	0.09	
								

Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference.

Figure 10: VAR Responses to Recession News Attention Shock



Note. Responses to a one standard deviation impulse to recession topic attention in various VAR specifications.

attention and plot results in Panel A of Figure 10. A one standard deviation rise in recession news attention associates with a drop in industrial production of 1.7% after twelve months. Employment experiences a maximum response of 0.8% after sixteen months. These effects are statistically significant, as shown by the 90% confidence bands (dashed lines).

The estimated responses reflect large economic magnitudes. As a benchmark for comparison, Panel B reports our replication of the impulse response to EPU (solid gray line) in the VAR model of BBD.[21] The response of industrial production to EPU and recession news attention are similar in magnitude, while employment responds nearly twice as strongly to a rise in recession news attention

[21] We replicate Figure 8 of BBD exactly, with the minor exception that we define the shock as a one standard deviation movement in the EPU measure for comparability with our attention shock definition. The resulting impulse responses to EPU are therefore slightly larger than those reported by BBD.

than it does to EPU.

The baseline impulse responses to recession news attention are conservative. Our news attention variable is ordered last in the baseline VAR, thus the impulse responses reflect the macroeconomic impact of recession news coverage after controlling for contemporaneous responses to all other variables. Panel B demonstrates the robustness of our baseline VAR conclusions to alternative specifications. First, we consider ordering news attention first in the VAR while otherwise preserving the order of covariates. This has a large impact our conclusions, more than doubling the recession news impulse response from production and almost doubling the effect on employment. If instead we alter the order of variables so that the news attention variable is second, after the S&P 500 index but before the Fed funds rate, production, and employment, we see slight increases in the magnitude of both industrial production and employment responses. The difference in responses for "news 1st" and "news 2nd" specifications highlights that news is often revealed simultaneously by media narratives and asset prices.

We then add BBD's EPU index (which we order first in the VAR, as done in BBD), which magnifies the drop in production to -2.2% at twelve months and the drop in employment to -0.9% at sixteen months following a recession news attention shock. Replacing EPU with the Michigan Consumer Sentiment Index (ordered first) attenuates the effect of recession news attention but leaves the baseline pattern unchanged and statistically significant. 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.

Our interpretation of these findings is that news attention to the recession topic captures useful information about future economic outcomes that is missed by other variables in the system. That such a readily available source of data—news text—can improve output and employment predictions above standard numeric macroeconomic series highlights fascinating possibilities for use of textual data in problems such as macroeconomic forecasting, which rely primarily on numerical data sources, and modeling the role of information transmission and media in the macroeconomy.

4.3 Narrative Retrieval

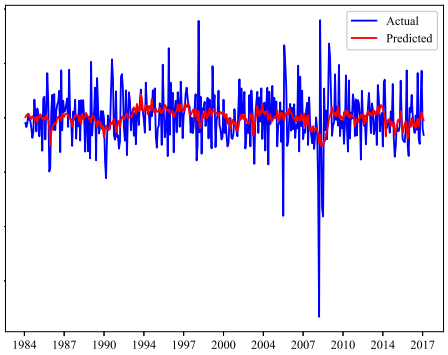
In the preceding analyses, we project a given numerical economic time series, x_t, onto topic attention to reconstitute macroeconomic time series in terms of news text. By combining estimates from this projection with the structure of the topic model, we can recover narratives corresponding to specific fluctuations in x_t or in conditional expectations of x_t. To illustrate our approach to narrative retrieval, we work with a lasso regression analysis motivated by the VAR in the previous section.

We first orthogonalize monthly industrial production growth with respect to its own lagged value in order to focus on growth rate innovations that are not forecastable with their own past values. Next, we forecast these innovations using news attention data in a five-variable lasso regression, following Section 4.1.

Table 4 reports the predictive regression results. The right panel plots industrial production growth AR(1) innovations (blue line) against their predicted values based on lagged news topic

Table 4: News-based Industrial Production Growth Forecasts

Topic	Coeff.	p-val.
Recession	-0.18	0.04
Health insurance	0.13	0.10
Russia	-0.09	0.80
Space program	-0.08	0.83
Oil market	-0.08	0.08
R^2	0.08	



Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference. Dependent variable is the AR(1) innovation in industrial production growth.

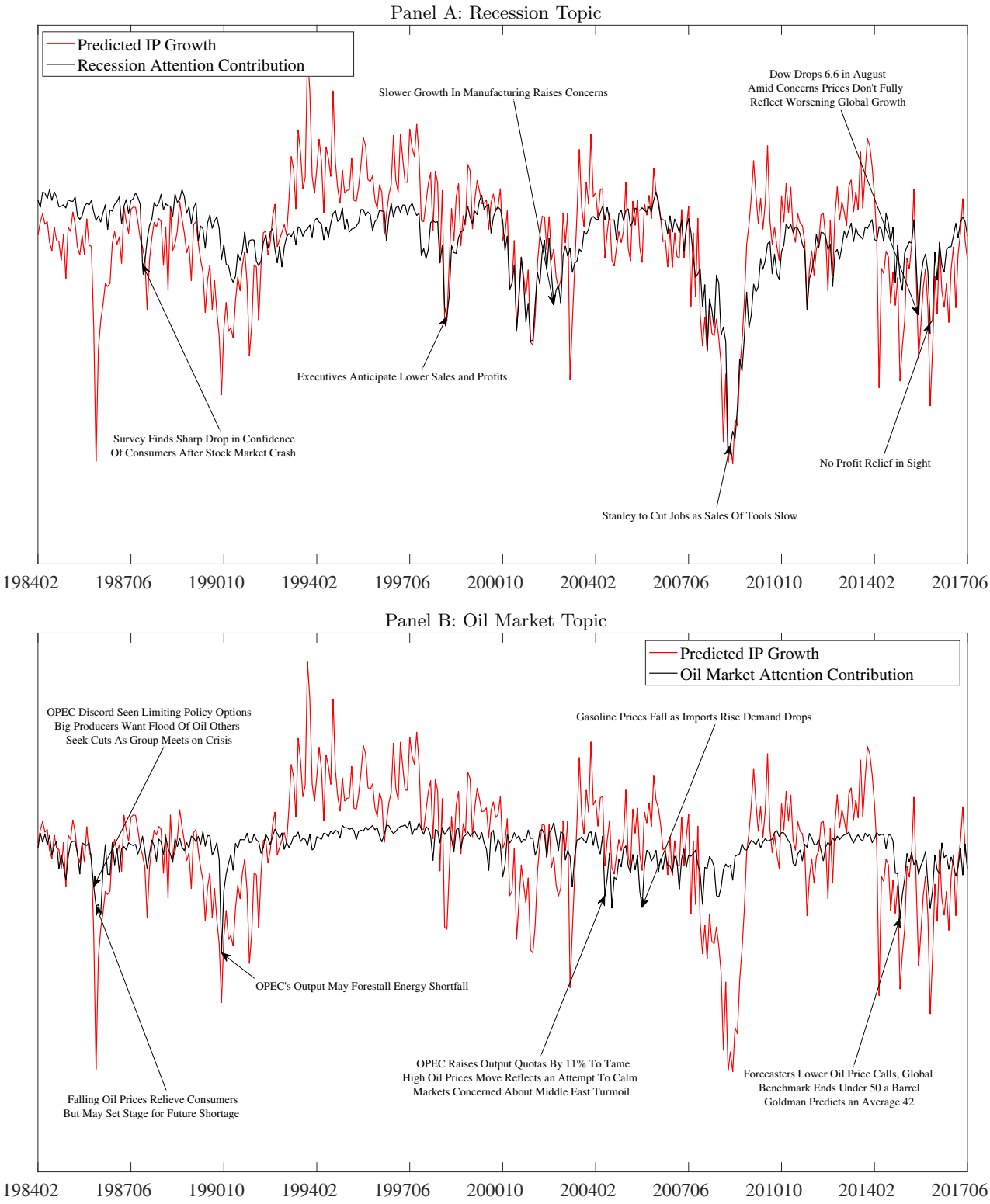
attention (red line). The left panel shows the five topics selected by the lasso, their regression coefficients, and their adjusted p-values for post-selection inference. In line with our VAR analysis above, the regression results show that the recession topic is a strong predictor of output growth. A one standard deviation increase in recession attention forecasts a 0.18 standard deviation drop in next month's production growth (after controlling for lagged growth rates). The other marginally significant forecaster of growth rates is attention to the oil market topic.

From the fitted forecasting model we can investigate which narratives are associated with changes in expected output. Both the top and bottom panel in Figure 11 show the news-based growth rate predictions in red, which correspond to the fitted values shown in the left panel of Table 4. Along with the news-based model predictions, we plot the contributions to the predicted values coming from the recession topic (top panel) and oil market topic (bottom panel). Prediction contributions for each topic are defined as the topic's prediction coefficient multiplied by its attention proportion each period.

Overlaying expected growth rates with forecast contributions shows how fluctuations in numerical economic time series can be mapped to news via topic attention regressions. From here, one can further map economic fluctuations to specific textual narratives by using topic model estimates. In each month t, we identify the *WSJ* article published during the month that is most representative of each topic k. This representative article is defined as the article i with the largest proportion of its content (θ_i) allocated to topic k.

The top panel of Figure 11 illustrates this mapping from growth rate expectations to narratives by annotating the six largest shifts in recession attention with the headlines of articles most indicative of recession news coverage at the time of the shift. For example, our model shows that narratives related to falling consumer confidence (November 1987), declining corporate sales and profits (October 1998, November 2008, and February 2016), and global growth concerns (September 2015) are key drivers of growth rate expectations in our model. Likewise, the lower panel shows that oil market narratives regarding changes in oil demand (October 2005), oil extraction (June 2004), and supply uncertainty

Figure 11: Narrative Retrieval for Industrial Production Growth Forecasts



Note. News-based growth rate predictions are in red, and contributions to the predicted values coming from the recession topic (top panel) and oil market topic (bottom panel) are in black. Each panel is annotated with the six largest shifts in attention based on the headlines of articles most indicative of topic coverage at the time of the shift.

(February 1986) influence growth expectations.

More generally, our proposed algorithm for narrative retrieval is the following. First, estimate the projection of numerical series x_t on the set of news topic attention proportions $(a_{1,t}, ..., a_{K,t})$ and recover projection coefficients $\beta_1, ..., \beta_K$ (many of which may be zero if using a penalized regression as in our earlier analysis). For a given event of interest, x_τ, identify the topics that most strongly contribute to that event through the projection. In the growth forecast example, we sought the topic that contributed most prominently to the shift in growth forecasts, corresponding to topic-dates characterized by large values of $\beta_i(a_{i,\tau} - a_{i,\tau-1})$. After identifying the key topics and events, we finally trace the event back to the original narrative text underlying that topic by finding the article published at time τ with the largest allocation to the key topics. For the purpose of Figure 11, we represented the event narrative as the headline of a single representative article. But retrieved narratives can be as rich as the researcher desires, including multiple articles with high attention on the topic of interest (rather than a single representative article) and leading sentences, paragraphs, or even full text (as opposed to just a headline).

4.4 Extensions

We conclude our analysis with three exploratory empirical analyses that highlight the breadth of questions that can be addressed with news text analysis.

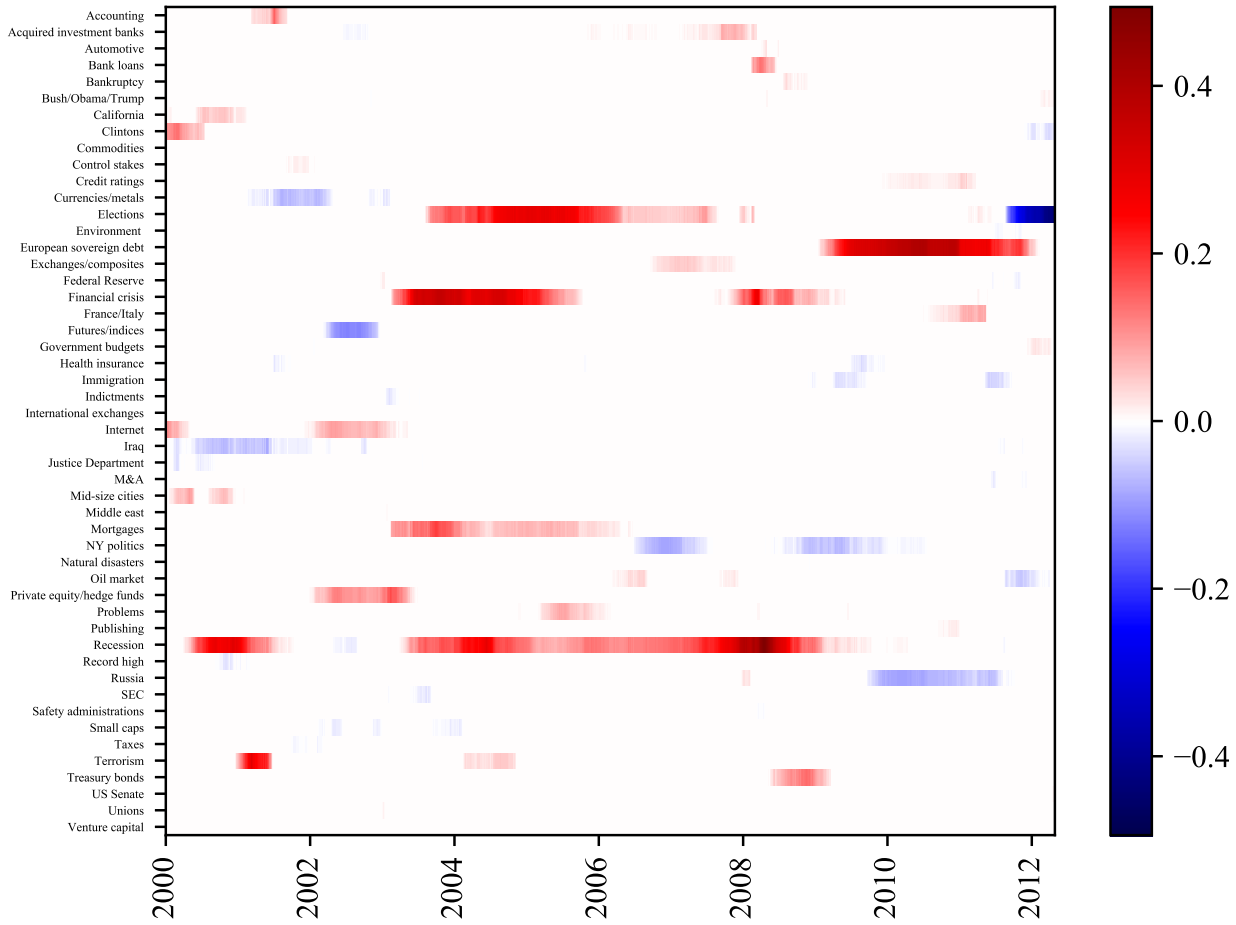
4.4.1 High Frequency Macroeconomic Modeling: A Financial Frictions Narrative

Few economists saw the 2007–2009 financial crisis coming, perhaps because it is not possible to predict such events (Cochrane, 2011), because of behavioral biases (Gennaioli and Shleifer, 2018), or because of strong beliefs in misspecified economic models (Krugman, 2009). When financial turmoil begins, economists, market participants, and policy makers gradually develop a narrative for the market's behavior (Shiller, 2017). This narrative is important as it can shape the policy response.

Recent literature has developed several measures of financial frictions that can quantify their severity in realtime. These measures can sound the alarm about emerging risks in the financial sector, but leave the problem of identifying their causes open. We next provide a method for identifying such causes by studying the time-varying correlation of a high frequency measure of financial frictions with the topics that appear in the *WSJ*.

We measure financial frictions by the spread between the price of five-year interest rate swaps and five-year treasury rates. Swap spreads represent the difference between the yield on a synthetic bond (receiving fixed on a swap) and the yield on a treasury bond of the same maturity. A larger spread captures the extent to which the financial intermediary that writes the swap is riskier than the treasury. Liu et al. (2006) have shown that this credit premium is primarily compensation for liquidity risk. Fleckenstein and Longstaff (2018) show that it is tightly related to balance sheet frictions and constraints faced by financial intermediaries. A major advantage of this measure is that it is readily available starting in mid-2000 on a daily basis. The high frequency nature of our daily newspaper data allows a higher frequency narrative for emerging risks than, for example, Hanley

Figure 12: Financial Frictions (Levels)



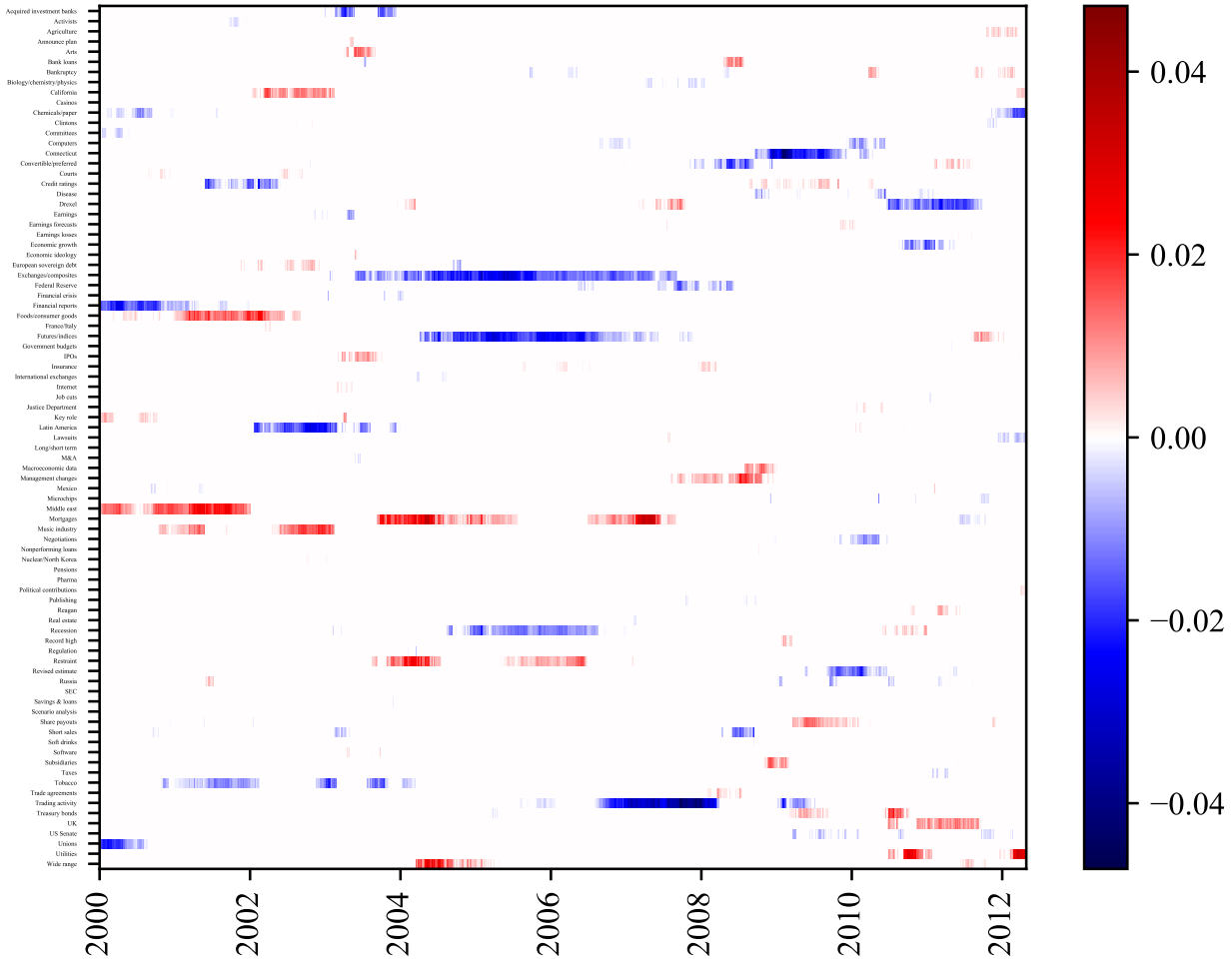
Note. We estimate which topics best explain financial frictions, as measured by the swap spread. The swap spread is the yield difference between five-year interest rate swaps and five-year treasury rates, both from FRED. We run rolling 1000-day window Lasso regressions of swap spreads on news topic proportions (levels on levels). We select a penalty that generates exactly five non-zero coefficients and plot a heatmap of these topic coefficients over time.

and Hoberg (2019) who rely on annual corporate filings.

We estimate which topics best explain intermediary frictions in rolling 1000-day lasso regressions of swap spreads on topic proportions. As before, we select a penalty that generates exactly five non-zero coefficients. We show a heatmap of topic coefficients over time in both levels (Figure 12) and innovations (Figure 13). The sparsity of the regression models provides a concise narrative for major events that exacerbated (redder) or alleviated (bluer) financial frictions.

Panel A shows that regressing levels on levels tends to focus on slow-moving trends associated with financial frictions. For example, "mortgages" and "financial crisis" light up in 2003, suggesting that concerns about mortgages and potential for an impending financial crisis partly explain the variation in swap spreads over the 2003–2006 period—well before the full blown crisis starts in the

Figure 13: Financial Frictions (Innovations)



Note. We estimate which topics best explain financial frictions, as measured by the swap spread. The swap spread is the yield difference between five-year interest rate swaps and five-year treasury rates, both from FRED. We run rolling 1000-day window Lasso regressions of swap spreads on news topic proportions (innovations on innovations). We select a penalty that generates exactly five non-zero coefficients and plot a heatmap of these topic coefficients over time.

summer of 2007. Other major topics associated with worsening liquidity risk are "recession" and "European sovereign debt". Interestingly, while coverage of the 2004 and 2008 elections coincided with worsening financial frictions, the 2012 elections are a period of improving liquidity.

A regression of changes on changes suggests a more nuanced narrative of financial frictions. Panel B shows that more topics become transiently important for understanding variation in swap spreads. For example, increased coverage of the middle east, concentrated around the 9/11 attack and the subsequent war in Iraq, is associated with increasing swap spreads. Greater coverage of "California" is associated with worsening financial frictions, long before it became an epicenter for delinquent mortgage-backed securities. Mentions of "acquired investment banks", "trading activity", and securi-

Table 5: *WSJ* Topics and Economic Policy Uncertainty Indices

Topic	Coeff	P-Value	Topic	Coeff	P-Value	Topic	Coeff	P-Value
Entitlement Programs			Financial Regulation			Fiscal Policy/Spending/Taxes		
Govt. budgets	0.45	0.00	Financial crisis	0.25	0.00	Govt. budgets	0.49	0.00
Euro. sov. debt	0.18	0.00	Competition	-0.25	0.00	Recession	0.20	0.00
Takeovers	-0.18	0.11	Bank loans	0.15	0.02	M&A	-0.16	0.00
Credit ratings	0.17	0.09	Problems	0.14	0.13	Competition	-0.16	0.33
Economic growth	0.08	0.15	Recession	0.13	0.03	Credit ratings	0.11	0.92
R^2	0.46		R^2	0.45		R^2	0.54	
Government Spending			Health Care			Monetary Policy		
Govt. budgets	0.54	0.00	Govt. budgets	0.37	0.44	Recession	0.24	0.06
Restraint	0.17	0.00	Economic growth	0.25	0.00	Airlines	0.22	0.00
M&A	-0.17	0.00	Takeovers	-0.18	0.96	Restraint	0.20	0.00
Problems	0.13	0.74	Bush/Obama/Trump	0.14	0.00	Record high	-0.18	0.00
Competition	-0.05	0.27	Euro. sov. debt	0.14	0.00	Problems	0.13	0.01
R^2	0.55		R^2	0.53		R^2	0.37	
National Security			Regulation			Sov. Debt/Currency Crises		
Iraq	0.53	0.00	Govt. budgets	0.27	0.11	Southeast Asia	0.41	0.24
Internet	-0.19	0.00	Recession	0.23	0.04	Euro. sov. debt	0.33	0.03
Airlines	0.17	0.00	Financial crisis	0.20	0.00	Music industry	-0.26	0.79
Record high	-0.17	0.00	M&A	-0.19	0.01	Movie industry	-0.13	0.15
US defense	0.09	0.06	Competition	-0.15	0.01	Problems	0.13	0.16
R^2	0.57		R^2	0.46		R^2	0.43	
Taxes			Trade Policy			Broad EPU		
Govt. budgets	0.43	0.00	Mexico	0.44	0.00	Problems	0.33	0.00
Recession	0.24	0.00	Trade agreements	0.26	0.00	M&A	-0.28	0.00
M&A	-0.15	0.00	Consumer goods	0.23	0.00	Govt. budgets	0.27	0.00
Competition	-0.15	0.00	Trading activity	0.15	0.06	Restraint	0.17	0.05
Credit ratings	0.12	0.08	Intl. exchanges	0.01	0.83	Competition	-0.09	0.07
R^2	0.49		R^2	0.50		R^2	0.53	

Note. Five-regressor lasso regression estimates and fits with p-values are adjusted for post-selection inference.

ties market "exchanges" are associated with improved liquidity, as are mentions of "Connecticut"—a hedge fund hub. Toward the end of our swap spreads sample, which ends on October 28, 2016, the greatest source of heightened financial frictions is attributed to "utilities," indicating a link between financial stability and attention on energy production and regulation around that time.

4.4.2 Unsupervised Economic Policy Uncertainty

In influential recent work, Baker et al. (2016) develop a broad index of economic policy uncertainty (EPU) based on news text, as well as a number of category-specific sub-indices that focus on particular aspects of economic policy like trade, taxes, and financial regulation. Our estimated topics are closely related to the EPU indices developed by BBD. Table 5 reports the five-regressor lasso regression of BBD's 12 EPU indices (11 catagorical EPU indices plus their broad EPU index) on our estimated news attention time series, and Figure 14 plots the model fits. For example, our five-topic

news attention regression explains 55% of the variation in the BBD "government spending" uncertainty index. This is driven primarily by our "government budget" topic {key terms: *spend cut, deficit reduction, grammrudman, federal budget, balance budget*}, which has a univariate correlation of 66% with the BBD index. The BBD "financial regulation" uncertainty index looks like a combination of our "financial crisis," "competition," "bank loans," and "recession" topics ($R^2 = 45\%$). It is interesting to note that "competition" is the only topic significantly negatively associated with financial regulation uncertainty. And the BBD "trade policy" uncertainty index behaves like a combination of our "Mexico," "trade agreements," and "consumer goods" topics ($R^2 = 50\%$).

To construct their main EPU index, BBD measure the fraction of articles in ten leading U.S. newspapers that contain the terms "economic" or "economy;" "uncertain" or "uncertainty;" and "Congress," "deficit," "Federal Reserve," "legislation," "regulation," or "White House." Their textual analysis amounts to a heavily restricted model, as the researchers decide on a few specific search terms from first principles prior to analyzing data. This approach has advantages and limitations. Thoughtfully choosing a small list of search terms ex ante has the advantage that a researcher can hone in on the text content of interest while easily avoiding the complexity and noise of surrounding text. But by requiring such detailed researcher input (i.e., priors), this textual analysis approach is especially prone to misspecification biases. Without some degree of flexibility in the text model, it is impossible to discern potentially important patterns in news outside of the pre-established term list. These consideration are a manifestation of the fundamental bias-variance tradeoff at the core of statistics.

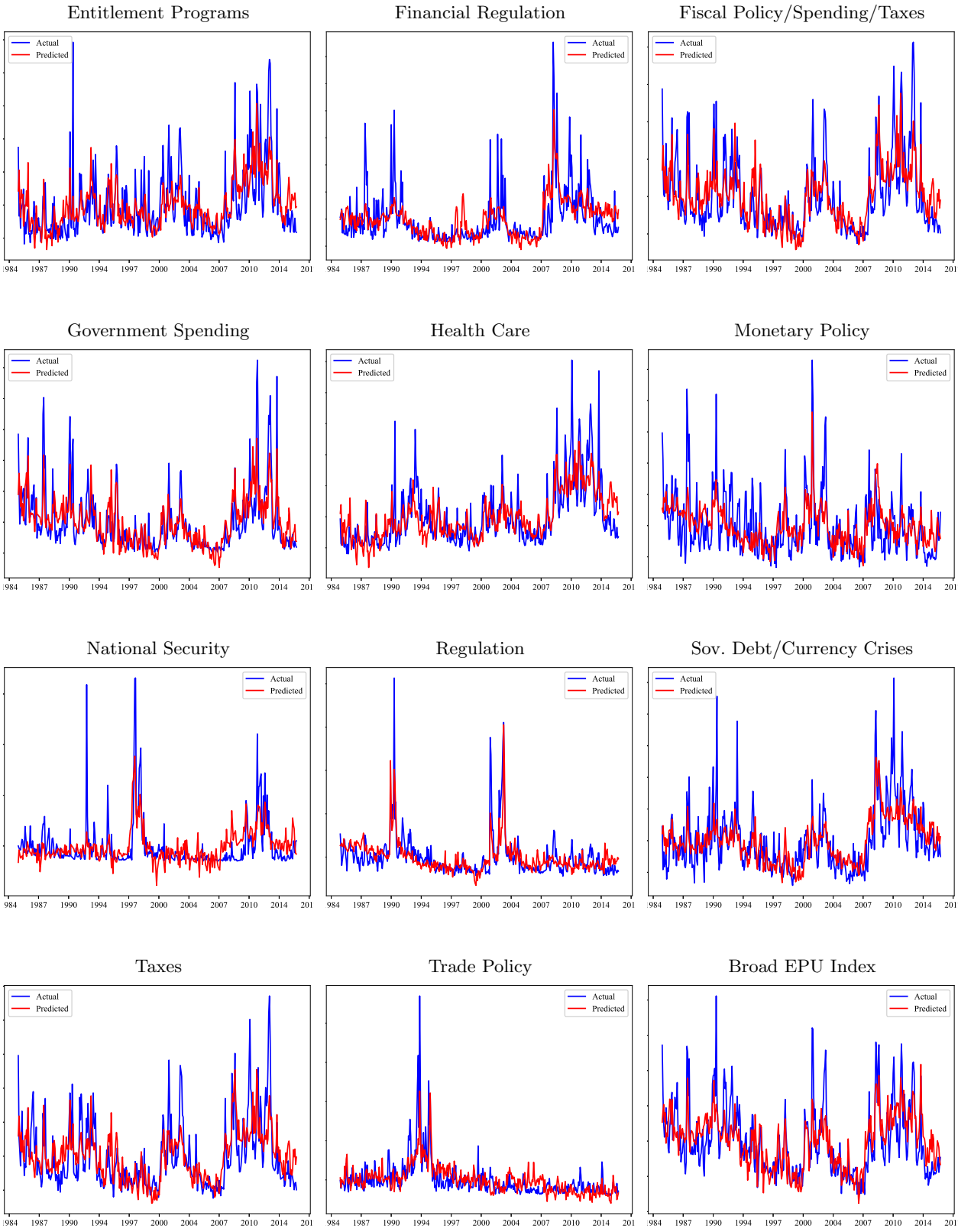
In contrast to BBD, our topic modeling approach is unsupervised and comparatively unrestricted. This also has its advantages and limitations. LDA lays out a general distributional framework for word counts that acts as a blank canvas for the data. This has the advantage of allowing the data to speak directly by projecting its underlying patterns onto this canvas (within the structure of the distributional assumptions). That is, our priors are less informative and therefore our approach is less biased. On the variance side of the tradeoff, our model has the limitation that, by modeling raw counts in their complex and high-dimensional form, it must wrestle with all varieties of noise in the data.

Our estimated news attention series share the great advantage of BBD's approach—the natural interpretability inherent in the use of text as data. While the BBD approach relies on careful researcher intervention to curate narrow search term lists, our results illustrate the exciting potential that topic models and other machine learning methods offer for constructing indices of economic conditions from news text with much less researcher intervention.

4.4.3 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

Figure 14: *WSJ* Topics and Economic Policy Uncertainty Indices



Note. Five-regressor lasso regression fits.

Table 6: 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.

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 6 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 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, viewing news media as a verbal mirror of the economy, we must also recognize that it comes with flawed refractions, often in the form of producer and consumer biases, erroneous inference, and unsubstantiated speculations (Gentzkow and Shapiro, 2006).

We estimate a topic model from the full text of *The Wall Street Journal*. Our estimates provide a complete taxonomy of its recurrent themes that describe 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 ac-

tivity, asset prices, and different measures of economic uncertainty. In the context of a standard macroeconomic VAR, we show that news topic attention contains substantial information about future economic outcomes above and beyond standard indicators.

Lastly, we propose a new econometric perspective on shock identification. 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 shock or fluctuation of interest. It thus isolates news-based narratives of those events that the researcher can "close-read" for detail and nuance about a given "shock" in numerical data.

References

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.

Blei, David, 2013, Probabilistic topic models, Lecture Slides, http://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf.

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

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.

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, How Did Paul Krugman Get It so Wrong?1, *Economic Affairs* 31, 36–40.

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* .

Fleckenstein, Matthias, and Francis A. Longstaff, 2018, Shadow funding costs: Measuring the cost of balance sheet constraints, Technical report, National Bureau of Economic Research.

Gennaioli, Nicola, and Andrei Shleifer, 2018, *A crisis of beliefs: Investor psychology and financial fragility* (Princeton University Press).

Gentzkow, Matthew, and Jesse M Shapiro, 2006, Media bias and reputation, *Journal of political Economy* 114, 280–316.

Griffiths, Thomas L, and Mark Steyvers, 2004, Finding scientific topics, *Proceedings of the National academy of Sciences* 101, 5228–5235.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic Interpretation of Emerging Risks in the Financial Sector, *The Review of Financial Studies* .

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.

Kelly, Bryan, Asaf Manela, and Alan Moreira, 2019, Text selection, *Working Paper* 14.

Krugman, Paul, 2009, How Did Economists Get It So Wrong?, *The New York Times* .

Liu, Jun, Francis A. Longstaff, and Ravit E. Mandell, 2006, The market price of risk in interest rate swaps: The roles of default and liquidity risks, *The Journal of Business* 79, 2337–2359.

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.

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.

Roll, Richard, 1988, Presidential address: R2, *Journal of Finance* 43, 51–566.

Shiller, Robert J., 2017, Narrative Economics, *American Economic Review* 107, 967–1004.

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

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.

A Construction 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. See the full list of excluded tags at www.structureofnews.com.
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). See the full list of excluded headline patterns at www.structureofnews.com.
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. See the full list of excluded stop words and URL patterns at www.structureofnews.com.
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 Key Terms

Table 7: Topic Key Terms

Metatopic label	Topic label	Key terms
Labor/transfers	Executive pay	*bonus, base salary, total compensation, pay package, compensation package, compensation committee, restrict stock, tyco, executive compensation, bonuses*
	Job cuts	*job cut, work force, fulltime employee, layoff, eliminate job, attrition, cut job, employment opportunity, employee work, unemployed*
	Unions	*uaw, teamsters, relation board, union member, national labor, labor relation, picket line, worker union, union leader, nonunion*
	Health insurance	*hmo, health plan, health coverage, health care, manage care, blue cross, health insurance, care act, blue shield, healthcare system*
	Pensions	*pension plan, employee retirement, benefit guaranty, pension, state pension, retirement system, retirement save, pension liability, retirement benefit, pension system*
	Government budgets	*spend cut, deficit reduction, grammrudman, federal budget, balance budget, budget plan, cbo, house budget, entitlement program, budget*
	Fees	*fee paid, pay fee, fee, charge fee, fee charge, higher fee, receive payment, license fee, company pay, upfront*
	Taxes	*irs, tax rate, deduction, income tax, tax bill, irss, capitalgains, capitalgains tax, tax court, taxable income*
Management	Connecticut	*stamford conn, stamford, greenwich conn, conn, haven conn, clark, greenwich, norwalk, norwalk conn, westport*
	C-suite	*executive officer, chief operate, operate officer, financial officer, chief financial, chairman president, officer john, officer robert, company chief, administrative officer*
	Mid-level executives	*vice, vice president, executive vice, senior vice, named senior, named vice, named executive, president finance, president market, group vice*
	Management changes	*succession plan, top job, departure, management shakeup, management style, shakeup, succession, heir apparent, top post, daytoday operation*
Transport/health/ defense	Natural disasters	*katrina, quake, hurricane katrina, fishermen, coast guard, storm, emergency management, tsunami, earthquake, boat*
	Police/crime	*inmates, police officer, violent crime, police department, gang, police, homicide, firearms, robbery, gunman*
	Mid-size cities	*kansas city, fla, lauderdale, palm beach, salt lake, lauderdale fla, lake city, fort lauderdale, beach fla, west palm*
	Disease	*disease control, aid virus, cdc, breast cancer, breast, infectious disease, heart disease, virus, flu, diagnosed*
	US defense	*rumsfeld, missile defense, marine corps, joint chief, donald rumsfeld, strategic defense, defense secretary, naval, senate arm, defense official*
	Pharma	*pfizer, generic, bristolmyers, lilly, generic version, drug company, generic drug, eli lilly, novartis, drug maker*
	Aerospace/defense	*force contract, lockheed, mcdonnell, mcdonnell douglas, northrop, pratt whitney, grumman, lockheed martin, airbus, general dynamic*
	NY politics	*blasio, cuomo, gov andrew, mayor bill, mayor michael, city council, andrew cuomo, city mayor, mayor, city hall*
	Automotive	*auto maker, domestic car, import car, ford, sportutility, chevrolet, chrysler, mazda, toyota, eestimated*
	Rail/trucking/shipping	*railroad, norfolk southern, conrail, csx, rail, union pacific, amtrak, burlington northern, norfolk, freight*
	Airlines	*airline, texas air, twa, american airline, airways, pilot association, pilot union, flight attendant, load factor, northwest airline*
	California	*los angeles, angeles, los, francisco, san francisco, diego, san diego, orange county, santa monica, angeles county*
	Rental properties	*square feet, square foot, vacancy rate, tenant, office space, landlord, office tower, squarefoot, retail space, highrise*
	Retail	*walmart, department store, store open, walmart store, penney, kmart, retail analyst, comparablestore, bentonville, bentonville ark*
Government	Political contributions	*donation, political contribution, philanthropy, campaign contribution, charity, nonprofit organization, nonprofit group, philanthropic, donor, donate*
	Regulation	*propose rule, rule govern, rule require, final rule, federal regulation, consumer protection, regulation, rule change, current rule, protect consumer*
	Environment	*epa, emission, protection agency, environmental protection, carbon dioxide, dioxide, greenhouse gases, clean air, environmental, environmental group*
	Private/public sector	*private sector, private enterprise, private business, public sector, private firm, private company, privatization, government aid, privatize, taxpayer money*
	State politics	*rhode island, state legislature, state include, west virginia, state legislator, schwarzenegger, north carolina, rhode, washington state, state lawmaker*
	Watchdogs	*inspector general, gao, account office, general account, accountability office, government accountability, page report, report conclude, inspector, review process*
	Utilities	*nrc, megawatt, nuclear regulatory, pge, nuclear plant, electricity, electric utility, public utility, utility commission, power company*
	Safety administrations	*safety board, traffic safety, highway traffic, safety administration, transportation safety, osha, national highway, occupational safety, health administration, accident*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	National security	*cybersecurity, security agency, hacker, intelligence committee, fbi, cia, espionage, homeland security, security expert, senate intelligence*
Legal	Justice Department	*spitzer, antitrust division, general eliot, receive subpoena, eliot spitzer, department antitrust, department investigation, justice department, internal investigation, general office*
	Indictments	*indictment, plead guilty, plead, plea agreement, mail fraud, face maximum, guilty plea, wire fraud, prosecutor allege, guilty*
	Courts	*supreme court, appeal court, scalia, federal appeal, chief justice, high court, justice william, william rehnquist, antonin, court rule*
	Lawsuits	*punitive damage, suit file, classaction, plaintiff lawyer, class action, suit allege, plaintiff attorney, lawsuit file, compensatory damage, classaction suit*
National politics	Clintons	*lewinsky, whitewater, rodham clinton, hillary rodham, rodham, clinton administration, president clinton, independent counsel, impeachment, janet reno*
	US Senate	*majority leader, senate bill, daschle, house bill, hastert, senate majority, appropriations committee, senate floor, senate version, gop leadership*
	Reagan	*contra, reagan administration, president reagan, shultz, george shultz, state george, regan, irancontra, james baker, secretary james*
	Bush/Obama/Trump	*obama administration, trump administration, president donald, president barack, bush administration, president dick, president bush, president joe, house press, barack*
	Elections	*sen obama, romney, sen mccain, dukakis, mitt romney, democratic primary, republican primary, mitt, gop primary, run mate*
	European politics	*opposition party, socialist party, ldp, parliamentary election, centerleft, parliamentary, rule party, centerright, rule coalition, main opposition*
Terrorism/ Middle east	Middle east	*palestinian, gaza, hamas, serb, gaza strip, israeli, plo, israel, hezbollah, palestinian authority*
	Nuclear/North Korea	*pyongyang, north korea, north korean, kim jong, nuclear program, daewoo, korean, seoul south, jong, kim*
	Terrorism	*taliban, karzai, afghan, qaeda, kabul, pakistani, islamabad, musharraf, hamid karzai, afghan government*
	Iraq	*iraqi, saddam, saddam hussein, kurds, baghdad, iraqi leader, iraqi government, erdogan, iraqi official, ankara*
International affairs	Russia	*yeltsin, moscow, putin, russian, kremlin, gorbachev, kiev, ukraine, russian president, boris yeltsin*
	Trade agreements	*gatt, wto, tariff, trade representative, antidumping, trade pact, protectionist, trade official, general agreement, freetrade agreement*
	Latin America	*argentina, chavez, latin, latin america, latin american, hugo chavez, cuban, colombia, president hugo, paulo brazil*
	Japan	*japanese company, tokyo japan, japan largest, japan biggest, shinzo, minister shinzo, abe, japanese business, strong yen, japan ministry*
	Canada/South Africa	*south africa, quebec, canadian government, toronto stock, british columbia, calgary, canadian cent, mandela, ottawa, calgary alberta*
	China	*yuan, beijing, xinhua, chinese official, chinese government, chinese company, chinese, zhu, beijing china, china sea*
	Southeast Asia	*jakarta, thailand, southeast asia, rupiah, southeast asian, suharto, indonesian, baht, indonesia, bhd*
	Germany	*east german, helmut kohl, bonn, chancellor helmut, helmut, german company, east germany, german official, west germany, chancellor gerhard*
	France/Italy	*generale, chirac, societe generale, french government, sarkozy, societe, french bank, belgium, francois, french company*
	UK	*gbp, scotland, hold plc, scotland group, plc british, british bank, british company, scottish, group plc, unit kingdom*
Entertainment	Music industry	*album, record label, song, record company, music industry, universal music, music group, music company, emi, musicians*
	Broadcasting	*clear channel, broadcaster, broadcast, cable channel, broadcast network, television station, cos nbc, nbc, viewership, liberty media*
	Publishing	*barnes noble, circulation, time mirror, random house, magazine publisher, daily news, gannett, newspaper publisher, daily newspaper, chicago tribune*
	Marketing	*saatchi, interpublic, omnicom, ddb, rubicam, saatchi saatchi, bbdo, wpp, young rubicam, publicis*
	Movie industry	*boxoffice, hollywood, paramount picture, motion picture, dreamworks, studio, sony picture, eisner, hollywood studio, mattel*
Social/cultural	Economic ideology	*letter letter, relate letter, hoover institution, moral, liberalism, enterprise institute, american enterprise, morality, hoover, franklin roosevelt*
	Schools	*public school, private school, high school, school, curriculum, test score, teacher union, school board, student, school district*
	Sales call	*conference room, salesman, salesmen, boss, phone call, hotline, colleague, receive call, desk, stunned*
	Cultural life	*dog, cowboy, deer, jacket, shirt, sweat, tshirt, cat, wear, shower*
	Arts	*sothebys, museum, auction house, art, collector, gallery, portrait, exhibit, thcentury, artistic*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Immigration	*immigrant, illegal immigrant, migrant, census, census bureau, immigration law, undocumented, immigration, citizenship, illegal immigration*
	Positive sentiment	*awful lot, pretty good, spend lot, good time, awful, good idea, making money, people feel, good deal, big money*
	Humor/language	*phrase, fool, word, humor, polite, translation, stupid, funny, emotion, tongue*
	Gender issues	*female, young women, mother, daughter, divorce, father, young children, sexual harassment, family member, wives*
Outlook language	Changes	*major shift, shift, fundamental change, big change, major change, significant change, rethink, rapid change, direction, change*
	Key role	*key role, play role, play key, play major, major role, important role, big role, play big, lead role, active role*
	Problems	*problem face, biggest problem, big problem, matter worse, major problem, real problem, bigger problem, similar problem, solve, fundamental problem*
	Challenges	*uphill battle, uphill, face tough, face pressure, battle, intense pressure, challenge face, status quo, quo, face challenge*
	Small possibility	*door open, time frame, frame, door, move quick, left open, wait, leave, ready, takes place*
	Spring/summer	*early march, late summer, close watch, late march, watch close, early sign, early stage, summer, watch, show sign*
	Long/short term	*long term, short term, long run, longer term, long time, term limit, short run, long period, long haul, long*
Science	Research	*survey conduct, respondents, survey show, survey, survey release, methodology, data compile, statistical, gallup, study release*
	Scenario analysis	*financial analysis, commentary, worstcase scenario, worstcase, analysis, foreseeable, foreseeable future, conventional wisdom, imply, scenario*
	Programs/initiatives	*program aim, goal, broader effort, top priority, pilot program, initiative, ambitious, achieve, program design, high priority*
	Biology/chemistry/physics	*gene, dna, mice, molecules, biology, scientist, physics, molecular, national laboratory, genetic*
	Space program	*nasa, space administration, national aeronautic, space station, space shuttle, spacecraft, astronauts, orbit, booster, mann*
	Systems	*task force, task, current system, centralized, decentralized, system work, complexity, efficient, inefficiency, workable*
	Size	*big chunk, chunk, large chunk, smaller, bigger, smaller company, rough double, rely heavily, size, rapid grow*
	Wide range	*wide range, vary wide, vast majority, broad range, wide variety, large number, limit number, large portion, vary, wide*
Activism	Activists	*icahn, carl icahn, investor carl, getty, pennzoil, carl, texaco, white plain, source close, holmes*
	Announce plan	*announce plan, plan announce, unveil plan, announce earlier, previous announce, yesterdays announcement, announce restructure, company announce, outline plan, latest move*
	Major concerns	*raise concern, major concern, express concern, concern, increase concern, voice concern, concern raise, cit, held, addition*
	Futures/indices	*spot index, dollar yen, brother treasury, commodity oil, oil barrel, treasury index, future index, bond lehman, stock volume, jonesaig future*
	Corrections/amplifications	*amplifications, correction amplifications, article correction, incorrect, news article, incorrect state, relate article, incorrect report, article, correction*
	Buffett	*buffett, berkshire, hathaway, berkshire hathaway, warren buffett, warren, omaha neb, omaha, neb, billionaire investor*
	Mexico	*mexican, salinas, mexican government, grupo, mexico, mexico city, ernesto, lopez, mexican peso, telefonos*
Announcements	Restraint	*temporary restrain, restrain order, postponement, postpone, delay, extension, indefinite, temporarily, expiration date, restrain*
	News conference	*news conference, told news, press conference, told reporter, conference, press release, news release, remark, attendees, public appearance*
	Company spokesperson	*call seek, seek comment, return call, spokesman confirm, return phone, spokesman decline, spokesman add, return telephone, company spokesman, comment spokesman*
	People familiar	*people familiar, familiar, person familiar, person close, people close, people involve, person, matter, individual familiar, executive familiar*
	Agreement reached	*definitive agreement, sign letter, sign definitive, definitive, reach definitive, agreement call, lone star, preliminary agreement, agreement include, agreement reach*
	Committees	*panel member, advisory committee, advisory panel, recommendation, proposal include, proposal call, advisory group, final decision, panel, vote unanimous*
	Negotiations	*stick point, negotiator, negotiate table, side agree, negotiation, impasse, reach deal, reach compromise, resolve dispute, unresolved*
Industries	Soft drinks	*coke, bottler, pepsi, softdrink, cola, cocacola, soft drink, pepsico, north america, beverage company*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Small business	*smallbusiness, small business, smallbusiness owner, business owner, national federation, independent business, small firm, small company, business administration, mediumsize business*
	Cable	*warner cable, time warner, aol time, warner, parsons, warner communication, levin, adelphia, aol, warner incs*
	Fast food	*franchisee, taco bell, burger king, fastfood, rite aid, starbucks, pizza hut, mcdonald, taco, fastfood chain*
	Competition	*goodyear, stiff competition, tire rubber, increase competition, fierce competition, goodyear tire, competitive advantage, competitive market, increase competitive, high competitive*
	Chemicals/paper	*union carbide, carbide, specialty chemical, georgiapacific, chemical company, pont, weyerhaeuser, chemical concern, chemical business, dupont*
	Venture capital	*venture, joint venture, venture capitalist, venture capital, venturecapital firm, venturecapital, form joint, venture partner, venture fund, venturecapital fund*
	Tobacco	*tobacco, smokers, tobacco company, cigarette, philip morris, nicotine, tobacco industry, morris, cigarette maker, reynolds tobacco*
	Subsidiaries	*owned subsidiary, wholly owned, spinoff, itt, parent company, unit include, separate company, wholly, noncore, business unit*
	Credit cards	*mastercard, credit card, debit card, debit, creditcard company, card, creditcard issuer, american express, creditcard, expresss*
	Couriers	*postal service, fedex, postal, unit parcel, ups, parcel service, post office, federal express, parcel, mail*
	Foods/consumer goods	*sara lee, general mill, nestle, hershey, avon, campbell soup, diapers, detergent, cereal, toothpaste*
	Insurance	*policyholders, life insurance, reinsurance, insurance department, insurance regulator, allstate, insurance commissioner, mutual life, life insurer, insurance business*
	Luxury/beverages	*lvmh, vuitton, hennessy louis, louis vuitton, moet, moet hennessy, lvmh moet, beer, brewer, wine*
	Casinos	*casino, atlantic city, marriott, hilton, carnival, hilton hotel, hotel chain, hotel room, inn, caesars*
	Revenue growth	*growth strategy, growth opportunity, annual revenue, boost revenue, growth area, company revenue, total revenue, company growth, revenue growth, growth potential*
Technology	Internet	*facebook, web site, search engine, alibaba, yahoo, ebay, youtube, home page, internet user, online advertise*
	Mobile devices	*iphone, smartphone, android, handset, ipad, blackberry, mobile device, nokia, apple incs, mobile phone*
	Electronics	*playstation, nintendo, console, kodak, digital camera, lcd, flatpanel, xbox, cdrom, electronic art*
	Phone companies	*att, sprint, longdistance, local phone, bell atlantic, phone company, baby bell, bell company, mci, longdistance company*
	Computers	*ibm, compaq, mainframe, dell, mainframe computer, machine corps, armonk, computer giant, packard, computer business*
	Software	*linux, window operate, ballmer, microsoft window, redmond wash, redmond, microsoft, oracle, software giant, lotus*
	Microchips	*amd, advance micro, memory chip, santa clara, semiconductor industry, clara, intel chip, intel, clara calif, chip maker*
Oil/mining	Steel	*steel price, steel industry, steelmaker, alcan, steel product, nucor, steel producer, steel, steel mill, steel company*
	Mining	*bhp, rio tinto, tinto, newmont, bhp billiton, mining company, miner, mining, billiton, newmont mining*
	Machinery	*machine tool, manufacture plant, whirlpool, manufacture facility, construction equipment, manufacture operation, manufacture capacity, appliance maker, production facility, cummins*
	Oil drilling	*exxon, drill, cubic feet, mobil, exxon mobil, drill rig, offshore oil, chevron, amoco, royal dutchshell*
	Agriculture	*genetically modified, farm, growers, poultry, madcow, farmer, ranchers, tyson food, agriculture secretary, tyson*
	Oil market	*opec, nonopec, opec member, oil minister, oil demand, world oil, brent crude, oil market, oil exporter, export country*
Economic growth	Record high	*record level, highest level, reach record, hit record, previous record, highest, pentup demand, high level, pentup, alltime*
	Economic growth	*gdp growth, productivity growth, gdp, domestic product, gross domestic, economy grew, national product, economy expand, gross national, growth forecast*
	Federal Reserve	*greenspan, alan greenspan, yellen, federalfunds rate, federalfunds, fomc, chairman alan, raise rate, janet yellen, inflation expectation*
	European sovereign debt	*ecb, imf, eurozone, euro zone, monetary fund, international monetary, draghi, eurozone country, trichet, ecb president*
	Recession	*economic downturn, steep drop, steep decline, downturn, bright spot, hardest hit, hit hard, weak economy, hit bottom, hard hit*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
	Product prices	*raise price, increase price, price reduction, price increase, price cut, wholesale price, higher cost, cut price, retail price, reduce price*
	Optimism	*cautious optimistic, remain optimistic, remain cautious, remain uncertain, uncertainty surrounding, positive sign, optimistic, overly optimistic, cautious, remain confident*
	Macroeconomic data	*seasonal variation, durable good, construction spend, seasonal adjust, nonresidential, business inventory, nondurable, durable, conference board, jobless claim*
Corporate earnings	Profits	*net profit, profit rose, profit fell, firsthalf net, profit increase, profit jump, operate profit, profit climb, profit surge, pretax profit*
	Revised estimate	*previous estimate, original estimate, preliminary estimate, revision, earlier estimate, revise downward, downward revision, initial estimate, revise, substantial increase*
	Earnings losses	*loss include, net loss, discontinue operation, operate loss, loss widen, pretax charge, discontinue, aftertax charge, post loss, restructure charge*
	Small changes	*typical require, average yield, decline slight, decrease, slight increase, remain unchanged, remain flat, sharp increase, part reflect, slight decline*
	Financial reports	*ended sept, forma, pro forma, ended march, latest fiscal, march compare, current fiscal, full fiscal, ended, company fiscal*
	Earnings forecasts	*adjust earn, analyst earn, analyst poll, forecast earn, earn forecast, earningspershare, earn expectation, beat expectation, earn guidance, earn estimate*
	Earnings	*cent dilute, dilute share, share revenue, share relate, callthomson financial, cent share, callthomson, call consensus, pershare figure, thomson financialfirst*
Markets	IPOs	*ipo, ipo market, initial public, public offer, ipo price, firstday, public stock, stock offer, roadshow, offer price*
	Bond yields	*yield rise, highyield bond, sell bond, yield fall, corporatebond, bond issue, treasury yield, issue bond, municipalbond, bond sale*
	Short sales	*short seller, short interest, short sell, borrow stock, borrow share, shortselling, shortsellers, short sale, sell borrow, seller*
	Small caps	*york smallcapitalization, poor smallcap, smallcapitalization stock, russell index, nyse rose, smallcap index, smallcap market, nyse fell, nasdaq gain, small-capitalization*
	Treasury bonds	*accept noncompetly, cusip, bid widen, price accept, noncompetitive tender, coupon equivalent, application accept, price rate, cusip number, bill mature*
	International exchanges	*dax stock, perspective geneva, australia world, switzerland australia, germany hong, firstsection, firstsection volume, international perspective, kong switzer-land, index calculate*
	Exchanges/composites	*composite trade, exchange composite, exchange trade, market composite, york stock, stock exchange, adrs, share close, board trade, composite*
	Currencies/metals	*dollar french, home mkt, frankfurt gold, dollar british, london toronto, comex base, tokyo zurich, comex london, yen swiss, ounce comex*
	Share payouts	*nyse symbol, share buyback, stock buyback, dividend increase, buyback, buyback program, share repurchase, share common, pay dividend, time estimate*
	Bear/bull market	*bear market, bull market, stock pick, stockpicking, priceearnings ratio, growth stock, pricetoearnings ratio, bull, priceearnings, pricetoearnings*
	Commodities	*cocoa exchange, comex division, sugar cocoa, coffee sugar, soybean future, coffee future, lifeofcontract, pork belly, wheat future, york cotton*
	Trading activity	*action stock, fell bond, market action, otc trade, volume total, investor report, issue advance, board volume, index exclude, transportation average*
	Options/VIX	*option exchange, volatility index, board option, contract compare, put trade, option strategist, option market, vix, cboe, option trader*
Buyouts	M&A	*combine company, merge company, propose merger, combine entity, merger plan, merger talk, planned merger, merger, planned acquisition, antitrust concern*
	Control stakes	*control stake, minority stake, majority stake, buy stake, abn, sell stake, remain stake, bought stake, minority shareholder, owns stake*
	Drexel	*drexel, leverage buyout, kohlberg kravis, drexel burnham, transaction finance, kohlberg, kravis robert, kravis, boesky, pickens*
	SEC	*sec file, sec, sec allege, sec official, commission file, sec enforcement, sec chairman, exchange commission, sec staff, sec investigation*
	Bankruptcy	*bankruptcy, unsecured creditor, bankruptcy court, federal bankruptcy, bankruptcy code, chapter company, creditor lawsuit, bankruptcy judge, reorganization plan, bankruptcy protection*
	Corporate governance	*annual meet, shareholder meet, proxy contest, corporate governance, independent director, dissident shareholder, board seat, proxy battle, annual shareholder, proxy fight*
	Takeovers	*poison pill, hostile takeover, takeover bid, hostile bid, share offer, hostile tender, share tender, higher offer, takeover battle, takeover attempt*
	Real estate	*real estate, reit, estate investment, realestate investment, estate market, estate, estate developer, estate development, commercial real, realty trust*
	Convertible/preferred	*common share, preferred share, convertible preferred, common stock, class common, convertible debenture, company common, national overthecounter, convertible subordinate, preferred*

Continued on next page

Table 7 – *Continued from previous page*

Metatopic label	Topic label	Key terms
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*
	Accounting	*andersen, account firm, kpmg, arthur andersen, fasb, deloitte, ernst young, audit firm, ernst, skadden*
	Investment banking	*csfb, sachs group, goldman, suisse group, sachs, goldman sachs, credit suisse, suisse, ubs, morgan*
	Acquired investment banks	*salomon smith, smith barney, barney, merrill, bear stearns, stearns, merrill lynch, lynch, stearns cos, painewebber group*
	Private equity/hedge funds	*hedge fund, privateequity, privateequity firm, hedgefund, blackstone, carlyle, privateequity fund, hedgefund firm, hedgefund manager, blackstone group*
	NASD	*nasd, customer account, security transaction, member firm, barred, disciplinary action, admit, disciplinary, misappropriated, deny*
Banks	Savings & loans	*fdic, thrift, loan association, federal deposit, federal save, deposit insurance, bank board, save loan, thrift industry, loan insurance*
	Nonperforming loans	*wells fargo, fargo, loanloss provision, bruyette, bankamerica, keefe bruyette, bruyette wood, nonperforming asset, loanloss, noninterest*
	Credit ratings	*debt rate, fitch rate, rate firm, rate service, creditwatch, rate agency, creditrating, taxes depreciation, junk status, notches*
	Financial crisis	*stress test, relief program, large financial, trouble asset, financial system, creditdefault swap, financial regulator, creditdefault, financial firm, financial crisis*
	Bank loans	*syndicate loan, revolving credit, credit line, libor, credit agreement, loan agreement, revolving, bank loan, creditor bank, rescheduling*
	Mortgages	*fixedrate mortgage, home price, home buyer, adjustablerate mortgage, mortgage rate, countrywide, federal house, mortgage company, mortgage lender, mortgage payment*

Figure 15: Topic Attention Proportions