Breaking the Word Bank:
Effects of Verbal Uncertainty on Bank Behavior

Paul E. Soto*

February 13, 2019

Abstract

Banks differ from non-financial firms as banks must communicate to both regulators and shareholders. Also, unlike non-financial firms, banks possess opaque and complex balance sheets and are the main providers of credit to the real economy. In this paper, I propose a new index to detect the idiosyncratic uncertainty banks face at the bank-quarter level by applying natural language processing techniques to earnings conference call transcripts. The index reveals which banks at a given quarter signal more uncertainty about their balance sheets. Higher uncertainty is associated with lower lending and higher trading the next quarter, suggesting active management of uncertainty. The active management of uncertainty is more pronounced during periods of high aggregate volatility and for banks with more skin in the game. Using loan level data and firm fixed effects, I control for demand-side factors and find that higher bank level uncertainty is associated with lower loan issuances the following quarter.

JEL Codes: G21, G3, E5, D8

Keywords: Uncertainty, Banking, Credit, Natural Language Processing, Machine Learning

*Federal Deposit Insurance Corporation (FDIC), Washington, D.C. E-mail: paul.soto@upf.edu. I am greatly indebted to my PhD advisers, José-Luis Peydró and Andrea Polo, for extraordinary research guidance. I would like to thank Nicholas Bloom, Stephen Hansen, Michael McMahon, Xavier Freixas, Christian Fons-Rosen, Puriya Abbassi, Filippo Ippolito, Manuel Garcia Santana, Rosalind Bennett, Christopher Kurz, Haelim Anderson, Christoph Kleiner, Claire Brennecke, Christos Makridis, Juan Imbet, Nick Garvin, Dmitriy Khametshin, Angelo Gutiérrez Daza, and participants at the seminars at the Universitat Pompeu Fabra, the Deutsche Bundesbank, and the Modelling with Big Data and Machine Learning conference for useful comments and suggestions. Special thanks to Kathy Zeidler for editorial assistance. Opinions expressed in this paper are those of the author and not necessarily those of the FDIC.
1 Introduction

The 2007 financial crisis introduced a period of prolonged uncertainty into the economy. Market participants were unsure which policies central banks and governments would introduce and how declines in asset prices would affect investment and productivity. They were also uncertain about the financial condition of banks entangled in disrupted markets. Banks play a unique and vital role in the real economy through credit provisioning (Kashyap et al. 2002), yet their fragility also can adversely affect the real economy (Bernanke et al. 1999). Evidence of how banks respond to uncertainty has been elusive given banks’ opaque and complex balance sheets (Morgan 2002, Mehran et al. 2011)\(^1\) and their tendency to window-dress assets under supervision (Abbassi et al. 2018).

Further, designing policies in the face of uncertainty remains particularly difficult for policy makers because uncertainty itself is not easy to measure. Current research on uncertainty largely focuses on aggregate measures. However, time-series measures of uncertainty are difficult to disentangle from the business cycle. A more granular bank level measure of uncertainty would allow researchers to understand which aggregate shocks affect bank level uncertainty and, more importantly, how banks respond to uncertainty while controlling for aggregate trends.

In this paper, I introduce a new method of measuring bank level uncertainty and illustrate how banks manage uncertainty through changes in balance sheet composition. I find that higher idiosyncratic uncertainty is associated with lower lending and higher trading, relative to total assets, the following quarter. The new measure is constructed by applying novel techniques in natural language processing and machine learning to earnings conference calls transcripts from 2002 to 2017. This paper complements extant aggregate uncertainty measures which are typically time-series with a measure at the bank level to allow for cross-sectional heterogeneity.

Creating the new measure involves two steps: (1) creating a list of *uncertainty* words and (2) counting the frequency of the *uncertainty* words in the transcripts. Many economic and finance papers using word counts involve either the researcher choosing a list of words they believe capture the topic they want to measure, or using a list of words identified from another paper. Rather than

---

\(^1\)Healthcare and pharmaceutical companies face similar issues through required communication between regulators and shareholders. Healy et al. 2002 and David et al. 2010 discuss the interaction between regulators and pharmaceutical firms.
relying on an existing dictionary that may not be suited for banks, I develop a new word list using a machine learning technique known as word embeddings. Word embeddings are numeric vector representations of words. In other words, every unique word in a set of documents has its own vector representation. The position of each word in the vector space is based on analyzing the contextual information of each unique word across documents, as similar words often appear in similar contexts. This methodology maps similar words of uncertainty into the same region of the vector space. Word embeddings provide researchers an objective framework to determine words of uncertainty based on semantic and syntactic similarities, rather than generating a list of words subjectively or with an existing lexicon from another paper. Figure 1 illustrates the two-dimensional projection of the word embeddings from the bank conference call transcripts in my sample. The rightmost region of the map shows words such as “anxiety,” “war,” and “fears” alongside “uncertainty,” suggesting semantically similar meanings for these words within the context of bank speech. I create a new uncertainty word list by clustering the words in this region.

I calculate bank level uncertainty as the frequency of these words in each quarterly conference call. Because the usage of uncertainty words will be related to aggregate conditions, such as the growth in volatility, economic policy uncertainty, and the term structure of corporate bond yields,
I filter out these variables from the frequency counts. The residual is the new measure I propose in this paper and represents the idiosyncratic uncertainty banks face at a given quarter.

Figure 2: Bank Level Uncertainty

\[ \text{BankUncert}_{b,t} \]

Note: This figure plots the time series median of the bank uncertainty variable, \( \text{BankUncert}_{b,t} \), I construct in this paper. \( \text{BankUncert}_{b,t} \) is the count of uncertainty terms from management responses of bank quarterly earnings conference calls after filtering out aggregate uncertainty measures, calculated at the bank-quarter level. The dashed blue lines represent the 25th and 75th percentiles.

Three results help validate the new measure as a proxy for uncertainty. First, the usage of uncertainty words is correlated to aggregate uncertainty measures of the economy. Though the measure is not associated with policy uncertainty, as measured by the Economic Policy Uncertainty (EPU) index from Baker et al. 2016,\(^2\) the frequency of uncertainty words in the conference calls is correlated with the Chicago Board Options Exchange (CBOE) Volatility Index (VIX) and changes in the term structure of corporate bond yields. This correlation is reassuring because the slope of the yield curve provides a measure of uncertainty about the economy (Collin-Dufresn et al. 2001). Because bank assets are predominantly made up of credit products, the fact that the usage of uncertainty words by banks moves hand in hand with an index pertaining to credit uncertainty

---

\(^2\)Baker et al. 2016 use a text-based approach to measure economic policy uncertainty by counting the number of newspaper articles mentioning the words “uncertainty” and “uncertain” in certain contexts and link their index to unemployment and firm-level investment.
is encouraging. Second, the time-series average of bank uncertainty lines up with events hampering the banking sector during the last business cycle. As seen in Figure 2, the index begins to rise at the onset of the recent financial crisis, in December 2007. Bank level uncertainty remained high following the collapse of Lehman Brothers in September 2008, peaking shortly after as regulators and banks sought solutions to the ongoing crisis. Toward the end of the crisis, uncertainty plummets. At the same time, cross-sectionally, banks differed largely in their assessment of idiosyncratic uncertainty, as shown by the dashed blue lines representing the 25th and 75th percentiles. Third, bank uncertainty is positively associated with post-call volatility but not post-call returns, suggesting that uncertainty through speech relays some form of risk and new information to market participants.

Encouraged by the evidence that the new measure is a proxy for uncertainty, I next assess how banks actively manage their balance sheets following changes in idiosyncratic uncertainty. While extant uncertainty measures are restricted to only the time dimension, the new measure allows me to exploit the cross-sectional heterogeneity of bank uncertainty at a given quarter. In particular, my measure allows me to include time fixed effects to control for interest rates, aggregate firm productivity, the stochastic discount factor, and other time varying characteristics related to credit and trading. I find that a one standard deviation increase in a bank’s uncertainty is associated with a 40 basis point drop in lending relative to total assets the following quarter. The drop in lending from idiosyncratic uncertainty is not due to time-varying aggregate conditions nor several relevant controls from bank balance sheets. Indeed, while banks cutback on credit after speaking about higher uncertainty, their balance sheets the following quarter report larger trading assets.

Next, I allow the coefficient of bank uncertainty to vary depending on bank leverage and the growth in aggregate volatility. Higher equity funding is associated with larger skin in the game, either on the part of shareholders or internally from bank executives and employees. Board members monitor and threaten bad and incompetent executives to incentivize diligent and conscientious actions. I find that declines in credit and increased trading assets from higher idiosyncratic uncertainty are more severe for banks with more skin in the game, suggesting bank CEOs more actively manage uncertainty under higher shareholder market discipline. Similarly, during times of high aggregate uncertainty, banks that speak using more uncertainty words tend not only to reduce credit but also to increase trading portfolios.
I then decompose the transcripts to identify topic-specific bank uncertainty. Using a technique in textual analysis and machine learning known as topic modeling, I partition the earnings conference call transcripts into two topics particularly relevant to banks: interest rates and housing. I remeasure bank level uncertainty using only the sections of the earnings call related to those topics. I find that uncertainty about interest rates leads to higher exposures of interest rate derivatives, while uncertainty about real estate is significantly associated with lower levels of real estate lending. The results are reassuring not only because they capture topic-specific uncertainty according to the established methodology, but also because they show how specific asset class exposures are partly driven by active management of uncertainty.

Finally, I provide evidence of the contraction in credit from bank level uncertainty at the loan level. A criticism of analyses done at the bank level is that firm-specific demand may drive uncertainty at banks and lower demand for credit by firms, resulting in the observed credit cutbacks. To mitigate these concerns, I use new issuances of loans at the bank-firm-quarter level and include firm-time fixed effects to control for observed and unobserved time varying firm-specific characteristics, especially firm demand. The identifying assumption requires one firm at a particular quarter to receive bank credit from two different banks. I find that banks which relay higher idiosyncratic uncertainty provide smaller-sized loans to firms, consistent with the results at the bank level. The loan level data provide further evidence that banks actively manage their uncertainty by reducing lending even after controlling for credit demand at the firm level.

This paper relates to several strands of literature. First, this paper adds to the growing literature on measuring uncertainty. Most similar to this paper is Baker et al. 2016 who develop a text-based economic policy uncertainty index by counting the frequency of newspaper articles with the words "uncertainty" and "uncertain" within certain contexts. Their measure spikes during tight presidential races, military conflicts, and the 2008 failure of Lehman Brothers. The economic policy uncertainty index is associated with declines in firm investments and output. One important distinction with Baker et al.’s measure and the index presented in this paper is that the economic policy uncertainty index is a time series and when used as a sole regressor can not be joined with time fixed effects. This paper looks at the cross-section of uncertainty facing banks, allowing me to accurately control for aggregate uncertainty. Berger et al. 2018 show that banks respond to higher economic policy uncertainty, as measured by the time series in Baker et al., by hoarding liquidity.
and reducing the supply of credit, which coincide with declines in credit caused by idiosyncratic uncertainty observed in this paper. 

Jurado et al. 2015 develop another measure of uncertainty that suggests uncertainty episodes, such as those found in Baker et al. 2016, appear more infrequently but persistently. Baley and Blanco 2019 develop a firm-level uncertainty index caused by a firm’s inability to disentangle temporary and permanent uncertainty shocks. Ludvigson et al. 2015 show that higher macro uncertainty is often an endogeneous response to the business cycle, while financial uncertainty is found to cause declines in real activity. Similarly, Berger et al. 2017 find that forward-looking measures of uncertainty have no real effects while measures derived from realized stock market volatility lead to contractions.

Second, this paper contributes to several studies analyzing the informativeness of earnings conference calls. Mayew 2008 and Bowen et al. 2002 show how new and valuable information is made public through these conference calls. Mayew and Venkatachalam 2012 apply voice recognition software to earnings calls to study the effect of vocal cues and emotions on stock recommendations. The informativeness of these calls affects post-call returns of public companies (Roychowdhury and Sletten 2012, Price et al. 2012). Davis et al. 2015 show that the manner in which individual managers speak affects investors’ interpretation of disclosures made in conference calls.

Last, I provide applications of natural language processing and machine learning techniques to economics and finance. Loughran and McDonald 2011 illustrate the nuance of economic and financial texts compared with text from other social sciences. Text analysis has been used to analyze central bank transparency (Hansen et al. 2018), regulatory uncertainty using public comments (Gissler et al. 2016), and stock market reactions to conference call transcripts (Demers and Vega 2014, Hassan et al. 2016). Hassan et al. 2016 measure political risk in earnings conference calls transcripts and assess the impact on volatility, investment, and political donations. Kozlowski et al. 2018 use word embeddings, a technique applied in this paper, to show how gender and class co-evolved over the last century.

The remainder of this paper is organized as follows. Section 2 provides a summary of the data sources. Section 3 describes the text-based methodology I use to extract idiosyncratic bank uncertainty from the conference calls. Section 4 provides predictions to guide the empirical results. Section 5 illustrates how banks actively manage uncertainty. Section 6 concludes the paper.
2 Data

2.1 Text Data from Conference Calls

In 2000, the U.S. Securities and Exchange Commission (SEC) implemented Regulation Fair Disclosure requiring public companies to disclose material information pertaining to business activities. The SEC encouraged public companies to post written transcripts of calls and webcasts on their websites for an appropriate period of time each quarter. The National Investor Relations Institute reports that the percentage of companies holding earnings calls increased from 80 percent in the mid-90s to 97 percent in 2014. The calls last anywhere between 30 minutes to 60 minutes, depending on the market capitalization of the company, current events, and analyst coverage. Earnings conference calls typically begin with the CEO and the CFO providing a summary of the recent quarter and what to expect the following quarter. Next, the line opens for questions from analysts and sometimes shareholders. Conference calls represent a unique opportunity to analyze information executives would like to share with market participants along with information that investors demand, providing a setting where new and valuable information is made public (Mayew 2008).

Earnings call transcripts were downloaded from Thomson Reuters. Since implementation of Regulation Fair Disclosure in 2002, most earnings call transcripts are made public around the date of the call. The calls for banks in my database are typically only available from 2002 onward. I downloaded the available transcripts for the banks in my sample from first quarter 2002 to third quarter 2017.

2.2 Financial Data

For blue balance sheet data I use the Federal Reserve Y-9C forms, which all banks with more than 1 billion dollars in consolidated assets are required to file with the Federal Reserve each quarter. Data are at the bank holding company level. I retrieve the following balance sheet characteristics: total loans ($BHCK2122$), trading assets ($BHCK1773+BHCK1754$), liquidity ($BHCK0081+BHCK0395+BHCK0397$), loans secured by real estate ($BHCK1410$), interest rate exposure ($BHCK8757$), non-interest income ($BHCK4079$), equity ($BHCP3210$), non-performing loans ($BHCK5525+BHCK5526$), tier-1 capital ($BHCK7205$), and total assets ($BHCK2170$). Table A1 in the appendix provides a

\footnote{This value was later changed to BHCA7205 in 2014. Following 2014, I use the latter definition.}
summary of the variables obtained from the Y-9C forms. All balance sheets variables, except for total assets and non-interest income, are calculated relative to total assets. Non-interest income is measured relative to total income, and total assets are transformed to the log of total assets.

The measures used for aggregate uncertainty are the Economic Policy Uncertainty dataset from Baker et al. 2016, the CBOE volatility index VIX, and corporate bond yields from the U.S. Department of the Treasury.

I use Dealscan data for the loan level analysis. The data contains new issuances of large, syndicated, commercial loans in the United States from 2002-2013 at a quarterly frequency. Saturating the loan level regressions with firm-quarter fixed effects reduces the cross-section of banks in the sample to 9 banks and 329 firms. The average size of these banks is larger, nearly $890 billion in total assets, with an equity ratio of 8.7% on average. These banks represent roughly 14% of total banking assets. The average size of the loans is $1.36 billion and the median is $750 million. The bank participation share is $90 million on average.

### 2.3 Final Dataset

I combine the FR Y-9C database with the earnings conference calls manually by matching observations using the names of the institutions. I use the RSSD9017 identifier from the call reports to merge with the name of the bank from the earnings call. In the event of a merger, I keep only the bank holding company and drop the acquired institution from the sample starting from the quarter of the merger. The resulting dataset is joined to the aggregate uncertainty variables. The sample covers the years 2002 through 2017. I restrict the sample to banks with at least 75 percent of the conference calls available during those years. The threshold was chosen so as to maintain a high cross-section of banks, but results are robust if the sample is restricted to only those banks for which data are available for every quarter.

Table 1 provides the summary statistics. In total, 56 banks were analyzed from first quarter 2002 to third quarter 2017 with 2,732 bank-quarter level observations. Total assets for all commercial banks in the United States averaged $9.5 trillion, at the industry level per quarter, during this time period. Total assets for the 56 banks in the sample averaged $8.4 trillion, representing roughly 88 percent of total industry assets. The average asset size of the banks is roughly $29 billion with an

---

4This dataset can be accessed at www.policyuncertainty.com.
### Table 1: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>P25</th>
<th>Median</th>
<th>P75</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loans</td>
<td>62.73</td>
<td>16.93</td>
<td>4.65</td>
<td>59.16</td>
<td>67.57</td>
<td>73.36</td>
<td>96.21</td>
</tr>
<tr>
<td>Trading Assets</td>
<td>19.85</td>
<td>10.3</td>
<td>0.1</td>
<td>13.05</td>
<td>17.42</td>
<td>24.01</td>
<td>61.75</td>
</tr>
<tr>
<td><strong>Independent Variables and Controls</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_B$</td>
<td>0.46</td>
<td>0.37</td>
<td>0.00</td>
<td>0.22</td>
<td>0.38</td>
<td>0.62</td>
<td>4.01</td>
</tr>
<tr>
<td>EPU</td>
<td>99.41</td>
<td>45.09</td>
<td>37.27</td>
<td>64.79</td>
<td>85.41</td>
<td>132.82</td>
<td>217.31</td>
</tr>
<tr>
<td>Corp. Bond Slope</td>
<td>1.56</td>
<td>0.83</td>
<td>-0.15</td>
<td>1.02</td>
<td>1.58</td>
<td>2.27</td>
<td>2.91</td>
</tr>
<tr>
<td>Log(Assets)</td>
<td>17.19</td>
<td>1.88</td>
<td>12.93</td>
<td>15.84</td>
<td>16.82</td>
<td>18.54</td>
<td>21.67</td>
</tr>
<tr>
<td>Equity</td>
<td>10.34</td>
<td>6.66</td>
<td>0.2</td>
<td>2.14</td>
<td>3.14</td>
<td>5.33</td>
<td>41.24</td>
</tr>
<tr>
<td>Profitability</td>
<td>29.29</td>
<td>15.72</td>
<td>12.57</td>
<td>19.42</td>
<td>27.34</td>
<td>36.57</td>
<td>84.14</td>
</tr>
<tr>
<td>Liquidity</td>
<td>5.41</td>
<td>0.00</td>
<td>0.2</td>
<td>2.14</td>
<td>3.14</td>
<td>5.33</td>
<td>41.24</td>
</tr>
<tr>
<td>Tier 1 Ratio</td>
<td>14.00</td>
<td>3.36</td>
<td>8.52</td>
<td>12.3</td>
<td>13.66</td>
<td>15.35</td>
<td>24.32</td>
</tr>
<tr>
<td>Non-Performing Loans</td>
<td>1.2</td>
<td>1.33</td>
<td>0.37</td>
<td>0.73</td>
<td>1.58</td>
<td>14.46</td>
<td></td>
</tr>
<tr>
<td><strong>Dealscan Sample</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facility Amount ($MM)</td>
<td>1.41</td>
<td>2.34</td>
<td>0.00</td>
<td>0.29</td>
<td>0.75</td>
<td>1.50</td>
<td>36.5</td>
</tr>
<tr>
<td>Bank Amount ($MM)</td>
<td>0.09</td>
<td>0.17</td>
<td>0.00</td>
<td>0.02</td>
<td>0.04</td>
<td>0.09</td>
<td>2.86</td>
</tr>
<tr>
<td>Equity</td>
<td>8.74</td>
<td>1.85</td>
<td>5.19</td>
<td>7.67</td>
<td>8.37</td>
<td>9.77</td>
<td>13.31</td>
</tr>
</tbody>
</table>

**Note:** This table reports the summary statistics of the variables used throughout the paper. Balance sheet variables are retrieved from the FR Y-9C forms from the Federal Reserve Board. The Loans variable is total quarterly commercial and industrial, agricultural, consumer, and foreign loans over total assets. Trading is the sum of available-for-sale and hold-until-maturity portfolios over total assets. $\sigma_B$ is the count of uncertainty terms of a bank’s quarterly earnings conference call. VIX is the CBOE VIX index. EPU is the Economic Policy Uncertainty index from Baker, Bloom and Davis 2016. Corp. Bond Slope is the difference in the ten year less two year corporate bond rate. The Log(Assets) variable is the logarithm of total assets measured in thousands. Equity is total equity capital over total assets. Profitability is the non-interest net income over total net income. Liquidity is the interest and non-interest bearing balances over total assets. Tier-1 Ratio is the total tier-1 capital over total assets. The Non-Performing Loans variable is total non-performing loans over lagged total assets. Facility Amount is the syndicate loan amount from the loan level sample, in billions of dollars. Bank Amount is the share of the facility lent by a particular bank, in billions of dollars.

average equity ratio of 10.3 percent. The average lending over total assets is about 62.73 percent, and trading assets account for 19.85 percent on average.

While the summary statistics report the nominal values, in all subsequent analyses, I demean the independent variables so the coefficients in regressions can be interpreted as a percentage change in the dependent variable in response to a one standard deviation increase in the independent variable. Unless otherwise noted, standard errors are clustered at the bank and quarter level.
3 Measuring Uncertainty

This section describes how bank level uncertainty is measured using text from the earnings conference calls. The process, illustrated in Figure 3, involves two steps. First, I use natural language processing and machine learning to construct a list of uncertainty words, $S_{uncertainty}$, using the entire corpus of conference call transcripts. Second, using the new list of words, I create a general bank level uncertainty measure counting the frequency of these words in the transcript for each bank at each quarter. I also generate topic-specific measures of uncertainty by isolating parts of the call dealing with particular topics, then counting the frequency of the uncertainty words using only those sections.

Figure 3: Process of Creating Bank Level Uncertainty

Step 1 (Section 3.1): Finding Words of Uncertainty, $S_{uncertainty}$

Step 2a (Section 3.2): General Bank Level Uncertainty Counting $S_{uncertainty}$ using entire conference call transcript

Step 2b (Section 3.3): Topic-Specific Bank Level Uncertainty Counting $S_{uncertainty}$ only in sections of call dealing with topic

3.1 Finding Words of Uncertainty, $S_{uncertainty}$

Most recent economic and finance research with text analysis involves word counts where the researcher determines the words which they believe capture a specific topic (such as “uncertain” and “uncertainty” for uncertainty) or chooses a pre-made list of words carefully vetted by subject matter experts. For example, Loughran and McDonald 2011 generate a list of uncertainty words by ciphering through 10-K filings of public companies, retaining only words that occur in at least 5 percent of the filings.

To mitigate subjectivity from the researcher choosing words that share similar meaning to a particular concept, or using a pre-made dictionary that may not be tailored for banking text, I rely on recent developments in machine learning and natural language processing to more objectively and automatically identify a new word list.
3.1.1 Word Embeddings

Word embeddings, developed in Mikolov et al. 2013a, are vector representations of words where distances preserve syntactic and semantic similarities between words. The embeddings serve not only as a dimension reduction tool for representing words, but also as a way of preserving the syntactic and semantic relationship in a Euclidean space. For example, word embeddings estimated in Mikolov et al. 2013a famously predicts the relationship: i) vec(Madrid) - vec(Spain) + vec(France) is closest to vec(Paris), and ii) vec(Germany) + vec(capital) is closest to vec(Berlin). In the appendix, I provide a detailed summary of the methodology from Mikolov et al. 2013a in the context of a neural network, the parameters, and the estimation strategy.

Word embeddings can be generated through the Skip Gram model, also introduced in Mikolov et al. 2013a. For example, suppose we have three documents as follows:

Document 1 : we think uncertainty about unemployment
Document 2 : we think fears about unemployment
Document 3 : we think fears and uncertainty about unemployment

The words uncertainty and fears share several similarities, so an ideal machine learning model would group these words together. The words share similar syntax as they are both preceded at least once by the word think and followed by the word about. Semantically, uncertainty and fears both evoke feelings of worry. The Skip Gram model attempts to project both words into the same region in some abstract vector space of dimension $H$. In the example of the three document examples above, seven unique words lead to seven unique word embeddings:

$$
\begin{bmatrix}
    u_{about}^1 & u_{uncertainty}^1 & u_{fears}^1 & u_{we}^1 & u_{unemployment}^1 & u_{think}^1 & u_{and}^1 \\
    u_{about}^2 & u_{uncertainty}^2 & u_{fears}^2 & u_{we}^2 & u_{unemployment}^2 & u_{think}^2 & u_{and}^2 \\
    u_{about}^3 & u_{uncertainty}^3 & u_{fears}^3 & u_{we}^3 & u_{unemployment}^3 & u_{think}^3 & u_{and}^3
\end{bmatrix}
= \begin{bmatrix}
    u_{about} & u_{uncertainty} & u_{fears} & u_{we} & u_{unemployment} & u_{think} & u_{and}
\end{bmatrix}
$$

In this example, $H$ has been set to 3, but this is a choice that should be determined by the researcher. Ideally, we would want the three-dimensional vector $u_{uncertainty}$ to be next to $u_{fear}$ as...
they share similar syntax and semantics. The Skip Gram model will learn the positioning of each word in the vector space by using the same technique used to learn the meaning of a new word: considering the context.

Mathematically, the Skip Gram model takes every word in every sentence and attempts to predict the words in the context. For example, with the word uncertainty in Document 1, the input is the word uncertainty, and the outputs will be the words in the context, think and about:

\[
\begin{align*}
\text{context (output)} & & \text{word (input)} & & \text{context (output)} \\
\text{we} & & \text{think} & & \text{uncertainty} & & \text{about} & & \text{unemployment}
\end{align*}
\]

The model maximizes the probability of the words think and about conditional on seeing the word uncertainty: \( P(\text{think} | \text{uncertainty}) \times P(\text{about} | \text{uncertainty}) \). The procedure works similarly to a multinomial logistic regression, where:

\[
P(\text{think} | \text{uncertainty}) = \frac{\exp(\beta'_{\text{think}} \mathbf{u}_{\text{uncertainty}})}{\sum_{j=1}^{V} \exp(\beta'_{j} \mathbf{u}_{\text{uncertainty}})}; \quad P(\text{about} | \text{uncertainty}) = \frac{\exp(\beta'_{\text{about}} \mathbf{u}_{\text{uncertainty}})}{\sum_{j=1}^{V} \exp(\beta'_{j} \mathbf{u}_{\text{uncertainty}})}
\]

with \( \beta'_{j} = \mathbf{u}'_{j} \). Initially, the word vectors are randomly initialized. The predictions for the probabilities \( P(\text{about} | \text{uncertainty}) \) and \( P(\text{think} | \text{uncertainty}) \) will be incorrect, so the model will adjust all coefficients (each \( \mathbf{u}_{j} \) and \( \beta_{j} \)) through a technique known as backpropogation, such that the probabilities of these two values are closer to 1 and the prediction of all other words moves closer to 0. The process continues using the newly adjusted \( \mathbf{u}_{j} \) word embeddings but with the next word in the sequence as input (about) and predicting the context (uncertainty and unemployment). This is done for every word, in every document, until a satisfactory error threshold is met.

The resulting word embeddings, \( [\mathbf{u}_{\text{about}}, \mathbf{u}_{\text{uncertainty}}, ..., \mathbf{u}_{\text{and}}] \) will preserve the structure, such that words which share similar meanings and which are used the same way in a sentence will be close to each other in the vector space.

I use the K-means algorithm to cluster the word embeddings into \( K \) disjoint sets to identify the cluster containing “uncertainty” and “uncertain.” K-means is an unsupervised learning algorithm that takes as input a set of vectors \( \{\mathbf{u}_{1}, \mathbf{u}_{2}, ..., \mathbf{u}_{V}\} \) and a hyperparameter, \( K \), representing the
number of clusters. The output is \( K \)-disjoint sets, \( \{S_1, S_2, \ldots, S_K\} \), composed of the input vectors.\(^5\)

### 3.1.2 Estimation

The Skip Gram model for the earnings conference call transcripts was estimated with \( H = 300 \) and a context window size of 10. The dimension size of 300 is similar to the size used in Mikolov et al. 2013a. The Mikolov et al. paper uses a context window size of 5, but the authors state that a larger window “results in more training examples and thus can lead to a higher accuracy, at the expense of the training time.” After inspecting several models with dimension sizes ranging from 100 to 300 and window sizes ranging from 5 to 10, \( H = 300 \), and a window size of 10 produced the most coherent representations.

I estimated the model with 2,732 earnings conference call transcripts that contained 18,617 unique words. Figure 1 illustrates the two-dimensional projection of the 300-dimensional word embeddings using a dimension reduction technique known as t-SNE (described in the appendix). Each point represents a unique word in the vocabulary across all 2,732 conference calls. The Euclidean proximity of two points proxies the similarity of the words both semantically and syntactically.

Words related to names, with “jpmorgan,” “barclays,” and “suntrust” identified in yellow, appear in the upper left portion of Figure 1. Due to their strong association with each other, “fannie” (from Fannie Mae) and “freddie” (from Freddie Mac) are shown nearly on top of each other in pink near the bottom middle section. The center region displays in turquoise the word embeddings of several political and economic personas, such as “mnuchin,” “paulson,” “yellen,” and “trump.” The bottom left portion displays words pertaining to forecasts, such as “forecast,” “outlook,” “projection,” and “estimate.” Last, on the right side of the figure, the word embeddings seem to associate the word “uncertainty” with “instability,” “fear,” “war,” and “illiquidity” due to their strong syntactic and semantic similarities.

While the two-dimensional image in Figure 1 is helpful for picturing the 300-dimensional word vectors on a plane, I use the original 300-dimensional vectors to cluster the words into 350 disjoint

---

\(^5\)Training the algorithm works iteratively in two steps. At the onset of training, a set of vectors, \( \{c_1, c_2, \ldots, c_K\} \) is chosen randomly as cluster centroids. The first step is cluster assignment in which each \( u_i \) is assigned to the cluster, which minimizes the distance between \( u_i \) and the cluster’s centroid. In other words, each \( u_i \) is assigned to cluster \( k \) if \( k = \arg \min \limits_k \text{dist}(u_i, c_k) \). The second step is updating the cluster centroids \( \{c_1, c_2, \ldots, c_K\} \) so that \( c_k \) is the average of all points assigned to cluster \( k \). These two steps are repeated until the sum of the squared errors is minimized.
adapted, amid, amidst, amplified, anxiety, attacks, austerity, backdrop, benign, bipartisan, brexit, ceiling, challenges, challenging, cliff, climate, clouded, commonwealth, concerns, conditions, confluence, confronting, congress, consumption, crash, crises, currents, cycles, deficit, deficits, deflation, deflationary, downturn, dysfunction, economic, election, elections, emerged, encountered, environment, environments, eu, euro, eurozone, face, faced, faces, facing, fears, fiscal, flash, fragile, franc, geo, geopolitical, governmental, governments, gridlock, gyrations, hampering, headwinds, heighten, illiquidity, immune, impasse, instability, intervention, iraq, lackluster, legislative, legislature, lingering, looming, ltro, macroeconomic, makers, midst, midterm, monetary, myriad, nafta, navigate, navigated, navigating, paralysis, persist, persisted, persistent, persistently, persists, peso, political, posed, presidential, prevailed, prevailing, prolonged, protracted, psychology, reactions, realities, recessionary, referendum, reforms, rhetoric, rican, ripple, sars, sequester, shutdown, sluggish, society, sparked, spite, stimulative, stimulus, stressful, struggles, surrounding, swiss, tariff, tariffs, tensions, terrorism, terrorist, threat, threats, tsunami, tumultuous, turbulence, turbulent, turmoil, uncertain, uncertainty, uneven, unfolded, unprecedented, unrest, unsettled, unstable, upheaval, war, weathered, weathering, withstand

Note: This table reports the new word list I generate by clustering words close to the word embeddings of “uncertainty” and “uncertain” using the K-means algorithm.

The cluster containing “uncertainty” and “uncertain,” which I refer to as the resulting “uncertainty” dictionary, $S_{\text{uncertainty}}$, is shown in Table 2. The list contains words associated with uncertainty, such as “fears,” “unprecedented,” and “instability.” Similarly, the methodology picks up events that are associated with rises in uncertainty, such as “brexit,” “terrorism,” and “war.” We also see words that are economic-specific and typically lead to uncertainty in financial markets, such as “illiquidity,” “recessionary,” and “crises.” The fact that words describing downturns, such as “recessionary,” “downturn,” and “crises,” appear in the list could explain the difficulty of measuring uncertainty separate from the business cycle. In the conference call transcripts, these words must appear in the same context as the word “uncertainty” according to their word embeddings, and could be perceived interchangeably by banks. Overall, the set appears to accurately resemble words related to uncertainty.

### 3.2 General Bank Level Uncertainty

The first measure used for bank level uncertainty is the frequency of uncertainty words, the cluster $S_{\text{uncertainty}}$ defined in the previous section, in bank conference calls. Formally, bank level

---

6The value of 350 was chosen after running a series of cross-validation tests for various cluster values and comparing residual scores.
uncertainty is defined as:

\[
\sigma_B = \frac{\sum_{t \in T} w_t \times \mathbb{1}(t \in S_{uncertainty})}{|T|}
\]  

(1)

where \(T\) is the set of all words in the earnings conference call; \(\mathbb{1}(t \in S_{uncertainty})\) is a dummy variable equal to 1 if word \(t\) from from the transcript is in the dictionary \(S_{uncertainty}\); and \(w_t\) is a term-frequency inverse document-frequency (TF-IDF) weight, common in text analysis, which provides higher weights to discriminative words in a document.

Table 3: Examples of High Uncertainty Responses

<table>
<thead>
<tr>
<th>Bank</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital One Financial Q3 2008</td>
<td>Like all banks, we face increasing cyclical economic headwinds and market uncertainties. We remain well positioned to navigate the near term challenges and to realize value-creating opportunities when the time is right. Now Gary and I will answer your questions. Jeff?</td>
</tr>
<tr>
<td>SVB Financial Group Q3 2013</td>
<td>All that said, we’re also well aware of the challenges facing us and the banking industry. The lackluster economy continues to create uncertainty. Headwinds from low interest rates continue to limit the pace of growth.</td>
</tr>
<tr>
<td>Northern Trust Q3 2016</td>
<td>In closing, the global macroeconomic environment continued to produce a difficult operating environment in the third quarter of 2016. Low and even negative interest rates around the globe; post-Brexit uncertainty; election uncertainty; and debate over central bank actions characterized the quarter. Despite that challenging backdrop, Northern Trust produced solid financial results, growing our earnings per share 13% year over year.</td>
</tr>
</tbody>
</table>

Note: This table reports the responses with high \(\sigma_B\).

The uncertainty measure is computed at the bank-quarter frequency, as earnings conference calls occur once per quarter per bank. It is useful, however, to look at examples for individual responses, such as how executive officers answer particular questions during the call. Table 3 shows examples of responses with high measures of uncertainty. For example, Capital One Financial discusses

\footnote{As common is textual analysis, I remove frequent words from the earnings conference calls. These include English pronouns, auxiliary verbs, and articles.}
in third quarter 2008 *cyclical economic headwinds and uncertainties* the bank faces. Other highly uncertain responses include *challenging backdrop* and *lackluster* as descriptors of the economy.

While $\sigma_B$ provides some indication of the uncertainty signaled by a bank at a particular quarter, I am interested in the uncertainty that is idiosyncratic to the bank independent of aggregate conditions. The bank may experience large uncertainty merely as a result of an increase in aggregate uncertainty; therefore, it is important to filter out the components of $\sigma_B$ explained by aggregate measures. I do this through the following regression:

$$ (\sigma_B)_{b,t} = \alpha + \beta_1 \Delta EPU_t + \beta_2 \Delta VIX_t + \beta_3 \Delta CorpSpread_t + \gamma_b + \epsilon_{b,t} \quad (2) $$

$\Delta EPU_t$ captures uncertainty from policy and is measured by the quarterly growth in the Economic Policy Uncertainty (EPU) index from Baker et al. 2016. $\Delta VIX_t$ is the quarterly growth of the VIX. $\Delta CorpSpread_t$ proxies uncertainty about the economy, measured by the growth in the ten-year less two-year corporate bond rate. The residual $\epsilon_{b,t}$ captures the component of bank uncertainty, $\sigma_B$ unexplained by aggregate uncertainty. Thus, idiosyncratic uncertainty is the residual,

$$ BankUncert_{b,t} = \hat{\epsilon}_{b,t} \quad (3) $$

### 3.3 Topic-Specific Uncertainty

While *BankUncert* represents the idiosyncratic uncertainty of the bank during the entire conference call, I am also interested in topic-specific sections of the call to better understand the uncertainty related to particular topics, such as housing and interest rates. To extract bank responses on specific topics, I implement a Latent Dirichlet Allocation (LDA) model. Introduced by Blei et al. 2003, LDA is an unsupervised algorithm that takes as input a set of documents; a hyperparameter $K$; and the number of topics. The LDA algorithm outputs two objects: (1) A document-topic matrix revealing the distribution over topics of each document and (2) a term-topic matrix showing a distribution over all words in the corpus for each topic. Similar to principal component analysis, the topics themselves do not have any objective meaning but rather are interpreted by the researcher. LDA models have recently been used in the economic literature to gain insight into speech data, as in Hansen et al. 2018.
Figure 4: Housing Topic (Topics 26 and 38)

Note: This figure displays the topics I use to distinguish responses pertaining to housing.

Figure 5: Interest Rates Topic (Topics 14, 15 and 40)

Note: This figure displays the topics I use to distinguish responses pertaining to interest rates.

I ran several models with 25 to 75 topics and carefully inspected the output of each, comparing their goodness-of-fit through statistical means such as perplexity and topic coherence. I settled on a topic model of 60 topics because of the low perplexity, high coherence measures, and interpretable topics. I isolated the topics related to housing and interest rates. The word clouds for these topics are displayed in Figures 4 and 5. Figure 4 groups together distributions of words related to real estate. A response is labeled as a “real estate” response if it predominantly contains words such as “mortgage,” “origin,” or “purchase” (Topic 26) or “commercial,” “estat,” “loan,” or “construct” (Topic 38). The interest rate topic identifies responses with “balanc,” “sheet,” or “risk” (Topic 14); “impact,” “effect,” or “rate” (Topic 15); or “deposit,” “margin,” “interest,” or “basi” (Topic 40). Similarly, I label a response as related to “interest rates” if the response contains words predominantly from the distribution of words shown in Figure 5.

Using these broader categories of topics, only responses predominantly related to either housing or interest rates are retained. For each of these two broader topics, the frequency of the occurrence of words from $S_{uncertainty}$ is counted, adjusting the denominator $T$ to reflect only sections of the call
Table 4: Topic-Specific Uncertainty Response Examples

**Interest Rates**

Independent Bank Q4 2011  
While all in all 2011 was a good year for us, and looking ahead we really must balance our natural enthusiasm and confidence in our franchise with the macro **uncertainty** that lingers across the economic, legislative, regulatory and political spectrums. Our industry is still facing meaningful revenue headwinds. Interest margins are clearly under pressure from the Fed’s concerted efforts to hold the rates down.

Boston Private Q4 2014  
As we enter the early portion of 2015, we do believe that a thin yield interest rate scenario is upon us. While the interest rate environment is volatile and the duration of this environment is hard to predict, amid concerns regarding global economic growth and heightened geopolitical risk, I do want to address how we’re thinking about performance in this environment.

**Housing**

Sun Bancorp Q3 2010  
Clearly this quarter presented tremendous challenges as our commercial real estate portfolios saw devaluation as the result of a profound slowdown in the commercial real estate market and an overall constricted economic environment in the markets we serve.

F.N.B. Corporation Q4 2012  
Slide four shows our 2012 quarterly loan growth trends. Our team has accomplished this consistent growth despite a challenging economic environment and historically low line utilization. Additional headwinds included significant reductions in the Florida portfolio and acceleration of prepayment speeds in the residential portfolio.

---

*Note:* This table reports examples of responses with high BankUncert in the interest rate and housing topics.

related to each topic. BankUncert for each conference call and each topic is then recalculated using equations (1) and (2). Thus, for each conference call, two new variables are generated: idiosyncratic uncertainty about housing and idiosyncratic uncertainty about interest rates.

Table 4 displays the top responses by bank management for the two topics analyzed. Reviewing the top responses for each topic helps validate that the topic modeling is able to objectively, and in an automated manner, select the portions of the earnings call transcripts specific to each topic.
4 Empirical Predictions

To validate $BankUncert$ as a measure of uncertainty, the measure should reflect changes in the second moment of a bank’s profitability. If this measure was negatively associated to stock returns, then it could be construed as another measure of bad news rather than uncertainty. Thus, if $BankUncert$ truly picks up the bank’s uncertainty, it should be positively associated with the second moment of the banks returns (i.e. post conference call volatility). Because all banks in my sample are public, I use their stock price returns and volatilities to proxy asset profitability.

Second, increased idiosyncratic uncertainty of a banks balance sheet could signal more dispersion in the distribution of their return on assets, which could be limited by regulators and shareholders. In order to meet constraints, the bank might respond by reducing lending. When profitable credit opportunities erode, banks shift the composition of their balance sheets in favor of credit rationing, securities, or search for yield (Stiglitz and Weiss 1992, Abbassi et al. 2016). The effect could be exacerbated by higher equity capital at risk or during times of higher aggregate volatility. In the appendix, I develop a stylized framework illustrating these hypotheses using a risk-neutral bank that is facing shareholder and regulator constraints.

The empirical predictions of the stylized framework are twofold. First, any measure of idiosyncratic uncertainty, given a level of investment in lending and equity capital, increases the volatility of returns. Second, the optimal level of investment in lending decreases amid a bank’s idiosyncratic uncertainty. The decline in lending as a result of higher idiosyncratic uncertainty is more severe for banks with larger equity capital funding and periods in which aggregate uncertainty is high.

5 Results

5.1 Validation

To interpret $BankUncert_{b,t-1}$, I start by reviewing how the frequency counts of $S_{uncertainty}$ relate to aggregate uncertainty variables. First, I regress $\sigma_B$ on the growth of these aggregate uncertainty measures, as in (2) to help validate the measure as a proxy for bank level uncertainty. Columns 1 and 2 in Table 5 show the results of the regression. $\sigma_B$ is statistically and positively related to the growth in the VIX as well as the term structure of corporate bond yields. As the financial sector
is largely responsible for allocating capital to the real economy through corporate lending, the
positive relationship suggests that uncertainty in the corporate bond market propagates to bank
speech. It is also useful to see how the main variables of interest on the bank balance sheet, lending
and trading, react to these aggregate uncertainty measures. Lending, shown in columns 3 and 4, is
positively and significantly correlated to growth in the VIX, while trading, shown in columns 5 and
6, is negatively and significantly correlated. The growth in the EPU index is negatively associated
with lending, which suggests that increasing economic policy uncertainty may lead banks to reduce
credit.

Table 5: Filtering Out Aggregate Market Conditions

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta VIX_t )</td>
<td>0.0010**</td>
<td>0.0011**</td>
<td>0.0186***</td>
<td>0.0190***</td>
<td>-0.0184***</td>
<td>-0.0186***</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td>(0.00)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>( \Delta EPU_t )</td>
<td>0.0005*</td>
<td>0.0005</td>
<td>-0.0102***</td>
<td>-0.0101**</td>
<td>0.0068*</td>
<td>0.0067</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>( \Delta CorpBond_t )</td>
<td>0.0001*</td>
<td>0.0001*</td>
<td>-0.0017</td>
<td>0.0010</td>
<td>0.0015</td>
<td>-0.0011</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank FE</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
<th>N</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.0282</td>
<td>0.2096</td>
<td>0.0032</td>
<td>0.9128</td>
<td>0.0064</td>
<td>0.7434</td>
</tr>
</tbody>
</table>

Note: This table reports regression results of bank variables on aggregate uncertainty measures. \( (\sigma_B)_{b,t} \) is the
count of uncertainty terms of a bank’s quarterly earnings conference call. Lending is total quarterly commercial
and industrial, agricultural, consumer, and foreign loans normalized by total assets. Trading is available-for-sale and
hold-until-maturity portfolios normalized by total assets. \( \Delta VIX \) is the change in the CBOE VIX index in the current
quarter. \( \Delta EPU \) is the change of the Economic Policy Uncertainty index from Baker et al. 2016 in the current quarter.
\( \Delta CorpBond \) is the growth in the ten-year less two-year corporate yield in the current quarter. Standard errors double
clustered at the bank and quarter level are reported in parentheses. * \( p < 0.10 \), ** \( p < 0.05 \), *** \( p < 0.01 \).

I use the residuals of column 2 in Table 5 as a proxy for idiosyncratic bank level uncertainty,
\( BankUncert \). Because in the main specification I regress lending and trading on these residuals,
applying the Frisch-Waugh theorem, lending and trading are also filtered for the same covariates.

In Table 6, I separate the sample into banks displaying high and low idiosyncratic bank uncertainty
at each quarter using the median value as the threshold. Bank-quarter observations with
high idiosyncratic bank uncertainty have higher uncertainty word counts on average (0.70 percent) than low idiosyncratic bank uncertainty (0.23 percent). Unconditionally, banks exhibiting high uncertainty report 3.51 percent fewer loans than banks with low uncertainty. In contrast, trading assets are higher among those with high $\text{BankUncert}_{b,t-1}$, suggesting banks compensate for the reduction in credit with more trading. The bank-quarter observations exhibiting high uncertainty correlate positively to growth in the VIX, EPU, and ten-year minus two-year yield.

### Table 6: High and Low Bank Uncertainty

<table>
<thead>
<tr>
<th></th>
<th>Low Uncertainty (N=1,377)</th>
<th>High Uncertainty (N=1,355)</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>$\sigma_B$</td>
<td>0.23</td>
<td>0.14</td>
<td>0.7</td>
</tr>
<tr>
<td><strong>Dependent Variables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Loans</td>
<td>64.47</td>
<td>15.54</td>
<td>60.96</td>
</tr>
<tr>
<td>Trading Assets</td>
<td>19.27</td>
<td>10.07</td>
<td>20.43</td>
</tr>
<tr>
<td><strong>Independent Variables and Controls</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIX Growth</td>
<td>1.27</td>
<td>40.59</td>
<td>1.49</td>
</tr>
<tr>
<td>EPU Growth</td>
<td>27.3</td>
<td>68.94</td>
<td>27.5</td>
</tr>
<tr>
<td>CorpSpread Growth</td>
<td>-3.77</td>
<td>166.95</td>
<td>-2.7</td>
</tr>
</tbody>
</table>

*Note:* This table divides observations in the sample between above and below the median bank uncertainty level. The rightmost column reports the differences in the mean. $\sigma_B$ is the count of uncertainty terms of a banks quarterly earnings conference call. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The time-series plot of idiosyncratic uncertainty picks up several well-known events hampering the banking sector and bank uncertainty over the last business cycle. Figure 2 plots the median value of $\text{BankUncert}_{b,t-1}$ at each quarter. As the recent financial crisis began to develop in December 2007, banks begin to exhibit larger idiosyncratic uncertainty. Bank level uncertainty remained high following the collapse of Lehman Brothers in September 2008, peaking the following quarter as the bailout of the financial system was announced. The dashed blue lines show the 25th and 75th percentiles of the uncertainty of the banks at a given quarter. Cross-sectionally, the boom period between 2004 and 2007 showed moderate disparities in uncertainty, while the beginning of the crisis in December 2007 suggests banks spoke similarly about uncertainty with small differences between the 25th and 75th percentiles. The interval rose subsequently during the peak and end of the recession, remaining wide thereafter.

Finally, to validate my measure of idiosyncratic uncertainty as a form of risk, I assess how the
Table 7: Bank Level Uncertainty and Post Call Volatility

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Volatility</strong>&lt;sub&gt;b,t+3&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BankUncert&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>0.1917***</td>
<td>0.1919***</td>
<td>0.1286**</td>
<td>0.1812**</td>
<td>0.1049</td>
<td>-0.0573</td>
</tr>
<tr>
<td>Previous Volatility</td>
<td>0.2342***</td>
<td>0.2329***</td>
<td>0.2317***</td>
<td>-0.1227***</td>
<td>-0.1238***</td>
<td>-0.1211***</td>
</tr>
<tr>
<td>Previous Return</td>
<td>-0.0110***</td>
<td>-0.0110***</td>
<td>-0.0108***</td>
<td>0.0114***</td>
<td>0.0113***</td>
<td>0.0111***</td>
</tr>
</tbody>
</table>

Bank Controls | N | Y | Y | N | Y | Y |
Bank FE | Y | Y | Y | Y | Y | Y |
Quarter FE | N | N | Y | N | N | Y |
N | 2,732 | 2,732 | 2,732 | 2,732 | 2,732 | 2,732 |
R<sup>2</sup> | 0.4317 | 0.4395 | 0.4641 | 0.0576 | 0.0658 | 0.1017 |

Note: This table shows the effect of bank uncertainty on post-call volatility. Columns 1-3 show results where volatility is the standard deviation of excess returns, as measured through a 3-factor Capital Asset Pricing Model, for <i>t</i> + 1 to <i>t</i> + 3 days after the conference call, while columns 4-6 show results for excess returns three days after the conference call. Previous Return (Volatility) is the excess return (volatility) during the quarter of the call. BankUncert<sub>b,t</sub> is the count of uncertainty terms of a bank’s quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes. Bank Controls include the following contemporaneous variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total asset, tier-1 capital over total assets and total non-performing loans over lagged total assets. Standard errors double clustered at the bank and quarter level reported in parentheses. * <i>p</i> < 0.10, ** <i>p</i> < 0.05, *** <i>p</i> < 0.01.

measure is associated with post-call volatility. First, I analyze the effect of idiosyncratic uncertainty on volatility three days after the call in columns 1-3 of Table 7. Column 1 illustrates how idiosyncratic uncertainty is positively and significantly associated with post-call volatility using only bank fixed-effects. Even when controlling for size, equity, non-performing loans, non-interest income, tier-1 ratio, and liquidity, the positive association between idiosyncratic uncertainty and short-run volatility remains statistically significant, as evident from column 2. By including the most stringent specification with quarter fixed effects, the impact of a one standard deviation increase in bank level uncertainty increases post-call three-day volatility 13 basis points, remaining statistically significant. The coefficients provide evidence that the content of the earnings call provides new information not previously priced into the banks market value.

Columns 4-6 repeat the exercise with the cumulative absolute excess returns three days after the call as the dependent variable. In the strictest regressions with full fixed effects and controls, short-term returns are not associated with BankUncert. Because BankUncert is more associated
with volatility than prices, which represent discounted future cash flows, the results of Table 7 suggest that uncertainty is not a euphemism for future losses and justifies using BankUncert as a proxy for bank uncertainty.

5.2 Balance Sheet Responses to Bank Level Uncertainty

Do banks cutback more on lending in lieu of increased trading when they emit higher uncertainty to the market? To answer this question, I regress lending and trading assets on BankUncert_{b,t-1}:

\[ Y_{Resid_{b,t}} = \beta_1 \text{BankUncert}_{b,t-1} + X_{b,t-1} + \gamma_b + \delta_t + \epsilon_{b,t} \quad (4) \]

\( Y_{Resid} \) and \( \text{BankUncert}_{b,t-1} \) are the residuals from (2), where the dependent variables are lending and trading assets. I include time fixed effects to control for observed and unobserved time varying characteristics. This allows me to control for all aggregate variables related to credit, such as market pricing, the stochastic discount factor, and aggregate firm productivity. The coefficient \( \beta_1 \) reports the change in \( Y_{Resid_{b,t}} \) as a response to an increase in bank uncertainty through speech in the earnings conference calls. Time fixed effects allow \( \beta_1 \) to represent the change in \( Y_{Resid_{b,t}} \) at a given quarter relative to other banks, as the time dimension is muted. Further, the vector of controls, \( X_{b,t-1} \), addresses potential endogeneity stemming from the fact that the relationship between what banks speak about at \( t - 1 \) and \( Y_{Resid_{b,t}} \) could be confounded by bank-specific characteristics in the last quarter, such as size, equity, non-performing loans, non-interest income, tier-1 ratio, and liquidity. Last, bank fixed effects, \( \gamma_b \), control for time-invariant bank characteristics such as corporate governance structures and risk appetites.

Results are reported in columns 1 and 4 of Table 8. With the full set of controls, higher BankUncert_{b,t-1} is associated with lower lending and more trading. In a given quarter, a bank with one standard deviation more uncertainty decreases lending next quarter by nearly 40 basis points relative to total assets. The regressions on trading assets imply movements away from lending into more liquid assets such as securities. A one standard deviation increase in idiosyncratic uncertainty is associated to a 38 basis point increase in trading assets, suggesting substitution between lending and trading when bank level uncertainty rises.
Table 8: Bank Level Uncertainty on Lending and Trading

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable:</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Loans&lt;sub&gt;b,t&lt;/sub&gt;</td>
</tr>
<tr>
<td>BankUncert&lt;sub&gt;b,t-1&lt;/sub&gt;</td>
<td>-0.0040**</td>
<td>-0.0036**</td>
<td>-0.0038**</td>
<td>-0.0039**</td>
<td>-0.0035**</td>
<td>0.0037**</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>BankUncert&lt;sub&gt;b,t-1&lt;/sub&gt; * Equity&lt;sub&gt;b,t-1&lt;/sub&gt;</td>
<td>-0.0043***</td>
<td>0.0038**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BankUncert&lt;sub&gt;b,t-1&lt;/sub&gt; * HighAggVol&lt;sub&gt;t-1&lt;/sub&gt;</td>
<td>-0.0023**</td>
<td>0.0022***</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bank Controls</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bank FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Quarter FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
</tr>
<tr>
<td>R&lt;sup&gt;2&lt;/sup&gt;</td>
<td>0.3070</td>
<td>0.3117</td>
<td>0.3089</td>
<td>0.3085</td>
<td>0.3118</td>
<td>0.3101</td>
</tr>
</tbody>
</table>

Note: This table reports the effect of bank uncertainty on bank balance sheet variables. Lending is total quarterly commercial and industrial, agricultural, consumer, and foreign loans normalized by total assets. Trading is the available-for-sale and hold-until-maturity portfolios normalized by total assets. BankUncert<sub>b,t-1</sub> is the count of uncertainty terms of a bank’s quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes. Equity is total equity capital over total assets. HighAggVol<sub>t-1</sub> is a dummy variable equal to 1 when the growth in the VIX index is above the median level in the sample and 0 otherwise. Coefficients of Equity<sub>b,t-1</sub> and HighAggVol<sub>t-1</sub> are estimated but not reported. Bank Controls include the following lagged variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total asset, tier-1 capital over total assets, and total non-performing loans over lagged total assets. Standard errors double clustered at the bank and quarter level reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

To better assess the extent to which uncertainty affects next-quarter credit and trading, I analyze the impact with two more sources of variation: equity and aggregate uncertainty. I run the following specification in Table 8:

\[ Y_{Resid,b,t} = \beta_1 \text{BankUncert}_{b,t-1} + \beta_2 \text{BankUncert}_{b,t-1} * D_{b,t-1} + X_{b,t-1} + \gamma_b + \delta_t + \epsilon_{b,t} \]

First I use \( D_{b,t-1} \) equal to equity over total assets of bank \( b \) at quarter \( t - 1 \). By exploiting the variation of leverage at the time of the conference call, this regression allows me to assess whether banks with more skin in the game, that is higher equity, align their actions more with their words. A bank with higher equity that speaks with more uncertainty has more incentive not to confuse investors, therefore enabling the bank to respond appropriately to higher uncertainty with fewer credit extensions. Columns 2 and 5 of Table 8 suggest that banks with higher equity reduce lending...
and increase trading when they emit a noisier signal through speech.

Next, I use $D_{b,t-1}$ equal to an increase in the VIX as a source of exogenous variation to see if alignments match more when aggregate volatility is high. $HighAggVol_{t-1}$ is a dummy variable equal to 1 when quarter $t - 1$ experiences an above-median increase in the growth of the VIX and 0 otherwise. The sign of $\beta_2$ is unclear. On one hand, when aggregate volatility is high, banks that are extremely uncertain may themselves be on the verge of insolvency, thus they may reach for yield in the next quarter by increasing lending to riskier borrowers. Thus $\beta_2$ could be positive. On the other hand, banks could be frank in their assessment of uncertainty, especially when uncertainty rises in the backdrop, and reduce credit more drastically when speaking with high uncertainty, leading to a negative coefficient.

Columns 3 and 6 show that the effects of $BankUncert_{b,t-1}$ on bank behavior are stronger during times of high aggregate uncertainty. Banks that speak with more uncertainty during times of high volatility reduce credit more than the average bank. The contrary is true for trading as seen by the positive and significant coefficient in column 4.

### 5.3 Topic-Specific Uncertainty

Using the two topic-specific uncertainty measures, real estate uncertainty and interest rate uncertainty, I next assess the impact on next-quarter real estate loans and interest rate derivatives, respectively. I run the following specification:

$$Y_{Resid_{b,t}} = \beta_1 BankUncert_{b,t-1} + \beta_2 TopicUncert_{b,t-1} + \beta_3 TopicAttention_{b,t-1} + \beta_4 TopicUncert_{b,t-1} \times TopicAttention_{b,t-1} + \beta_5 X_{b,t-1} + \gamma_b + \delta_t + \epsilon_{b,t}$$

where $TopicUncert_{b,t-1}$ is the TF-IDF weighted count of uncertainty words $S_{uncertainty}$ in sections of the call dealing with a particular topic, filtered for aggregate uncertainty as in (2). Because the effect of uncertainty on a particular asset class might depend on emphasis of the topic during the call, I interact the topic-specific uncertainty with the proportion of the call devoted to the topic, $TopicAttention_{b,t-1}$, as estimated by the topic model.

Table 9 shows how topic-specific bank uncertainty affects particular asset classes. The dependent
### Table 9: Topic-Specific Uncertainty

<table>
<thead>
<tr>
<th>Topic:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable:</td>
<td>Housing RealEstate&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td>Interest Rates IntRateExposure&lt;sub&gt;b,t&lt;/sub&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$BankUncert_{b,t-1}$</td>
<td>-0.0015</td>
<td>-0.0016</td>
<td>0.0007**</td>
<td>0.0007**</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>$TopicUncert_{b,t-1}$</td>
<td>-0.0044**</td>
<td>-0.0048***</td>
<td>0.0010**</td>
<td>0.0010**</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>$TopicAttention_{b,t-1}$</td>
<td>0.0126***</td>
<td>0.0117***</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>$TopicUncert_{b,t-1} \times TopicAttention_{b,t-1}$</td>
<td>-0.0027***</td>
<td>-0.0029***</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
</tbody>
</table>

Bank Controls

<table>
<thead>
<tr>
<th>Bank FE</th>
<th>N Y</th>
<th>N Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quarter FE</td>
<td>Y Y</td>
<td>Y Y</td>
</tr>
<tr>
<td>N</td>
<td>2,732</td>
<td>2,732</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2526</td>
<td>0.3145</td>
</tr>
</tbody>
</table>

Note: This table estimates the effect of topic-specific bank uncertainty on bank balance sheet variables. Real Estate is total loans secured by real estate. Interest Exposure is total interest rate exposure over total assets. $BankUncert_{b,t-1}$ is the count of uncertainty terms of a bank’s quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes. Columns 1 and 2 report the coefficient of $BankUncert_{b,t-1}$ only for responses during the earnings call pertaining the real estate cloud of Figure 4; 3 and 4 of the interest rate cloud of Figure 5. Bank Controls include the following lagged variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total asset, tier-1 capital over total assets, and total non-performing loans over lagged total assets. Standard errors clustered at the bank and quarter level reported in parentheses.

The uncertainty in real estate responses is associated with reductions in loans secured by real estate. This effect increases in the attention given to real estate and housing, as seen by the negative and significant coefficient on $TopicUncert_{b,t-1} \times TopicAttention_{b,t-1}$. Similarly, higher uncertainty of interest rates increases the exposure of bank interest rate positions, suggesting hedging activities or speculation. This effect is constant regardless of the attention paid to the topic of interest rates as seen by the insignificance of the double interaction. These results suggest topic-specific uncertainty may allow for a better lens to understand bank behavior.

---

Each dependent variable is the residual of a regression of the balance sheet variable on the covariates of (2).
5.4 Evidence at the Loan Level

The results in the previous sections suggest bank level uncertainty reduces credit the following quarter. However, data at the bank level cannot rule out the reverse causality that banks are uncertain precisely because they know firms’ demand for credit will be low, resulting in a negative relationship between bank uncertainty and lending.

To mitigate this concern, I use loan level data from Dealscan. The previous analysis using the Federal Reserve Y-9C form incorporated not only commercial lending but also loans to agricultural producers, consumers, and foreign firms. Although the type of lending in Dealscan is only a subset of bank lending, two features of the data help provide evidence of active management of uncertainty through credit cutbacks. First, the Federal Reserve Y-9C reports the stock of all bank lending, while Dealscan reports new loan issuances. Second, and perhaps more importantly, loan level data from Dealscan allows for the inclusion firm-time fixed effects. These fixed effects control for observed and unobserved characteristics at the firm level each quarter. The identifying assumption in firm-time fixed effects regression is that each firm at each time period must receive new loans from two different banks. I assume that each firm included in the estimation has a positive demand for credit during the quarter, allowing me to better estimate the effect of bank level uncertainty on the amount of the new loan issuances after controlling for demand-side factors. Summary statistics for the sample are shown in Table 1.

I run the following specification in Table 10:

\[
\text{Log} (\text{Loan})_{b,f,t} = \beta_1 \text{BankUncert}_{b,t-1} + X_{b,t-1} + \gamma_b + \delta_{f,t} + \epsilon_{b,f,t}
\]  

(6)

where \( \text{Log} (\text{Loan})_{b,f,t} \) is the log loan amount of \( b \) to firm \( f \) at time \( t \) and \( \delta_{f,t} \) is firm-time fixed effects to control for firm demand at time \( t \). Along with bank controls as described above, I saturate the model with an array of loan controls that include dummy variables for: the purpose of the loan, whether the syndicate contains multiple lead arrangers, whether the loan is a term loan or revolver loan, and whether bank \( b \) is a lead arranger. Column 1 excludes all firm and time fixed effects and shows a negative relationship between bank uncertainty and the amount of the loan. Columns 2 and 3 include year fixed effects to control for yearly trends and firm fixed effects to control for unobserved firm characteristics affecting demand for credit, such as time-invariant risk appetite
Table 10: Bank Level Uncertainty and Loan Level Data

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{BankUncert}_{b,t-1}$</td>
<td>-0.0582**</td>
<td>-0.0972**</td>
<td>-0.0823**</td>
<td>-0.0753**</td>
<td>-0.0501**</td>
</tr>
<tr>
<td></td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Bank Controls: Y Y Y Y Y
Loan Controls: Y Y Y Y
Bank FE: Y Y Y Y
Year FE: N Y
Quarter FE: N N N Y
Firm FE: N N Y
Firm-Quarter FE: N N N N Y
N: 1,094 1,094 1,094 1,094 1,094
$R^2$: 0.1284 0.1664 0.7638 0.7761 0.7818

Note: This table estimates the effect of bank uncertainty on corporate loan issuances from 2002-2013. The dependent variable is the log of the amount issued by bank $b$ to firm $f$ at quarter $t$ using Dealscan data. $\text{BankUncert}_{b,t-1}$ is the count of uncertainty terms of a bank’s quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes. Loan Controls include dummy variables for the purpose of the loan, whether the syndicate contains multiple lead arrangers, whether the loan is a term loan or revolver loan, and whether bank $b$ is a lead arranger. Bank Controls include the following lagged variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total assets, tier-1 capital over total assets and total non-performing loans over lagged total assets. Standard errors clustered at the bank level reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

and corporate structure. The coefficient on $\text{BankUncert}$ is negative and significant. In column 4, quarter fixed effects reduce the coefficient slightly to -0.0753 while remaining statistically significant. Column 5 shows the strictest regression including firm-time fixed effects. The coefficient drops in magnitude to -0.0501, suggesting that a one standard deviation increase in bank level uncertainty is associated with a nearly 5 percent drop in the loan amount of new credit issuances. Table 10 shows the importance of firm level fixed effects to account for demand, as columns (2)-(4), which exclude the granular firm-quarter fixed effects, report a downward biased estimate on the cutbacks in credit associated with larger uncertainty. In fact, the increase in the coefficients between columns 4 and 5 where firm-quarter fixed effects are included suggests $\text{BankUncert}$ is not orthogonal to observed and unobserved characteristics of the firm, especially firm demand (Altonji et al. 2005). Thus, while firm demand tomorrow plays an important role in determining bank uncertainty today, the negative coefficient suggests that even after controlling for firm demand, $\text{BankUncert}$ still predicts lending at the loan level.
Table 11: Bank Level Uncertainty and Lending During the Financial Crisis

<table>
<thead>
<tr>
<th>Dependent Variable:</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{Log(LoanShare)}_{b,f,t} )</td>
<td>( \text{BankUncert}<em>{b,t-1} \ast \text{LehmanConnection}</em>{b} \ast \text{Crisis}_{t} )</td>
<td>0.1375</td>
<td>-0.2754***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.19)</td>
<td>(0.06)</td>
</tr>
<tr>
<td>( \text{LehmanConnection}<em>{b} \ast \text{Crisis}</em>{t} )</td>
<td>-0.1943</td>
<td>0.0264</td>
<td>0.1188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.26)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>( \text{BankUncert}<em>{b,t-1} \ast \text{Crisis}</em>{t} )</td>
<td>-0.3258</td>
<td>0.0584</td>
<td>-0.7379**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.17)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>( \text{BankUncert}<em>{b,t-1} \ast \text{LehmanConnection}</em>{b} )</td>
<td>-0.3910</td>
<td>0.2933*</td>
<td>-0.2281</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.24)</td>
<td>(0.14)</td>
</tr>
<tr>
<td>( \text{Crisis}_{t} )</td>
<td>0.1379</td>
<td>-0.2456***</td>
<td>-0.9180***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.12)</td>
<td>(0.05)</td>
</tr>
<tr>
<td>( \text{BankUncert}_{b,t-1} )</td>
<td>0.2348</td>
<td>-0.3154**</td>
<td>-0.4498**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.17)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>Bank Controls</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Firm FE</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bank FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>( N )</td>
<td>241</td>
<td>241</td>
<td>241</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.0471</td>
<td>0.8976</td>
<td>0.8995</td>
</tr>
</tbody>
</table>

**Note:** This table reports the effect of bank uncertainty on corporate loans from 2005 to 2009. The dependent variable is new loan issuances of commercial loans from Dealscan at the bank-firm-quarter level. \( \text{BankUncert}_{b,t-1} \) is the count of uncertainty terms of a banks quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes. \( \text{LehmanConnection}_{b} \) is the percentage of loans \( b \) had syndicated with Lehman Brothers before the financial crisis during 2005-2006. \( \text{Crisis}_{t} \) is a dummy variable for the years 2008 and 2009. Bank Controls include the following lagged variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total asset, tier-1 capital over total assets, and total non-performing loans over lagged total assets. Standard errors clustered at the bank level are reported in parentheses.

The Dealscan data are practical for understanding the decline in lending during the financial crisis of 2007 to 2009. Ivashina and Scharfstein 2010 provide evidence that banks with connections to Lehman Brothers, the investment bank that failed on September 15, 2008, was positively related to credit cutbacks. The mechanism proposed is that a bank with larger syndicates with Lehman produced more uncertainty around the financial condition of that bank, as the bank’s assets could be tightly linked and similar in quality. In Table 11, I run equation (5) during a smaller window of
time to incorporate the financial crisis, 2004 to 2009. I interact the bank-uncertainty measure with *LehmanConnection*\(_b\)*, the percentage of loans bank \(b\) syndicated with Lehman Brothers before the crisis from January 2005 to December 2006, and also a dummy variable for the crisis in years 2008 and 2009. Column 3 provides the strictest specification with both firm fixed effects, bank fixed effects, and bank controls. The coefficient of *BankUncert*\(_{b,t-1}\) remains negative and significant similar to Table 10, and yet, the reduction in credit from higher uncertainty during the crisis is larger as seen from the negative and significant coefficient on *BankUncert*\(_{b,t-1} \times Crisis*\(_t\). The reduction in credit from higher bank uncertainty is larger during the crisis for banks with tighter connections to Lehman Brothers, as seen in the significantly negative coefficient on the triple interaction term.

In general, Dealscan allows for an analysis at the loan level while controlling for firm demand and new issuances. While this exercise includes only a subset of banks and firms, the loan level evidence remains suggestive of active bank management of uncertainty through lower amounts of new loan issuances.

### 5.5 Robustness

While the dictionary produced using the word embeddings could lead to ambiguous words possibly unrelated to uncertainty, I recompute the frequency of uncertainty words using the simple approach where \(S_{uncertainty} = \{"uncertainty" \text{ and } "uncertain"\}\). Table A2 in the appendix shows results of the regressions of volatility, returns, lending, and trading. While counting only the words “uncertainty” and “uncertainty” to proxy uncertainty simplifies the process of selecting a dictionary, there is no clear indication that this approach captures uncertainty or a signal of future profitability. The insignificance of *BankUncert*\(_{b,t}\) in column 1, where the dependent variable is three-day post call volatility and the large coefficient in column 2, where the dependent variable is returns, suggests this measure of uncertainty may proxy bad news. The lagged variable is slightly significant in predicting lending and positively significant for trading. The slight similarity of this parsimonious method of using a simple dictionary of “uncertainty” and “uncertain” compared with using the word embeddings approach bolsters my confidence that choosing a dictionary based on machine

\(^9\)Due to the low number of observations and to preserve the degrees of freedom, this table cannot include the several loan controls used in Table 10.
learning and natural language processing techniques could provide for a more scientific approach to dictionary-based textual analysis.

6 Conclusions

In this paper I propose a complimentary measure to existing uncertainty variables by using speech to quantify bank level uncertainty. Using bank earnings call transcripts, I generate a new list of uncertainty terms that is not based on existing dictionaries, but rather generated from the transcripts themselves to better capture semantic and syntactic similarities to the word “uncertainty.” The measure of bank uncertainty counts the frequency of uncertainty terms within a given conference call, filtering out aggregate uncertainty, such as growth in the term structure of corporate bond yields, VIX, and the economic policy uncertainty index. Banks communicating larger uncertainty decrease future lending while simultaneously increasing trading. By developing a measure at the bank level, my analyses include not only bank fixed effects and important lagged variables but also time fixed effects to control for observed and unobserved time-varying characteristics. The results suggest that this new measure proxies well the idiosyncratic uncertainty facing a bank.

It is my hope that this new measure can be used to track the flow of credit to the real economy by identifying banks most likely to cut back credit during the business cycle and deviate to more liquid assets such as trading. The new bank level measure could provide another layer of transparency to an otherwise opaque financial industry.
References


Buch, C. M., Buchholz, M., and Tonzer, L. Uncertainty and international banking. In *Conference paper for the conference on International Banking: Microfoundations and Macroeconomic...*
Implications organized by the International Monetary Fund (IMF) and the Dutch National Bank (DNB), 2014.


Demers, E. A. and Vega, C. The impact of credibility on the pricing of managerial textual content. 2014.


# Appendix

## 7.1 Appendix Tables and Figures

Table A1: Description of Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \sigma_B )</td>
<td>Reuters</td>
<td>TF-IDF weighted count of uncertainty words in earnings conference call transcripts</td>
</tr>
<tr>
<td>Interest Rate</td>
<td>FR-Y9C</td>
<td>Interest Rate Exposure (BHCK8757) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Exposure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lending</td>
<td>FR-Y9C</td>
<td>Total Loans and Leases (BHCK2122) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>(Loan Level)</td>
<td>Dealscan</td>
<td>Logarithm of Bank Allocation multiplied by Facility Amount</td>
</tr>
<tr>
<td>Real Estate Loans</td>
<td>FR-Y9C</td>
<td>Loans Secured by Real Estate (BHCK1410) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Returns</td>
<td>CRSP</td>
<td>3-Day Cumulative Absolute Returns from 3-Factor CAPM Model</td>
</tr>
<tr>
<td>Trading Assets</td>
<td>FR-Y9C</td>
<td>Available-for-Sale Securities (BHCK1773) + Held-to-Maturity (BHCK1754) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Volatility</td>
<td>CRSP</td>
<td>3-Day Volatility of Excess Returns from 3-Factor CAPM Model</td>
</tr>
<tr>
<td><strong>Independent Variables and Controls</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( BankUncert )</td>
<td>Reuters</td>
<td>Residual of ( \sigma_B ) projected on 90-day growth in VIX, EPU and 10-year minus 2-year</td>
</tr>
<tr>
<td>Equity</td>
<td>FR-Y9C</td>
<td>Total Equity Capital (BHCP3210) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>( HighAggVol )</td>
<td>CBOE</td>
<td>Dummy variable equal to 1 if VIX growth above median value between 2002-2017</td>
</tr>
<tr>
<td>Liquidity</td>
<td>FR-Y9C</td>
<td>Noninterest-Bearing Balances, Currency, Coin (BHCK0081) + Interest-Bearing Balances in U.S. Offices (BHCK0395) + Interest-Bearing Balances in Foreign Offices (BHCK0397) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Non-Perform. Loans</td>
<td>FR-Y9C</td>
<td>Non-Performing Assets (BHCK5525+BHCK5526) over Lagged Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Profitability</td>
<td>FR-Y9C</td>
<td>Non-interest Net Income (BHCK4079) over Total Net Income (BHCK4079+BHCK4107)</td>
</tr>
<tr>
<td>Size</td>
<td>FR-Y9C</td>
<td>Logarithm of Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>Tier-1 Ratio</td>
<td>FR-Y9C</td>
<td>Tier-1 Capital (BHCK7205) over Total Assets (BHCK2170)</td>
</tr>
<tr>
<td>( TopicAttention )</td>
<td>Reuters</td>
<td>Percentage of Transcripts devoted to topic as estimated by a 60-topic topic model</td>
</tr>
<tr>
<td>( TopicUncert )</td>
<td>Reuters</td>
<td>( BankUncert ) measured with only portion of call devoted to particular topic using 60-topic topic model</td>
</tr>
</tbody>
</table>
Table A2: Bank Level Uncertainty with the Uncertainty Word List, $S_{uncertainty}$, equal to “[“uncertainty”, “uncertain”]"

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Volatility$_{b,t:t+3}$</td>
<td>Returns$_{b,t:t+3}$</td>
<td>Loans$_{b,t}$</td>
<td>Trading$_{b,t}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BankUncert$_{b,t}$</td>
<td>0.0711</td>
<td>-0.1539</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.07)</td>
<td>(0.13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BankUncert$_{b,t-1}$</td>
<td></td>
<td></td>
<td>-0.0020*</td>
<td>-0.0020*</td>
<td>-0.0020*</td>
<td>0.0027**</td>
<td>0.0027**</td>
<td>0.0027**</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>BankUncert$<em>{b,t-1}$ *Equity$</em>{b,t-1}$</td>
<td></td>
<td>-0.0019*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BankUncert$<em>{b,t-1}$ *HighAggVol$</em>{t-1}$</td>
<td></td>
<td></td>
<td>-0.0005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00)</td>
</tr>
<tr>
<td>$N$</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
<td>2,732</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.4639</td>
<td>0.1019</td>
<td>0.3032</td>
<td>0.3041</td>
<td>0.3033</td>
<td>0.3064</td>
<td>0.3067</td>
<td>0.3065</td>
</tr>
<tr>
<td>Bank Controls</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Bank FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Quarter FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

Note: This table reports the effect of bank uncertainty on bank variables. Column 1 shows results where volatility is the standard deviation of excess returns, as measured through a 3-factor Capital Asset Pricing Model, for $t+1$ to $t+3$ days after the conference call, while column 2 is the excess returns three days after the conference call. Lending, shown in columns 3-5, is total quarterly commercial and industrial, agricultural, consumer, and foreign loans normalized by total assets. Trading, shown in columns 6-8, is the available-for-sale and hold-until-maturity portfolios normalized by total assets. BankUncert$_{b,t}$ is the frequency count of “uncertainty” and “uncertain” in a bank’s quarterly earnings conference call after filtering out the effects of aggregate uncertainty changes at quarter $t$. Equity is total equity capital over total assets HighAggVol is a dummy variable equal to 1 when the growth in the VIX index is above the median level in the sample and 0 otherwise. Coefficients of Equity$_{b,t-1}$ and HighAggVol$_{t-1}$ are estimated but not reported. Bank Controls include the following lagged variables: log assets, total equity capital over total assets, non-interest net income over total net income, interest and non-interest bearing balances over total asset, tier-1 capital over total assets, and total non-performing loans over lagged total assets. Standard errors double clustered at the bank and quarter level reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. 


7.2 Why Might Uncertainty Affect Bank Balance Sheet?

In this section, I illustrate reasons a bank would respond to increased idiosyncratic uncertainty by changing their balance sheet composition. I start with a stylized framework, based on Buch et al. 2014, as a useful foundation for establishing testable hypotheses in the following sections. Then, I discuss other possible mechanisms at play.

7.3 Stylized Example

A risk-neutral bank $b$ must decide at time $t$ the proportion of its assets to invest into lending, $\alpha_t$, with the remainder $1 - \alpha_t$ invested into bonds. Bonds return 1 with zero risk, while the return on lending, $r^L_{t+1}$ is random. Macroeconomic forecasts suggest a prior over the return on lending such that $r^L_{t+1} \sim N(\mu, \sigma_B^2)$, where $\sigma_B^2$ represents aggregate uncertainty. In addition, the bank guides investors by signaling its own beliefs on the uncertainty of $r^L_{t+1}$ in a conference call, by providing another signal of $r^L_{t+1}$:

$$r^L_{t+1} \sim N(\mu, \sigma_B^2)$$

The speech during the conference call provides a new variance on the expected return of lending. $\sigma_B^2$ represents the bank level uncertainty. Importantly, $\sigma_B^2$ is composed of an aggregate component and an idiosyncratic component. Because conference calls contain important insights into the profitability of the firm, I assume $\sigma_B^2$ is increasing in the idiosyncratic component (Roychowdhury and Sletten 2012, Mayew 2008, Hassan et al. 2016). I abstract from strategic communication because the empirical analysis below demonstrates that banks act as if they relay truthful information in $\sigma_B^2$. Thus, investors use Bayesian updating to form a posterior of $r^L_{t+1}$:

$$\tilde{r}^L_{t+1} = r^L_{t+1} | r^L_{t+1} \sim N(\mu, \tilde{\sigma}^2)$$

with the posterior variance $\tilde{\sigma}^2 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_\nu^2}$. The expected shareholder capital, $k_{t+1}$, can be written as:

$$k_{t+1} = \alpha_t (1 + \tilde{r}^L_{t+1}) + (1 - \alpha_t) - d_t = 1 - d_t + \alpha_t \tilde{r}^L_{t+1} = k_t + \alpha_t \tilde{r}^L_{t+1}$$

In an analysis of non-financial firms, Demers and Vega 2014 find managers use uncertain language to truthfully convey the uncertain prospects of their firm. For an analysis of cheap talk in banking, see Ahnert and Nelson 2016.
Banks seek to maximize shareholder capital tomorrow $k_{t+1}$. As a result, banks want to choose $\alpha$ as high as possible. However, a Value-at-Risk (VaR) constraint limits the amount the bank can invest in the risky technology, lending. The VaR constraint can be interpreted as limiting the probability that equity capital tomorrow is negative:

$$Pr(\tilde{r}_{t+1}^L < -\frac{k_t}{\alpha_t}) = 1 - p$$

Regulators also have information on the distribution of $\tilde{r}_{t+1}^L$, so the constraint can be rewritten as a function of the first and second moments\(^{11}\):

$$Pr(\tilde{r}_{t+1}^L < \mu - \phi \frac{\sigma_B \sigma}{\sqrt{\sigma_B^2 + \sigma_B^2}}) = 1 - p$$

I assume regulators are more conservative in their estimates of $\tilde{r}_{t+1}^L$, making $-\frac{k_t}{\alpha_t} \leq \mu - \phi \frac{\sigma_B \sigma}{\sqrt{\sigma_B^2 + \sigma_B^2}}$.\(^{12}\) Because the bank wants to set $\alpha$ as high as possible, the inequality will bind. Equating the moments and the ratio of capital to lending leads to the optimal level of lending:

$$\alpha_t = \max \left\{ 0, \min \left\{ 1, \frac{k_t}{\phi \frac{\sigma_B \sigma}{\sqrt{\sigma_B^2 + \sigma_B^2}} - \mu} \right\} \right\}$$

### 7.4 Hypotheses

First, I derive how the posterior variance $\tilde{\sigma}^2$ is affected by increases to uncertainty $\sigma_B^2$:

$$\frac{d\tilde{\sigma}^2}{d\sigma_B} = \frac{2\sigma_B^4 \sigma}{(\sigma_B^2 + \sigma_B^2)^4} > 0$$

Thus, in order to validate a measure of bank level uncertainty, a positive relationship to post-call volatility would be reassuring as the posterior variance increases in $\sigma_B$. Next, I derive how optimal

\(^{11}\)I assume the banks assets are marketable. Note that the banks I study in the empirical analysis are all publically traded.

\(^{12}\)Fahlenbrach and Stulz [2011] find during the financial crisis, banks took large risks predominantly because shareholders wanted to.
lending $\alpha_t$ changes depending on bank level uncertainty.

$$\frac{d\alpha}{d\sigma_B} = -\frac{\sigma^3 k_t \phi}{\sqrt{\sigma^2 + \sigma^2 B (\sigma^2 B \phi - \mu \sqrt{\sigma^2 + \sigma^2 B})^2}} < 0$$

As risk increases through a larger variance, banks actively manage their balance sheet compositions in favor of less risky investments in lending. The effects are more severe for banks with higher capital, as seen by differentiating with respect to $k_t$:

$$\frac{d^2\alpha}{d\sigma_B dk} = -\frac{\sigma^3 \phi}{\sqrt{\sigma^2 + \sigma^2 B (\sigma^2 B \phi - \mu \sqrt{\sigma^2 + \sigma^2 B})^2}} < 0$$

Lastly, active management of bank level uncertainty is influenced by the level of aggregate uncertainty $\sigma^\nu$.

$$\frac{d^2\alpha}{d\sigma_B d\sigma^\nu} = \frac{\sigma^2 \sigma^2 B k_t \phi (\sigma^\nu \sigma^2 B \phi - 3 \mu \sqrt{\sigma^2 + \sigma^2 B})}{(\sigma^2 + \sigma^2 B)^{3/2} (\mu \sqrt{\sigma^2 + \sigma^2 B - \sigma^V B \phi})^3}$$

The cross-partial derivative will be negative as long as $\frac{\sigma^\nu \sigma^\mu}{\sqrt{\sigma^2 + \sigma^2 B}} > 3 \mu$. If the language in the conference call relays significant uncertainty (such that the posterior standard deviation is over 3 times larger than the prior mean), then during times of high macroeconomic uncertainty banks will reduce lending more from responses to idiosyncratic uncertainty.

The empirical predictions of the stylized model are as follows. Any measure of idiosyncratic uncertainty, given a level of investment in lending and equity capital, increases the volatility of the returns. Second, the optimal level of investment in lending is decreasing in a bank’s idiosyncratic uncertainty. The decline in lending as a result of higher idiosyncratic uncertainty is more severe for banks with larger equity capital funding and periods in which aggregate uncertainty is high.

### 7.4.1 Alternative Explanation

Risk-neutrality serves as a convenient benchmark for understanding the decisions of the bank as a whole. However, at the individual level, bank managers may very well exhibit risk aversion, for example through lending to only favorable clients or prioritizing above all else shareholder desire to avoid bankruptcy (Froot and Stein 1998). As a result, another explanation for the reduction of lending in response to higher uncertainty could be that bank managers reduce lending when their
signals of $r_{t+1}$ are noisy because of risk-averse preferences. This reduction in lending would reduce cash flows for investors and share price, resulting in higher volatility. While the empirical results in this paper illustrate the effects of higher idiosyncratic uncertainty on balance sheet compositions, I am agnostic of the mechanism and use the hypotheses presented here as guidance for the empirical section.
7.5 Detailed Explanation of Mikolov et al. 2013a

Figure A1: Skip Gram Model

This figure illustrates the Skip Gram model proposed by Mikolov et al. 2013a using one training example: (uncertainty, think). The Input Layer is a one-hot encoded vector of length $V$, the number of unique words across all documents. A one is placed at the index of word uncertainty and a zero elsewhere. Using an $H$-by-$V$ matrix, $U$, the Input is transformed into an $H$ dimension vector in the Hidden Layer, $u_{\text{uncertainty}}$. Lastly, $u_{\text{uncertainty}}$ is transformed back into a vector of length $V$ using a $V$-by-$H$ matrix $V$. The output vector is normalized using the softmax function to create $y_o$. The errors, the difference between $y_o$ and the one-hot encoded vector of the target words in the context, are then used to update the weight matrices $U$ and $V$ using gradient descent.

The Skip Gram model takes as input a word represented as a one-hot encoded vector\(^\text{13}\). The output is the words in the context of $w_i$. The context is the $M$ words to the left and $M$ words to the right of $w_i$. Let $w_{i+m}$ represent the word which lies $|m|$ words to the left (right) of word $w_i$ for $m < 0$ ($m > 0$) in the original document.

The transformation of the input to the output happens in two steps. Let $x_{w_i}$ be a one-hot representation of word $w_i$ and $V$ the total number of unique words across all documents. First, $x_{w_i}$ is projected onto an $H$-Dimensional space with $U$, an $H$-by-$V$ matrix, to create $u_{w_i}$. Second, $u_{w_i}$ is projected back onto a $V$ dimensional space using a $V$-by-$H$ matrix $V$. To obtain a probability

\(^{13}\)A one-hot encoded vector for word $w_i$ is a vector of length $V$ - the number of unique words across all documents. A value of 1 is placed at the index of $w_i$ and 0 elsewhere.
distribution over all $V$ words in the vocabulary, the softmax function\(^{14}\) is applied to the resulting vector. The neural network, illustrated in the appendix in Figure A1 with a simplified training example, can be summarized by the following equations:

\[
\begin{align*}
\text{Input} & \quad x_{w_i} \\
\text{Hidden Layer (Word Embedding)} & \quad u_{w_i} = U x_{w_i} \\
\text{Output} & \quad x_o = V u_{w_i} = \left[ v'_1 u_{w_i} \quad v'_2 u_{w_i} \ldots \quad v'_{V} u_{w_i} \right]' \\
\text{Output Probabilities} & \quad y_o = \text{softmax}(x_o)
\end{align*}
\]

$y_o$ will be a distribution over all $V$ terms and will be the same output for all $2 \times M$ words in the context of $w_i$.

The important object of the Skip Gram model is the matrix $U$. $U$ provides the word embeddings where each word will be represented as a vector in $\mathbb{R}^H$. Because $U$ is an $H$–by–$V$ matrix, the word embedding of word $w_i$ will be the column in $U$ pertaining to $w_i$. $V$ can also be interpreted as another space of word embeddings, where each row $i$ represents the word embedding of word $w_i$ with different valued elements. However, the semantic and syntactic differences across words are still preserved, hence typically only $U$ is used for the word embeddings.

The model is trained by finding $U$ and $V$ such that the average probability of context words is maximized. The objective function to maximize is

\[
\frac{1}{Z} \sum_{i=1}^{Z} \sum_{-M \leq m \leq M, m \neq 0} \log(p(w_{i+m}|w_i)),
\]

where $Z$ is the total number of words across all documents and $p(w_{i+m}|w_i)$ is the element in $y^o$ pertaining to word $w_{i+m}$. Taking just one word $w_i$, we can change the objective function to a loss function to minimize:

\[
E = - \sum_{-M \leq m \leq M, m \neq 0} \log(p(w_{i+m}|w_i)) = - \sum_{-M \leq m \leq M, m \neq 0} \log\frac{\exp(v'_{w_{i+m}} u_{w_i})}{\sum_{j'=1}^{V} \exp(v'_{j} u_{w_i})}
\]

\[
= - \sum_{-M \leq m \leq M, m \neq 0} v'_{w_{i+m}} u_{w_i} + \log\left(\sum_{j'=1}^{V} \exp(v'_{j} u_{w_i})\right)
\]

\(^{14}\)The softmax function maps a vector of $K$ elements to a range of $[0,1]$ with elements summing to 1. If $z_i$ is element $i$ of vector $\mathbf{z}$, the softmax function is given by $\text{softmax}(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{k=1}^{K} \exp(z_k)}$. 

44
where \(v_{w_{i+m}}\) is the row of \(V\) pertaining to \(w_{i+m}\).

Training the model requires obtaining optimal estimates of \(U\) and \(V\) through gradient descent. Basically, given some values of \(U\) and \(V\), we observe how far off the current estimates of the probabilities for the words in the context of \(w_i\) were. Then, depending on the error we move around the elements of \(U\) and \(V\) such that the error improves.

More formally, first, every word in every sentence is fitted with the model to obtain the prediction \(y_o\). Second, the errors are calculated using the gradient of \(E\) with respect to \(U\) and \(V\) to update the two matrices:

\[
V^{\text{new}} = V^{\text{old}} - \alpha \frac{\partial E}{\partial V^{\text{old}}}
\]
\[
U^{\text{new}} = U^{\text{old}} - \alpha \frac{\partial E}{\partial U^{\text{old}}}
\]

While normal gradient descent could be applied to discover optimal parameters, Mikolov et al. 2013b discuss significant improvements to training with respect to sampling for parameter updates. Mikolov et al. 2013b address the computational complexity of using basic gradient descent by describing two techniques known as negative sampling and subsampling of frequent words.

For each training example, the output will be a sparse vector with 1 only at the index of the context word and 0 in the tens of thousands of indices which are not in the context. Negative sampling means only updating the weights of a sample of the words (columns) in \(V\) which should output a 0, or negative sample.\(^{15}\) The positive word, the target word in the context whose output should be 1, is also updated in \(V\). The sampling distribution for the negative words is given by:

\[
P(w_i) = \frac{f(w_i)^{0.5}}{\sum_v f(w_v)^{0.5}}
\]

where \(f(w_i)\) is the frequency of word \(w_i\).\(^{16}\) Mikolov et al. suggest choosing 5-20 negative samples so only 0.04 percent of the millions of weights need to be updated for each training example.

Another improvement to the computational speed of estimating the neural network is subsampling

\(^{15}\)Note the update in \(U\) will only be the word embedding of the input word \(w_i\). This is evident when looking at the update function of \(U^{\text{new}}: \frac{\partial E}{\partial U^{\text{old}}} = V^T[x^o \circ \sigma']^T x^h x^i T\). The input vector \(x^iT\) is a sparse vector with a 1 only at the index of word \(w_i\).

\(^{16}\)The authors mention trying different functional forms of the sampling distribution and find this one to perform the best.
of frequent words. The Skip Gram model can learn much from the co-occurrence of words such as terrorism and uncertainty, as these are relatively infrequent. However, words such as the are relatively uninformative as a multitude of words can precede or follow them. Thus, each input word in the training set are kept with probability $P(w_i) = \sqrt{\frac{t}{f(w_i)}}$, where $t$ is a chosen threshold, typically 0.001.$^{17}$

$^{17}$This functional form was chosen by Mikolov et al. because it attributes higher probabilities to more frequent words while preserving the ranking of the frequencies. The C-code for estimation provided by the authors, however, uses $P(w_i) = (\sqrt{\frac{f(w_i)}{t}} + 1) \frac{1}{f(w_i)}$, which is a slightly more convex form of the initial suggestion.
7.6 t-Distributed Stochastic Neighbor Embedding (Maaten and Hinton 2008)

The goal of t-SNE is to map a set of $V$ vectors, $X = x_1, x_2, ..., x_v$, from $\mathbb{R}^M$ into another set of vectors, $Y = y_1, y_2, ..., y_v$, from $\mathbb{R}^N$ such that $N < M$. First, distances are measured between each vector in $X$ as probabilities. This is done by centering a Gaussian distribution at each vector, $x_i$ to compute a probability for every other vector in $X$. That is, the probability/distance measure is calculated as follows:

$$p_{ij} = \frac{p_{ji} + p_{ij}}{2V} \quad \text{where} \quad p_{ji} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{j' \neq i} \exp(-||x_i - x_{j'}||^2/2\sigma_i^2)}$$

Each $p_{ij}$ provides a measure of probability between points which is proportional to their similarity. If two points are close together in $\mathbb{R}^M$, then the probability is high. t-SNE seeks to find the set $Y$ whose distances are similar to all $p_{ij}$s. The distances in $\mathbb{R}^N$ are calculated similarly, except the t-distribution density is used.

$$q_{ij} = \frac{q_{ji} + q_{ij}}{2V} \quad \text{where} \quad q_{ji} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_k \sum_{l' \neq k} (1 + ||y_l - y_k||^2)^{-1}}$$

Ideally, the $p_{ij}$s and $q_{ij}$s will be similar to each other. If so, then distances in the high-dimensional space would be preserved in the low-dimensional space. As a result, the objective function to minimize is the divergence between the $p$ and $q$ distributions, which is commonly computed as the Kullback-Leibler divergence:

$$KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

The Kullback-Leibler divergence is well suited for this task as large penalties are produced when $p_{ij}$ is large and $q_{ij}$ is small. Thus, it will aim to associate large $ps$ with large $qs$, hence the difference between PCA and t-SNE. t-SNE will seek to preserve vectors close together (high $ps$), while PCA preserves vectors further apart.

t-SNE is then estimated via gradient descent. The points in $Y$ will move around until the Kullback-Leibler is sufficiently small. Maaten and Hinton 2008 provide more details about the estimation and algorithm for optimal performance.