Dynamic Interpretation of Emerging Risks in the Financial Sector

Kathleen Weiss Hanley and Gerard Hoberg *

October 15, 2018

ABSTRACT

We use computational linguistics to develop a dynamic, interpretable methodology that can detect emerging risks in the financial sector. Our model can predict heightened risk exposures as early as mid 2005, well in advance of the 2008 financial crisis. Risks related to real estate, prepayment, and commercial paper are elevated. Individual bank exposure strongly predicts returns, bank failure and return volatility. We also document a rise in market instability since 2014 related to sources of funding and mergers and acquisitions. Overall, our model predicts the build-up of emerging risk in the financial system and bank-specific exposures in a timely fashion.

*Lehigh University and The University of Southern California Marshall School of Business, respectively. Hanley can be reached at kwh315@lehigh.edu. Hoberg can be reached at hoberg@marshall.usc.edu. We thank the National Science Foundation for generously funding this research (grant #1449578) and Christopher Ball for providing extensive support regarding our use of the metaHeuristica software platform and advice on the computational linguistic methods. We appreciate comments from Tobias Adrian, Allen Berger, Harry DeAngelo, Greg Duffee, Robert Engle, Paul Glasserman, Hugh Kim, Naveen Khanna, Tse-Chun Lin, Seung Lee, Andrew Lo, Frank Olken, Raluca Roman, Maria Zemankova, participants at the Macro Financial Modeling Winter 2017 Meeting, the MIT Golub Center Third Annual Conference, the 28th Annual Conference on Financial Economics and Accounting, the Midwest Finance Association, Scientific Network: Textual Analysis in Economics and Finance (Mannheim), the 2018 Jacob Levy Equity Management Center Conference, and seminar participants at City University of Hong Kong, London Business School, Michigan State University, Office of Financial Research, Pennsylvania State University-Harrisburg, Seoul National Unversity, UC Davis, University of Cambridge, University of Calgary, University of Delaware, University of Georgia, University of Miami, University of Minnesota, University of Rochester, University of South Carolina, and Warwick Business School for excellent comments and suggestions.

> Fundamentally, in a system where knowledge of the relevant facts is dispersed among many people, prices can act to coordinate the separate actions of different people ...The whole acts as one market, not because any of its members survey the whole field, but because their limited individual fields of vision sufficiently overlap so that through many intermediaries the relevant information is communicated to all.
>
> Hayek (1945)

I Introduction

The events of the global financial crisis of 2008 revealed the need for regulators to understand the sources of emerging risks in the financial sector in order to intervene before they contribute to financial instability. In response, the academic literature has proposed a number of quantitative measures to monitor the escalation of risk in the financial system. Bisias, Flood, Lo, and Valavanis (2012) provide a survey of over 30 risk metrics and this list has continued to grow since its publication. Examples include macroeconomic indicators, network measures, illiquidity and solvency metrics, and the probability of financial distress, to name a few (see also Giglio, Kelly, and Pruitt (2016)). The large number of proposed methods to determine potential risks is due to the fact that there are many ways of defining risk in a complex financial system. Indeed, the authors argue that "a robust framework for monitoring and managing financial stability must incorporate both a diversity of perspectives and a continuous process for re-evaluating the evolving structure of the financial system and adapting systemic risk measures to these changes."[1]

This is a challenging endeavor for two reasons. First, many quantitative metrics, such as network analysis (Billio, Getmansky, Lo, and Pelizzon (2012), Allen, Babus, and Carletti (2012) and Elliot, Golub, and Jackson (2014)), principal components (Kritzman, Li, Page, and Rigobon (2011)), and bank risk exposure (Adrian and Brunnermeier (2016) and Acharya, Pedersen, Philippon, and Richardson (2012)), may indicate the build-up of risk system-wide but do not provide information on economic channels. Second, other measures that do examine specific channels such as liquidity mismatch (Brunnermeier, Gorton, and Krishnamurthy (2014)), housing sector risk (Khandani, Lo, and Merton (2012)) and consumer credit (Khandani, Kim, and Lo (2010)) are often backward-looking and the risks

[1]Despite the plethora of research since the financial crisis, a single definition of systemic risk is still not widely accepted. Bisias, Flood, Lo, and Valavanis (2012) acknowledge "the truism that 'one cannot manage what one does not measure' is especially compelling for financial stability since policymakers, regulators, academics, and practitioners have yet to reach a consensus on how to define 'systemic risk'." In general, systemic risk has the property of causing severe instability across a number of institutions. Although the risks we seek to identify could affect the financial system as a whole, they may not cause a systemic event if regulators and industry participants engage in activities to mitigate their impact. Thus, we are careful not to term our emerging risks as systemic. For another discussion of the many ways academics and regulators have defined systemic risk, see also (Schwarcz 2008).

associated with the next crisis might not be similar to past events.

We propose a new approach to detect emerging risks in the financial sector that uses big data to crowdsource information from both banks and investors. For banks, we use computational linguistics applied to annual bank 10-K filings. For investors, we use daily stock returns and estimate the bank-pair covariance matrix. In order for a risk to be considered as emerging, we require three conditions to be met: (1) the risk is pervasively disclosed by a large number of banks,[2] (2) investor trading patterns indicate abnormalities in the covariance matrix relative to past quarters, and (3) our model indicates that these covariance abnormalities are significantly related to banks' common risk disclosures.

We process bank risk disclosures using two text analytic methods in tandem. First, we run Latent Dirichlet Allocation (LDA) separately for each year of our sample. Similar to principal components analysis for numerical data, LDA identifies a small number of verbal themes that best explain the variation in text across our sample. This step limits detection to include only those risks that are systematically present and thus relevant to many banks.[3] Second, we use semantic vector analysis (SVA), a method based on neural networks that detects semantic relatedness, to convert the output from LDA into a set of interpretable risk factors that are stable over time.

On the investor side, a number of academic papers show that investors both aggregate and produce information (Hayek (1945), Grossman and Stiglitz (1980), Glosten and Milgrom (1985), Kyle (1985), Chen, Goldstein, and Jiang (2007), and Bond, Edmans, and Goldstein (2012)) that influences the co-movement of asset prices (Veldkamp (2006)). Our framework requires that at least some fraction of investors see an opportunity to profitably trade the most highly exposed banks in advance of a risk becoming manifest.[4]

We use the covariance in asset returns to measure the commonality of risk exposure between banks, which is motivated by Veldkamp (2006), who argues that information pro-

[2]Note that even though there is a debate in the literature regarding the optimal level of bank disclosure opacity (see Diamond and Dybvig (1983), Gorton and Pennacchi (1990), Flannery, Kwan, and Nimalendran (2013), Gorton and Ordonez (2014) and Dang, Gorton, Holstrom, and Ordonez (2016)), we note that mandated risk disclosure allows us to identify bank-specific levels of bank disclosure. We find that these disclosures are highly informative.

[3]As evidence that these disclosures are informative, Campbell, Chen, Dhaliwal, Lu, and Steele (2014) document a positive association between specific risk factor disclosures in the 10-K and ex ante measures of risk exposures. In a related context, Kravet and Muslu (2013) find that industry-wide risk factors in the 10-K are more important than firm-specific disclosures in explaining post-filing return volatility and trading volume.

[4]In the context of banking, Bui, Lin, and Lin (2016) find that changes in short interest can predict banks' stock returns during crises. Also supportive of this framework, Peristian, Morgan, and Savino (2010) find that the market can distinguish *ex ante* between banks that did and did not have a capital gap before stress test results are released.

duction in one asset can affect the prices of related assets, causing them to covary.[5] A key econometric contribution of our study is to determine when risk is emerging in each quarter by examining the link between the candidate risk factors and the stock return covariance matrix computed using daily stock returns in the given quarter. This methodology allows us to examine if investors are potentially trading on the candidate risks, indicating they are suddenly value-relevant, which could be an indication that these risks may become manifest in the near future.

It is theoretically possible that observing bank disclosures alone (assuming perfect disclosure) or investor trading alone might be adequate to identify emerging risks. However, robustness tests around the financial crisis strongly indicate that neither produces strong evidence of emerging risks until it is too late. As we explain in detail later regarding bank disclosures alone, this is because the likelihood of false inference is high if both banks and investor signals are not jointly evaluated. For example, new disclosure requirements can change bank disclosures in a systematic way even when material risks are not present. Our approach screens out such false positives because we further require that the given risk explains investor trading behavior, making it timely and value-relevant. Information from investors alone also cannot identify emerging risks because bank disclosures are necessary to provide the interpretable risk exposures.

It is also important to note that our research program was not influenced by data-mining or specification-search.[6] To the best of our knowledge, this type of risk assessment has not been possible using pre-existing methods. Our study also provides researchers and regulators alike a new tool to monitor and evaluate risks to the financial system.

Our first main contribution is an aggregate measure of emerging risks in each quarter. This measure, displayed in Figure 1, shows the extent to which the aggregate risk exposure of the financial sector is building. Note that the level of the index becomes highly elevated in the second quarter of 2005, far in advance of the financial crisis and reaches its peak by the fourth quarter of 2006. In contrast, other indicators of emerging risk such as VIX or aggregate volatility, do not become elevated until the crisis begins in 2008. Equally important, Figure 1 shows that the aggregate emerging risk score does not increase during other episodes of market volatility that did not severely affect financial institutions. Two

[5]Commonality in returns are used as a measure of interconnectedness in a number of studies including Billio, Getmansky, Lo, and Pelizzon (2012), Brunetti, Harris, Mankad, and Michailidis (2018), and Diebold and Yılmaz (2014).

[6]Our framework is the result of a 2014 National Science Foundation grant application that was submitted before any tests were run. Importantly, the bank and investor crowdsourcing implementation we propose is the same as in the initial proposal (see National Science Foundation grant #1449578).

such examples are the bursting of the technology bubble of 2000 and the events surrounding 9/11/2001, both of which did not have serious spillovers to financial intermediaries. We conclude that our model does not produce elevated indications of risk simply when markets are volatile.

Our second major contribution is the implementation of a flexible model that uses natural language processing to identify risks. This is important because many of the risks that banks face are disclosed only through verbal discussions in the risk factor section of the 10-K and thus, models based on quantitative data alone might miss potential emerging risks. A primary benefit of our model is that it can identify emerging risks regardless of whether they are disclosed quantitatively or qualitatively. To ensure additional flexibility in the identification of risks, we develop two models that differ on the level of human intervention versus automation in selecting candidate emerging risk themes.

The first model, the "static" model, considers 31 semantic risk themes identified from the manual evaluation of LDA output that correspond to the types of financial sector risk established in the academic literature (see Appendix C for a review of the literature). The static model is informative because many determinants of financial instability are similar from one event to another (see Calomiris and Gorton (1991), Goldstein (2005), Reinhart and Rogoff (2008), Allen, Babus, and Carletti (2012), and Goldstein and Razin (2015)).[7] By examining the econometric link between individual semantic themes and covariance matrix abnormalities in each quarter, we find that themes related to real estate (Herring and Wachter (1999), Reinhart and Rogoff (2009) and Mian and Sufi (2009)), prepayment risk (Roberts and Sufi (2009)), commercial paper (Kacperczyk and Schnabl (2010) and Covitz, Liang, and Suarez (2013)), dividends (Acharya, Gujral, Kulkarni, and Shin (2011)), operational risk (Aebia, Sabatob, and Schmid (2012)) and credit cards (Mian and Sufi (2011)) are elevated as early as 2005.

A virtue of the static model is its flexibility to further query the model when a specific emerging risk has multiple potential sub-channels. As an example, consider the situation

[7]For example, many of the risks that lead to the recent global financial crisis are similar to those of the Asian financial crisis. Discussing the Asian crisis, the *Report of G7 Finance Ministers to G7 Heads of State or Government*, May 1998 states "These weaknesses included over-extended lending to the property sector, the build up of large off-balance sheet positions, excessive exposure to highly leveraged borrowers, policy directed loans and excessive reliance on short-term borrowing in foreign currency. Had information about these developments been more widely available earlier, the international markets and International Financial Institutions might have been better placed to assess the risks in Asia and elsewhere. ...There is therefore a need for strengthened mechanisms to ensure appropriate risk analysis. This points to the need for enhanced international surveillance and improved prudential standards, and to the need to encourage internationally active financial institutions to act prudently on available information."(http://www.g8.utoronto.ca/summit/1998birmingham/g7heads.htm)

where the model indicates that real estate risk is elevated in a particular quarter. Because our semantic method automatically outputs a complete vocabulary that characterizes any given risk theme, a researcher can decompose the broad topic of real estate risk into granular sub-themes such as subprime, mortgage-backed, HELOC, and foreclosure. These sub-themes can then be included in the static model to assess their relative importance. When we run the extended static model on the real estate sub-themes noted above, we find that these specific risks became elevated before the crisis period. Additionally, in a second test motivated by the sovereign debt crisis, we find that risks relating to the Eurozone, International Monetary Fund, sovereign debt, Brazil and Greece become heightened in late 2008/early 2009, just prior to the onset of the events in Europe.

We recognize, however, that the financial sector is both complex and constantly changing, posing a challenge for those who monitor risk. The limitation of the static model is that it requires the researcher to choose which risks to include. Yet if a source of risk is truly new and unexpected, the researcher would not know to enter the given theme into the static model, thereby allowing a risk to financial stability to go undetected. In order to overcome this limitation, we propose a second "dynamic model" that automates the identification of candidate risks in each year. The only input this model takes from the researcher is the removal of boilerplate bigrams to enhance computational efficiency and reduce noise. The benefit of automation is that it can detect emerging risks even if the researcher is entirely unaware of its potential relevance. This issue is critical because unexpected risks could remain undetected for long periods of time, leaving policy makers little time to react.

Our dynamic model reassuringly finds many of the same emerging risks identified by our static model. However, the model also reveals new risks outside of the static model that may have been unanticipated. An example is the theme "weather events" in 2013. Another, is the bigram "education loans" in the third quarter of 2011, around the time President Obama made two changes to the federal student loan program. Our results suggest that both models provide key insights regarding the global financial crisis and also elevated risks in more recent years.

In addition to identifying risks common to the entire financial system, our method can be used to assess the potential impact of individual banks' exposure to emerging risks in each period. We find that banks with higher ex ante overall exposures to our static risks experience three ex post negative outcomes. These outcomes include more negative stock returns during the financial crisis, higher bank failure rates[8], and higher stock price volatility

[8]Other studies predicting bank failures include Sarkar and Sriram (2001), Cole and White (2011), and

lasting up to 36 months. These results are consistent with Fahlenbrach, Prilmeier, and Stulz (2012), who find that stock returns during previous crises can predict bank performance in future crises. This suggests that "some aspects of their (bank) business model could make them more sensitive to crises, but so could their risk culture."

Another key attribute of our model is that it can be used in real time. When we examine the aggregate risk index through the beginning of 2016, we document a new build-up of potential risk. The static model illustrates that risks relating to mergers and acquisitions, real estate, taxes, and short-term funding emerge strongly by early 2013. Exposure to these emerging risks also predicts bank-specific negative stock returns from December 2015 to February 2016 (when financial firms were particularly volatile). Although not all emerging risks will necessarily materialize, we believe our approach offers important insights regarding either potential vulnerabilities in the financial sector when faced with an exogenous shock (such as the LCTM collapse) or the build-up of risk within banks that could contribute to a systemic event (as occurred in the recent financial crisis).

It is important to note that since all U.S. listed firms are required to report risk factors in Item 1A of their 10-Ks, the model we propose may be used in any industry in which risk identification through time is important. For example, insurance companies, like banks, may contribute to financial instability and our method can isolate risks unique to that industry that may be important to regulators. Furthermore, the usefulness of our method is not limited to industries whose activities may affect the economy as a whole. It can also be employed by analysts to identify and track industry-specific risks. Thus, an energy analyst may be able to use the disclosures of energy companies to ascertain the potential impact and risk of fracking and shale gas production.

Our results also suggest that information about emerging risks is slow moving, and may take many months to fully impact asset prices. Theoretical models in this area assume that there are only two states of nature: normal times when there is no information production, and crisis periods that induce information production. For example, Gorton and Ordonez (2014) argue that the banking sector is more efficient when there is little or no information production on the quality of bank assets, as this economizes on information costs, and in so doing, leads to lower borrowing costs and greater economic growth. Yet opaqueness that is optimal in normal times exposes the economy to periodic crises following aggregate negative shocks to collateral values. Information production to ascertain collateral quality will then increase for a period of time until the crisis is resolved.

DeYoung and Torna (2013).

However, the path from stability to crisis is clearly not instantaneous given real world frictions. Slow information diffusion in stock returns may occur because of disagreements on the economic value of information (Miller (1977)), the existence of short sale constraints (Diamond and Verrecchia (1987)), limits to arbitrage (Shleifer and Vishny (1997)), information processing (Merton (1986)), and/or limited investor attention (Barber and Odean (2007)). Regarding how long the market takes to correct mispricing, Debondt and Thaler (1985) show that overreaction can impact stock returns for periods as long as three to five years. To the extent that bank stocks are by nature opaque (as theory predicts above), and crises are by nature unexpected, it seems plausible that the market's awareness of emerging risks can play out over a long period of time. Practically speaking, we suggest that there exist three states of information production: (1) no information production (normal period), (2) some information production as risk is building (transition period), and (3) high information production (crisis period). The existence of the transition period implies that our model can be used as an early warning system of specific risk channels that drive financial instability and can be informative regarding preemptive corrective actions.[9]

II Text Analysis and Methodology

We propose a methodology that both dynamically measures the aggregate risk exposure of the banking industry as a whole, and also the specific underlying sources of risk. It is designed to address two primary limitations of standard computational linguistic methods. First, some textual analysis approaches, while useful in certain contexts, cannot isolate underlying economic risks. These approaches include textual tonality (Tetlock (2007)) and readability (Loughran and McDonald (2014)). The negative tone of a 10-K, for example, may be related to specific bank outcomes but is not particularly helpful in understanding the source of the negativity. Second, other methods require the researcher to explicitly specify the potential emerging risks as input. An example of this approach is Baker, Bloom, and Davis (2016), who create an index of economic policy uncertainty (EPU) using a self-generated list of words. They find that the index can predict firm-level attributes such as stock price volatility and investment activities in industries that relay on government contracts. EPU, however, is very specific and cannot capture other risks that are not

[9]Our time line of emerging risks is consistent with Bussiere and Fratzscher (2006), who suggest that an early warning system that incorporates "three regimes (a tranquil regime, a pre-crisis regime, and post-crisis/recovery regime) can provide a substantial improvement in the forecasting ability" of these models. Other early warning systems include Huang, Zhou, and Zhu (2009), Giesecke and Kim (2011), Estrella and Mishkin (2016), Frankel and Saravelos (2012), and Duca and Peltonen (2013).

included in the index. This latter requirement is critical in our setting as the sources of financial instability are inherently unpredictable, and might be unknown ex ante to the researcher.

Our approach is novel and it crowdsources signals about emerging risks from both investors and banks. Regarding investors, we use daily return data and stock price comovement to uncover signals about investor information. Relating information produced through trading activity and mandated disclosures is motivated by the literature such as Bond, Edmans, and Goldstein (2012), who argue that "A financial market is a place where many speculators with different pieces of information meet to trade.....Prices aggregate these diverse pieces of information and ultimately reflect an accurate assessment of firm value." Motivated by Veldkamp (2006), we use covariance as a measure of the comovement of the industry when investors trade on information about specific risks. In the paper, she proposes a model in which news generates comovement in stock prices. In the model, a signal must have two features to produce comovement. First, it must contain information about the value of many assets and second, it must be observed by many investors. Our method encompasses both of these attributes.

Regarding banks, we use their collective risk disclosures in their 10-Ks to uncover which risks are potentially important. Our sample of 10-K's is extracted by web-crawling the Edgar database for all filings that appear as "10-K," "10-K405," "10-KSB," or "10-KSB40."[10] The document is processed for text information, fiscal year, and the central index key (CIK). Although all of the text-extraction steps outlined in this paper can be programmed using familiar languages and web-crawling techniques, we utilize text processing software provided by metaHeuristica LLC. The advantage of doing so is that the technology contains pre-built modules for fast and highly flexible querying, while also linking the queries to analytics including Latent Dirichlet Allocation and Semantic Vector Analysis, which are critical to our model's vision and implementation.[11] We use all available fiscal years in the metaHeuristica database from 1997 to 2015.

One benefit of using metaHeuristica is that the discussion of risk factors in the 10-K is difficult to extract using standard programming methods. Prior to 2005, most firms discussed risk factors in many different parts of the 10-K with heterogeneous subsection labels.

[10]Following convention, we only use the initial 10-K filed in each fiscal year, and do not consider amended 10-Ks which can be filed at a much later time.

[11]For interested readers, the metaHeuristica implementation employs "Chained Context Discovery" (see Cimiano (2006) for details). The database supports advanced querying including contextual searches, proximity searching, multi-variant phrase queries, and clustering.

Starting in 2005, risk factors became more standardly placed in Item 1A. metaHeuristica's dynamic querying tools allow us to identify and directly query sections and subsections of the 10-K having titles or headers containing the word root "risk" regardless of where they are in the 10-K.

The output from these metaHeuristica queries is the full set of paragraphs that contain discussions of risk factors for all banks in our sample from 1997 to 2015. Examples include interest rate[12], capital adequacy,[13] and mortgage loan risk[14]. Each paragraph is then linked to key identifiers including the bank's central index key (CIK), the file date of the given 10-K, the bank's fiscal year end, and the filer's SIC code. In its raw form, the text is in paragraph form and is high-dimensional (thousands of paragraphs and unique words). Such complexity makes it difficult to detect emerging risks without dimensionality reduction algorithms that we discuss next.

A Latent Dirichlet Allocation

Our first natural language processing technique is LDA, which is a dimensionality reduction algorithm used extensively in computational linguistics (see Blei, Ng, and Jordan (2003)). LDA assumes an underlying model in which each document is generated from a probability distribution over topics.[15] To understand the intuition of LDA, suppose that there are a fixed number of T topics that banks draw upon when writing their risk factors. Potential topics might include real estate risk, deposit risk, and risks relating to sources of funding. Each of these topics will have a vocabulary related to the discussion and through probabilistic modeling, LDA discovers the different topics that the documents contains and how much of each topic is present in the document.

A key virtue of LDA is that it auto-derives the set of risk exposures and requires only one input: the number of topics T to be generated. To maintain parsimony, we focus on 25 topics (although we consider 50 topics for robustness and find similar results). The choice of 25 topics reflects the expected granularity of risk factor text and is consistent with 25

[12] "In a sustained rising interest rate environment the asset yields may not match rising funding costs, which may negatively impact interest margins."

[13] " Republic's failure to maintain the status of "well-capitalized" under our regulatory framework, or "well-managed" under regulatory exam procedures, or regulatory violations, could compromise our status as a FHC and related eligibility for a streamlined review process for acquisition proposals and limit financial product diversification."

[14] "Our interest-only mortgage loans may have a higher risk of default than our fully-amortizing mortgage loans and, therefore, may be considered less valuable than other types of mortgage loans in the sales and securitization process."

[15] We provide only a summary level discussion of LDA in this paper. We refer more advanced readers who are interested learning more about LDA to the original study by Blei, Ng, and Jordan (2003), or to Appendix A in Ball, Hoberg, and Maksimovic (2016).

important risk factors being present.

LDA produces two data structures as output. The first describes the distribution of systematically important topics discussed by each bank in each year. These specific "topic loadings" for each bank-year in our sample are word vectors of length 25 representing the extent to which a bank discusses each of the 25 risk topics. This data structure thus reduces the dimensionality of 10-K risk factors to just 25.

The second data structure is a set of words and their frequencies for each topic. This data structure contains 25 topic vectors of individual word lists with corresponding word frequencies/probabilities. For parsimony, we retain the top 100 words and the top 100 commongrams for each topic. These word lists can be evaluated to determine the content of each topic.

Figure 2 displays a summary of the LDA output for our sample of banks in 2006. The figure shows that bank risk factors contain many topics that imply sensible risk disclosures by banks. These include discussions related to interest rate risk, economic conditions, real estate loan risk, regulation risk, fair value, and corporate governance. Because each of the 25 topics is simply a word list, the specific subject of the LDA topic needs to be assessed manually topic-by-topic.

To illustrate, the topic labeled "r-10" in Figure 2 is an example of a highly interpretable emerging risk, as this source of risk appears to be related to real estate loans. The list contains phrases such as "real estate," "loan portfolio," and "commercial real estate". However, some topics have unclear, blended interpretations. For example, the topic "r-08" contains phrases such as "fair value," "interest rate risk," and "financial instruments." Although any one of these items might indicate an interpretable risk factor, the blending of these terms in one topic indicates ambiguity making it difficult to assign a specific risk. In reviewing many topics, we find this ambiguous interpretation problem to be pervasive and magnified when examining many topics year-over-year.

Thus, the primary strength of LDA is its ability to identify vocabulary that is systematically present for a large number of banks. This characteristic is due to LDA's objective, which is to explain a large fraction of the corpus using few degrees of freedom (akin to principal components for numerical data). However, there are obvious limitations of LDA that impact its effectiveness in isolating specific risks. First, as we note above, the economic interpretation of individual topics is often ambiguous, as LDA does not generate a specific label for each topic. For example, if the selected number of topics is too few, or if

factors are too correlated, they will be grouped by LDA into a single risk topic, making the interpretation unclear. If, on the other hand, the selected number of topics is too many, individual risks might be split into more than one factor. Second, when LDA is run on the corpus of documents in each year, the composition of the 25 LDA topics will likely change, making it difficult (if not impossible) to track specific risks through time. Despite these weaknesses, the real strength of LDA in our setting is that topic models are very good at identifying words and bigrams that are highly present in a large number of bank risk disclosures. The content we extract from LDA therefore automatically satisfies the key criterion of being pervasive or systematically important to many banks. In the next section, we use these individual words and bigrams as inputs to a second natural language processing technique we call "Semantic Vector Analysis" (SVA). This step will preserve LDA's critical strength (limits attention to pervasive vocabulary only), while addressing its limitation (interpretability).

B Semantic Vector Analysis

The SVA algorithm draws upon research in the area of "Distributional Semantics." The intuition is that "a word is characterized by the company it keeps" (Firth (1957)). SVA is a probabilistic approach used to uncover the semantics of natural language, and it is often used in search engines to score documents even when they contain related words but not the actual word(s) used to define the search.

SVA uses a neural network to predict the distributional occurrence of each word as defined by the other words with which it normally co-occurs. The resulting model is a set of n-dimensional vectors, with a single n-vector representing each word (or in our case each bigram). The result is that similar words are grouped together in vector-space. This means that, unlike both LDA and other basic word list methods, SVA is not a bag-of-words algorithm, as the relative position of words in a document matters.

SVA is implemented in two stages: (1) a "training stage" that only needs to be run once, and (2) a "mapping stage" where specific risks are mapped to vocabularies that best represent each given risk semantically. Each SVA vocabulary has a specific label, greatly reducing ambiguity in interpretation.

The word vectors are trained using a two-layer neural network that learns the contextual use of each word based on the distribution and ordering of words in the corpus of 10-Ks. The use of 10-Ks as input to this calculation ensures that the mapping of concepts to vocabularies takes into account the general style of discussion used in 10-K regulatory

filings. The information from the 10-K is then stored in high-dimensional vectors that retain information garnered from examining semantic relatedness through the proximity of each word to other words appearing in each document. Intuitively, this is simply a mapping from each word or commongram to a specific semantic vocabulary.[16] The approach ultimately maps a single word or commongram to a vocabulary that is frequently associated with the given word or commongram.

Like the classic SAT test, SVA attempts to predict word associations, e.g. boy is to girl as man is to x? A unique example of the ability of this technique to predict word choice is in Mikolov, Chen, Corrado, and Dean (2013). Using SVA, the authors attempt the Microsoft Sentence Completion Challenge, a recently introduced task for advancing natural language processing. The challenge is very simple. The tasks consists of 1040 sentences, where one word is missing in each sentence. The goal is to select the most reasonable choice for the missing word given a list of five alternatives. The authors use Google News to train the word vectors and by feeding the phrase into SVA, can select the correct missing word with a high degree of accuracy.

SVA can also be used in reverse. In our case, we know the "missing word" and are interested in the vector of words (each having a weight indicating importance) that best represents the particular word's theme and meaning. To illustrate, we enter any word such as "governance' or a commongram such as "real+estate" as input, and SVA will return a corresponding semantic vocabulary. Table I displays the output vectors for four of our semantic themes. For example, the theme "Real Estate" is usually associated with words that include "foreclosure", "property", "lien", etc. "Regulatory Capital" is associated with "prompt corrective", "adequacy guidelines", "maintain", and "regulatory agencies". Thus, the word lists associated with each semantic theme are directly interpretable with a unique label that allows the user to examine the importance of specific risks over time.

C Mapping SVA Themes to Individual Banks

In order to determine the bank's discussion of a specific risk, we need to map the bank's text to each word vector from SVA. We do this by computing the cosine similarity between the vocabulary list associated with each SVA risk theme, and the raw text of each bank's overall risk factor disclosure. The procedure is as follows:

For each year t, suppose there are $n_{i,k,t}$ unique words in the union of bank i's risk

[16] This process is often referred to as a "word-to-vec" mapping (Mikolov, Chen, Corrado, and Dean (2013) and Mikolov, Sutskever, Chen, Corrado, and Dean (2013)).

disclosure and theme k. We represent the bank i's risk factor disclosure as a vector with $n_{i,k,t}$ elements, which we denote $W_{i,t}$. Each element is populated by the number of times bank i uses a given word in its risk factor disclosure in year t, and the vector is normalized to have a length of 1. For any word that appears in SVA theme k but not in bank i's risk disclosure, the element is set to zero.

Analogously, theme k is a vector also with $n_{i,k,t}$ elements, which we denote $T_{k,t}$. Each element of this vector contains the numerical theme loadings, as shown in Table I, for words that are part of the theme and this vector is again normalized to length 1. For any word that appears in bank i's risk disclosure but not in SVA theme k, the element is set to zero. Note that the vectors $W_{i,t}$ and $T_{k,t}$ have the same length.

We compute bank i's loading on semantic theme k in year t as $S_{i,k,t}$ as the normalized cosine distance:[17]

$$S_{i,k,t} = \frac{W_{i,t}}{||W_{i,t}||} \cdot \frac{T_{k,t}}{||T_{k,t}||} \tag{1}$$

After computing the loading for bank i for each of the semantic vectors, we have a complete panel database with one observation being a bank-year, and containing one column of bank-specific loadings for each semantic theme ($S_{i,k,t} \forall$ SVA themes k). The resulting data structure allows us to observe the intensity of every bank's discussion of each emerging risk theme in each year.

III Determination of Emerging Risks

To determine which semantic risk themes are emerging in a given quarter, we run quarterly regressions at the bank-pair level where the dependent variable is a bank-pair covariance.[18] Our bank-pair panel regressions use more information than is available using simpler bank-level regressions and allows us to detect comovement in the banking industry. This is important for maximizing power so that only system-wide emerging risks can be detected early, even when the number of investors trading based on these risks may be modest.

For each bank pair i and j, we examine the link between the pair's common risk exposures and the pair's return covariance. Our central hypothesis is that as investors produce information on an emerging risk to which both i and j are exposed, the stock return co-

[17] Cosine similarity is bounded between 0 and 1 with observations closer to one indicating greater similarity between the SVA theme and the bank's risk factor disclosure. Thus, if a particular SVA theme's cosine similarity with bank i's risk factor disclosure is close to one, this means that the bank's discussion of the theme is highly relevant.

[18] We have 76 quarterly observations that span 1998 to 2016. Although our disclosure sample is available in 1997, our covariance regressions start in 1998 because we lag all independent variables by at least one year.

variance of the pair will become abnormally high. In normal times, this relationship will be weak or non-existent. We compute pairwise loadings as the product of bank i and j's individual exposures that capture the extent to which banks i and j are jointly exposed to a given risk theme k:

$$S_{i,j,k} = S_{i,k}\ S_{j,k} \tag{2}$$

We then regress the quarterly return covariance of i and j on each of the N joint risk exposures using ex ante data from the prior fiscal year $t-1$. We also include controls for industry, size, and accounting variables using the following regression equation:[19]

$$Covariance_{i,j,t} = \alpha_0 + \beta_1 S_{i,j,t-1,1} + \beta_2 S_{i,j,t-1,2} + \beta_3 S_{i,j,t-1,3} + ... + \beta_N S_{i,j,t-1,N} + \gamma \mathbf{X_{i,j,t-1}} + \varepsilon_{i,j,t}, \tag{3}$$

This model produces N β coefficients for each of the N semantic themes, and also a set of γ coefficients for the industry and bank characteristics. These slopes are computed separately in each quarter.

In the time series analysis, we consider the R^2 from the above regression and decompose it into parts. First, we compute the R^2 attributable to the industry and accounting variables $X_{i,j,t}$ by running the regression (3) without the semantic themes:

$$Covariance_{i,j,t} = \alpha_0 + \gamma \mathbf{X_{i,j,t-1}} + \varepsilon_{i,j,t}. \tag{4}$$

We compute the marginal R^2 that is attributable solely to the semantic themes by taking the R^2 from Equation (3) and subtracting the R^2 from Equation (4). The resulting marginal R^2 is a time-series variable, as the regression is run once per quarter.

We consider two variations of the covariance model that differ in the level of human intervention versus automation needed to identify the candidate risks. Our main model is a static model that identifies core economic risks in the financial sector through a manual inspection of the entire time-series of LDA topics. The benefit of this approach is that the user can track the importance of each economic risk year-over-year and, because the model is static, one can create an aggregate risk index that can be used to identify the elevation in overall financial sector risk through time.

Furthermore, the static model can be extended to include user-identified risks. For example, a bank examiner may notice a pattern in the risks of certain banks and would

[19] We estimate pairwise control variables as the dot product of the variable for banks i and j and winsorize the covariance estimates in each quarter at the 1/99% level.

like to understand if such risks are affecting a larger number of institutions. In order to determine if this is the case, the user would simply include the risk in the static model. Another use of the user-identified risk model is to permit the user to examine a specific set of risks more granularly to better understand the manifestation of a general risk. The drawback to the static approach is twofold: First, it uses some human judgment in the choice of LDA risks to include[20] and second, it only includes risks that are previously known.

In order to reduce the amount of human intervention, we propose a dynamic model that is almost fully automated (with the exception of manual removal of boilerplate for computational efficiency and noise reduction) and updates the set of candidate risk factors each year based on salience. This model uses a larger, more granular set of possible risks as input than the static model to identify new risks. While the primary benefit of such a model is to allow previously unidentified risks to emerge, it is not designed to track the elevation of individual risks over time. Each model has benefits and limitations, but in combination, can be powerful in identifying a set of risk factors that are both interpretable and specific.

A Central Hypothesis

Our approach requires that banks produce and disclose some information on their risk exposures, and that investors produce information on the underlying state of the economy. One way to illustrate the intuition is to assume, for example, that bank returns are determined by a simple factor model. Without loss of generality, consider a single risk factor model:

$$\tilde{r}_{i,t} = \beta_i \tilde{K}_t + \tilde{\epsilon}_{i,t} \tag{5}$$

The return of bank i in quarter t has an emerging risk component common to all banks ($\tilde{K}_t$) and an idiosyncratic component ($\tilde{\epsilon}_{i,t}$). In our setting, the semantic theme loadings from SVA identify the bank's exposure to the emerging risk (β_i). Note that knowledge of this exposure is not adequate to determine the bank's outcome because information about $\tilde{K}_t$ is also needed. We propose that stock market investors produce information about $\tilde{K}_t$. An increase in R^2, therefore, emerges only when investors become informed and trade on $\tilde{K}_t$. When this occurs, the covariance matrix of stock returns $\tilde{r}_{i,t}$ becomes significantly related to emerging risk exposures. Assuming random variables are independent standard normals, the following arises in expectations regarding banks i and j:

[20] As we note below, we try to minimize the amount of human judgment needed to select the risks

$$Cov[\tilde{r}_{i,t}, \tilde{r}_{j,t}] = \beta_i \beta_j \tag{6}$$

This relationship echoes our covariance model's functional form as we regress pairwise covariance on the product of bank i and bank j's risk exposures (see Equation 4). This leads to our central hypothesis:

Central Hypothesis: When risk is building in the financial sector, a regression of pairwise covariance on the risk themes will become significant and produce an elevated R^2 in Equation 3. When no emerging risk is present, this R^2 will be close to zero.

IV Sample and Data

Our initial sample of publicly traded financial institutions has SIC codes in the 6000-6199 range and is identified using the Center for Research in Security Prices (CRSP) and Compustat databases. To be included, a bank must have an available link between its Compustat gvkey and its central index key (CIK), the unique identifier used by the Securities and Exchange Commission. The gvkey to CIK links are obtained from the SEC Analytics database. Observations must also have a machine readable 10-K risk factor discussion as identified by the metaHeuristica software. To satisfy this requirement, we query metaHeuristica for any 10-K section titles, or subsection titles, containing the word "risk" or "risks".

Our final sample contains 10,558 bank-year observations from 1997 to 2015 that satisfy these requirements. We have an average of 587 publicly traded banks per year in our sample and Figure 3 displays the composition of our sample over time. The number of banks is 583 in the first year of our sample, peaking in 1999 at 703 banks. One reason for this initial increase is that banks did not consistently disclose risk factors in the first two years of our sample, but did so more reliably after 1999. After the peak in 1999, the number of banks in our sample slowly declines to 564 by the onset of the financial crisis in 2008 and further declines steeply to 398 by the end of our sample in 2015. This reflects the well-known finding that many banks failed or were acquired in the aftermath of the crisis.

A Financial Market Variables and Bank Characteristics

Our primary financial market variable of interest is the pairwise covariance of financial firms in a given quarter. In each quarter, we calculate the daily return covariance matrix for all of the banks in the sample using data from CRSP. In order to compare the performance of our aggregate risk index, we collect information on six additional indicators of overall

market risk or uncertainty. The first measure is the cross-sectional standard deviation of monthly returns for all stocks in the CRSP database in a given quarter. The second is an analogous measure based on financial firms only. The third is the implied volatility of European-style S&P 500 index options (VIX) from Yahoo Finance. The fourth is the average quarterly pairwise covariance of banks in our sample. The fifth and sixth are two measures of economic policy uncertainty from Baker, Bloom, and Davis (2016). One is based on text-based searches of words related to policy uncertainty in 10 newspapers, and the other augments the news-based measure with measures of uncertainty based on tax code changes, CPI disagreement, and federal versus state and local purchase disagreement.[21]

We collect bank characteristics from Call Reports following the literature (Cole and White (2011) and Cornett, McNutt, Strahan, and Tehranian (2011)), which we use as control variables in our covariance model. We also separately explore the extent to which these accounting variables predict emerging risks. We aggregate Call Report data at the holding company level if the bank has a parent ID, otherwise, data is at the individual commercial bank level. In order to identify a specific bank in our data, we merge the RSSD ID in the Call Report Data with the New York Federal Reserve's list of public institutions to obtain a CRSP PERMCO. This field allows us to merge Call Report data with our sample.

We then construct a number of bank characteristics (all but *Assets* are scaled by assets). *Cash* and *CatFat* from Berger and Bouwman (2009) as measures of liquidity.[22] *Loans* and *Ln(Assets)* are used as indicators for the size of the bank. *Non-Performing Assets*, the sum of loans that are 30 days and 90 days past due, and *Loan Loss Prov & Allow*, the sum of loan loss provision and allowances capture potential problem lending. *Bank Holding Co. Dummy* is an indicator variable equal to one if the bank has a parent, zero otherwise. *Negative Earnings Dummy*, an indicator variable equal to one if net income is negative, zero otherwise, is a measure of profitability. *Capital*, the ratio of equity to assets, has been shown to predict subsequent bank performance (Berger and Bouwman (2013) and Cole and White (2011)). Finally, *Bank Age* is the time since the bank's first appearance in Compustat.

We augment the database with Compustat industry data, which is based on SIC codes, and with textual network (TNIC) industry data from Hoberg and Phillips (2016). Because our framework naturally controls for industry, these additional controls are conservative, and allow us to control for variation in product market offerings within the sample of banks

[21] We thank the authors for providing economic policy uncertainty data on their website http://www.policyuncertainty.com/. The website also provides details regarding the construction of each index.

[22] Generously provided by Christa Bouwman at https://sites.google.com/a/tamu.edu/bouwman/data.

(our results are robust to excluding these controls). Overall, the purpose of examining bank and industry characteristics is to provide an array of controls that should absorb known variation in bank-pair-quarter covariances. Hence, any emerging risk factors we find can be seen as abnormal relative to normal drivers of covariance.

B Summary Statistics

Panels A and B of Table II display summary statistics for the bank characteristics from Compustat and the Call Reports, respectively. Most of the financial institutions in our sample, 84.2%, are bank holding companies. The average bank has loans to assets of 50.7%. Loan loss provision and allowances as well as non-performing assets (NPA) are both close to zero (0.02% and 0.05%, respectively). On average, banks have a capital ratio of nearly 10% and 9.1% have negative net income.

Panel C reports statistics for our bank-pair-quarter variables. Because the number of quarterly bank pair permutations is large, there are roughly 19.7 million observations during our sample period. The panel shows that the average pair of banks, not surprisingly, has a positive covariance. Because our sample is limited to financial institutions, 86.9% are in the same two-digit, 50% in the same three-digit and 46.9% are in the same four-digit SIC code. The average TNIC pairwise similarity from Hoberg and Phillips (2016) is 0.104, indicating a material amount of product similarity among the banks in our sample. As a basis for comparison, the average pairwise similarity in the baseline TNIC database is 0.064.

V Static Risk Methodology

In the static risk model, we specify a set of high-level risk factors that remain constant over time. Our candidate SVA risk themes are derived from an examination of the most frequent words and commongrams from LDA. We identify these most frequent terms in each year and remove any boilerplate such as "balance sheet" or "million December." We group the remaining terms into broad categories of risks that are fundamental to the banking sector aided by a review of the literature (see Appendix C for a brief review of the themes identified from the literature.)[23] As an example, there are many bigrams in the top terms from LDA that we associate with the risk theme "mortgage loans" (residential mortgage, mortgage backed, mortgage servicing, etc.) and "credit cards" (card loans, card receivables, card products). After eliminating any themes that are highly correlated, we have 31 final

[23] All themes identified from the academic literature on the financial crisis are already included in the LDA topics but may not have been included in the final list of risks due to multicollinearity.

risk themes in the static model.[24] Note that these themes may not be all-inclusive but the model allows additional topics to be included if one so desires.

We begin by exploring the cross-sectional characteristics of financial institutions that are most exposed to each of the 31 static themes. The results are displayed in Table IV, where the dependent variable is a bank's loading on each of the 31 static themes and the independent variables include the accounting-based bank characteristics used in the covariance model. In the spirit of brevity, we discuss the determinants of the real estate, mortgage-backed, and regulatory capital themes. We find that banks have a higher loading on real estate risk when they are smaller and are more likely to have negative earnings. This pattern is also evident in the sub-theme related to mortgage-backed securities in Panel B. This could imply that less profitable banks are at risk for high exposure to risky loans. In contrast, larger banks, with better performing loans, more capital, negative earnings and non-performing assets are most likely to be exposed to regulatory capital risk.

A Aggregate Indicator of Emerging Risks

We next analyze whether our measures are informative in predicting the build-up of emerging risk. To focus uniquely on the contribution of the semantic themes in explaining covariance, we compute the marginal adjusted R^2 of the textual risk factors alone by subtracting the adjusted R^2 from Equation (4) from the adjusted R^2 from Equation (3). The result is an quarterly time series of marginal adjusted R^2 statistics attributable only to the semantic risk themes.

In order to then determine whether a given quarter's semantic themes R^2 is elevated, we use the initial part of our sample (1998 to 2003) as a baseline period to compute the quarterly mean and standard deviation of this time series. For all remaining quarters (2004 to 2015), we then compute the z-score for each quarterly observation by subtracting the baseline mean and dividing by the baseline standard deviation. We consider a high z-score as indicative of the presence of emerging risks. The time-series of our aggregate risk measure is presented in Figure 1

Table II of Panel D displays summary statistics for the quarterly time-series variables We show that the average adjusted R^2 is 8.1% for the covariance model that includes

[24] We originally identified 60 themes but reduced the number to 31 after noting that a number of themes were highly correlated with other themes. This reduction resulted in variance inflation factors below 10.0 for our static covariance model, indicating that multicollinearity is unlikely to be a concern. See Appendix A for a list of the original themes. For robustness, we also consider a variation where we use the 25 LDA topic loadings instead of the 31 semantic theme loadings and obtain similar results. This indicates that the 31 semantic themes are correctly capturing the information in the LDA loadings.

only accounting variables (bank characteristics and industry). The incremental R^2 of the inclusion of the 31 semantic themes to the accounting variables-only covariance model is 1.2%. Hence, the additional textual information improves explanatory power by a material 15%. The covariance model with only accounting variables has a higher R^2 because it also includes the industry controls and bank size, which are first-order drivers of covariance.

Another observation from Panel D is that both R^2 variables have substantial variation. For example, the marginal R^2 from the inclusion of the 31 semantic themes ranges between 0.1% and 3.0%. This variation illustrates the crucial property of our emerging risk model: covariance explanatory power can deviate substantially from its long term average.

We also include summary statistics on other risk metrics. The average VIX index during our sample is 20.9, and it reaches a high of 51.7 in the 4th quarter of 2008. The mean pair covariance over the sample period is 1.062. The average cross-sectional standard deviation of monthly returns is 15.4% for all firms, and 9.1% for banks only. The lower result for banks only is because (A) firms in a specific industry have less cross-sectional variance due to the common industry component and (B) banks are highly regulated and insured. The average economic policy uncertainty from Baker, Bloom, and Davis (2016) is relatively low at 110 but has a maximum of 215 in the latter period of our sample. Using only the news component, this variable is slightly higher.

Table III displays Pearson correlation coefficients for our different time series variables that may capture the emergence of risk. The standard time series variables used in past studies (VIX, cross sectional return volatility, and average covariance) are strongly positively correlated. For example, the average pairwise covariance for our sample of banks, as well as both metrics of average cross-sectional standard deviation are more than 50% correlated with the VIX. In contrast, the two R^2 variables from our risk model have lower and often negative correlations with the other variables. Our later results will show that this is because our static risk model R^2 variables substantially lead these other measures in time series. This dramatically impacts their simultaneous correlations.

We compare our aggregate risk score to other metrics of emerging risk in Figure 4. Panel A displays the levels of four well-known emerging risk variables: the VIX, quarterly average pairwise covariance among bank-pairs and the quarterly average standard deviation of returns for all firms and financial firms. In addition, we include the EPU index from Baker, Bloom, and Davis (2016).[25] Panel B plots analogous results for variations of our covariance model: including only accounting and bank characteristics (no text), the static

[25]See replication files at http://www.policyuncertainty.com/.

model (as reproduced from Figure 1), the dynamic model and an LDA-only model.

In Panel A, the VIX, average covariance, both measures of cross sectional return volatility, and the EPU generally do not become elevated above baseline levels until 2008, close to Lehman Brothers' failure. It is noteworthy, however, that the EPU index increases from roughly 50 to roughly 100 in late 2007, indicating some increase in potentially relevant media coverage in late 2007. However, we also note that even at this time, the EPU index is not particularly alarming as it remains near its sample-wide mean of 111. Overall, we conclude that although these existing measures are relevant for a myriad of reasons, none become particularly elevated before 2008.[26]

In Panel B, the model with only accounting variables becomes significantly different from the baseline period just after the first quarter in 2007. It remains elevated from the end of the second quarter of 2007 through the first quarter of 2009. We conclude that our covariance model offers advantages relative to the aggregated measures in Panel A even before we include the text-based risk variables.

More importantly, Panel B also shows that the R^2 due to the textual semantic themes emerges significantly earlier. In particular, risk becomes elevated in the second quarter of 2005 and more so by early 2006. This is well before the crisis emerges, and before R^2 of the model with only accounting variables emerges. In each model, the aggregate risk remains elevated as the crisis materializes in 2008, and tapers off thereafter. In the last panel, using the R^2 due to LDA topics rather than SVA themes, results in a similar pattern in aggregate, but the specific LDA factors lack interpretability. Thus, for the remainder of the paper, we concentrate on SVA textual themes.

B Individual Emerging Risks

The preceding analysis provides evidence that semantic themes may provide an early warning of an increase in financial instability. A primary advantage of semantic themes, as compared to accounting or aggregate financial market variables, is the ability to interpret the specific channels that may be contributing to emerging risk build-up. For example, it is not clear what action should be taken to monitor emerging risk if volatility suddenly explains a significant amount of comovement. However, verbal information directly relating to potential risks such as real estate, commercial paper, or rating agencies can be assessed more directly. In this section, we discuss the contribution of each of the 31 relevant static

[26] We acknowledge that the recent financial crisis is only one data point, and hence it is difficult to determine the precise efficacy in predicting a financial crisis.

semantic themes in explaining the covariance of bank-pair returns. By doing so, we can identify specific channels driving emerging risks in each quarter.

As with the aggregate time series results in Figure 4, we first compute the marginal R^2 contribution of each individual semantic theme in each quarter using our model in Equation (3). This is done by computing the adjusted R^2 of the full model including bank characteristics, and then recomputing the adjusted R^2 with a single semantic variable excluded. We then take the difference between the two. This calculation is done separately for each of the 31 semantic themes, and the result is a single quarterly time series of R^2 contributions for each semantic theme.

In each quarter, we thus compute the marginal adjusted R^2 contribution from the 31 risks. As we did for the aggregate emerging risk score, we define the initial part of our sample (1998 to 2003) as a calibration period, and use this period to compute a baseline mean and standard deviation for each theme's quarterly marginal R^2. In each of the subsequent quarters from 2004 to 2015, we compute a z-score based on how many standard deviations the current value is from the baseline mean and plot the resulting quarterly z-score. An elevated z-score indicates an emerging risk factor.

Appendix B reports a fully detailed set of figures displaying the time-series of z-scores for each of our 31 text-based emerging risk factors. In Figure 5, we restrict attention to only the most prominent emerging risks in the period leading up to the 2008 financial crisis. The figure shows a large increase in the z-scores for the semantic theme "real estate" (Reinhart and Rogoff (2009)) consistent with the build-up of risk in mortgage credit preceding the crisis (Mian and Sufi (2009)). Demyanyk and Hemert (2011) suggest "that the seeds for the crisis were sown long before 2007, but detecting them was complicated by high house price appreciation between 2003 and 2005 - appreciation that masked the true riskiness of subprime mortgages." Notably, our methodology detects the emergence of these risks in 2005, well before delinquencies in the 2006 and 2007 loan vintages became apparent (in Subsection C we use an extended model to examine this theme further).

A related theme to real estate is prepayment risk, which shows very high elevation in the second quarter of 2005. This theme has component words such as "mortgage-backed" and "penalties" and likely captures the propensity of borrowers to refinance existing loans such as mortgages and credit cards. It may also reflect prepayment risk of corporate borrowers who are likely to renegotiate existing loans (Roberts and Sufi (2009)).

We also observe elevated risks for commercial paper in the first quarter of 2007, indicative

of worries by some investors regarding the quality of these securities during the crisis. This may be due to concerns about mortgage-backed securities and the liquidity of various short-term assets (Kacperczyk and Schnabl (2010), Covitz, Liang, and Suarez (2013) and Acharya, Schnabl, and Suarez (2013)). The timing of the risk elevation for this theme is in advance of the bankruptcy filing of the two Bear Stearns' hedge funds on July 31, 2007.

Risks related to credit cards become prominent in late 2006. Issuance of credit card securitizations steadily increased in 2006 and 2007, eventually reaching $94 billion in 2006.[27] Mian and Sufi (2011) find "that homeowners with high credit card utilization rates and low initial credit scores have the strongest tendency to borrow against an increase in home equity." However, they find no evidence that borrowers use the funds from refinancing to reduce credit card balances. This can increase default exposure for issuing banks.

The semantic theme related to dividends is also prominent in the pre-crisis period. Acharya, Gujral, Kulkarni, and Shin (2011) show that banks, even at the height of the financial crisis, continued to pay dividends to equity holders. The paying of dividends further depletes regulatory capital at precisely the time banks are experiencing losses.

It is well-known that credit rating agencies played a role in the crisis. For example, the literature documents problems with the rating process such as ratings shopping (Benmelech and Dlugosz (2009), Skreta and Veldkamp (2009), Bolton, Freixas, and Shapiro (2012), and Griffin and Tang (2012)), ratings catering (Griffin, Nickerson, and Tang (2013)), rating agency competition (Becker and Milbourn (2011)), and rating coarseness (Goel and Thakor (2015)). While we find an emergence of this risk before the Lehman bankruptcy (September 2008) in the first quarter of 2008, the timing suggests that foreknowledge of the importance of this risk arrives later than some other factors, closer to the onset of the financial crisis.

The operational risk theme in Figure 5 is heightened as early as 2004 and remains elevated until early 2009. This factor is less specific than others and likely captures overall concerns about banks' ability to manage increased exposure to risk, and the extent to which they have robust risk management procedures. This theme is important because the mitigation of risk is often discussed in conjunction with the disclosure of such risks, making it a prominent leading indicator. Indeed, the most prominent word in the SVA vector that is not related to the title of the theme is "manage."

Three other themes from Appendix B are also elevated but not presented in Figure 5. The first two are short-term funding themes: cash and deposits. Fahlenbrach, Prilmeier,

[27] See Report to the Congress on Risk Retention at https://www.federalreserve.gov/boarddocs/rptcongress/securitization/riskretention.pdf.

and Stulz (2012) find that banks that relied more on short-term funding, including deposits, performed poorly during the crisis. The third theme is governance. A number of papers have suggested that risk-taking is related to ownership control (Laeven and Levine (2009)) and governance structure during the financial crisis (Pena and Vahamaa (2012) and Aebia, Sabatob, and Schmid (2012)).

It is noteworthy that some risks do not emerge around the time of the 2008 crisis. In Appendix B, we do not find elevated risks related to counterparty risk, derivatives, securitization, or executive compensation even though some of these risks were identified as contributing to the crisis ex post. For example, concerns about executive compensation were raised based on the prediction that bank managers might have engaged in excessive risk taking because government guarantees can hedge downside risk. Alan Blinder "refer(s) to the perverse incentives built into the compensation plans of many financial firms, incentives that encourage excessive risk-taking with OPM – Other People's Money."[28] Our finding that executive compensation risk does not emerge is consistent with Fahlenbrach and Stulz (2011), who find no evidence that worse compensation incentives were correlated with bank performance during the crisis.

Derivative and counterparty risk (Duffie and Zhu (2011) and Acharya, Philippon, Richardson, and Roubini (2009a)) are only slightly elevated prior to the crisis despite the fact that counterparty risk associated with credit default swaps may have enabled an "unsustainable credit boom" that lead to excessive risk-taking by financial firms (Stulz (2010)).

The lack of significance for securitization as a standalone theme may be due to the inclusion of other themes based on securitized assets such credit cards, student loans, and real estate. One interpretation is that the act of securitization, by itself, is not necessarily a risk factor, as one needs to know the asset class involved in order to determine whether banks have heightened risk exposure.

Although we use the financial crisis as an experiment, the ultimate viability of the approach depends on being able to identify future emerging risks. In this spirit, we note that Figure 1 shows that there is a major decline in early 2009 in the contribution of the static semantic themes, consistent with the ultimate recovery that was observed, and with government interventions to reduce systemic risk. However, as can be seen in both Figure 1 and Figure 6, a substantial number of new risks appear to be emerging from 2011 to 2015. Predicting future events in real-time is a rare achievement for academic research, but the

[28] Crazy Compensation and the Crisis, *Wall Street Journal*, May 28, 2009 http://www.wsj.com/articles/SB124346974150760597.

possibility of building models with this potential is a key motivation for this research.

In Figure 6, we see a heightened risk associated with mergers and acquisitions in more recent years. According to the St. Louis Fed in June of 2014, "bankers have expressed renewed interest in strategic partnerships after weathering the financial crisis."[29] This continues into 2016 as analysts expect it to be a busy year for bank mergers as "small and medium banks look for ways grow, cuts costs, and survive in an industry in which bigger increasingly seems better."[30]

We also find that themes related to funding sources, such as cash, emerge strongly in this period. Although not shown, we find a similar emergence for other short-term assets such as certificates of deposit, credit cards, and deposits (see Appendix B). This might indicate that conditions such as very low interest rates posed challenges for the traditional funding sources of banks. For example, deposits typically earn rates of interest well-below those of short-term treasuries, making it harder for banks to attract customers. Maintaining this profitable deposit-to-bond differential is difficult when bond yields are effectively zero as they would require negative deposit rates. Related to this issue, the *Wall Street Journal* notes that earnings for banks in the first quarter of 2016 were expected to decline 8.5% from the same period last year.[31]

Real estate risk declines after the financial crisis but re-emerges in late 2013 as the housing market begins to rally once again. Backlogs of foreclosures continued to rise during this time, creating uncertainty in the balance sheets of financial institutions.[32]

A plethora of lawsuits aimed at banks were announced in late 2015 and 2016. For example, seven large banks settled a private lawsuit accusing them of rigging the ISDA benchmark in fixed income.[33] Thus, the increase in this risk exposure in 2016 may presage additional lawsuits in the financial sector.

The tax semantic theme is intermittently heightened in the latter part of the sample.

[29]Gary Coner, Is High Tide Approaching for Bank Mergers and Acquisitions? *Banking Insights* (https://www.stlouisfed.org/bank-supervision/supervision-and-regulation/banking-insights/is-high-tide-approaching-for-bank-mergers-and-acquisitions)

[30]Deirdre Fernandes, Bank Mergers Increase as Profits are Squeezed, *Boston Globe* (https://www.bostonglobe.com/business/2016/02/10/bank-mergers-increases-profits-are-squeezed/SWpRTme0Tml1HK4ucmkVDK/story.html).

[31]Kuriloff, Aaron, Miserable Year for Banks: Stocks Suffer as Rates Stay Low, *Wall Street Journal* April 10, 2016.

[32]See http://www.nytimes.com/2013/10/10/us/real-estate-boom-in-phoenix-brings-its-own-problems.html?_r=0 and http://www.forbes.com/sites/morganbrennan/2013/01/17/worst-of-foreclosure-crisis-is-over-but-problems-remain/#13bac1435748.

[33]Jonathan Stempel, Seven Big Banks Settle U.S. Rate-Rigging Lawsuit for $324 Million, *Reuters*(https://www.reuters.com/article/us-banks-rigging-settlement/seven-big-banks-settle-u-s-rate-rigging-lawsuit-for-324-million-idUSKCN0XU2B5).

The increase in the importance of this theme could be due to the fact that, in 2014, the legislators put forth a proposition to increase taxes on bank liabilities rather than profits to reduce bank reliance on debt.[34] Tax reform was also a central topic in the period leading up to the presidential election.

Counterparty risk has been a focus for financial regulators recently. Federal Reserve chair, Janet Yellen notes "in the 21st century, a run on a failing banking organization may begin with the mass cancellation of the derivatives and repo contracts that govern the everyday course of financial transactions."[35] The increase in the importance of this theme in late 2014 is consistent with concerns over the importance of this risk.

The semantic theme related to operational risk is also highly elevated in this period, suggesting that risk management may be at the forefront of investors' concerns.

We conclude this section by noting that some of these recent risks may not portend financial instability in the future either because regulators intervene before the risk becomes apparent or because concerns surrounding these risks are alleviated by economic conditions. The findings of this section, however, show how the model can help regulators assess the build-up of particular risks and to then strategically use data from supervisory activities to monitor the banking sector.

C User-Defined Risks

The static model has the flexibility to easily query and extend the SVA model using additional key phrases of interest. These themes may used in hypothesis testing by researchers and/or regulators to explore the emergence of granular risks through time. This is particularly useful because the examination of the associated sub-themes of a broad static theme (see Table I) allows one to understand the specific manifestations of risk over time. In addition, the model can be similarly extended to include cases where a user is interested in the time-series importance of a specific risk identified by academic theory, or in the case of a regulator, through prudential supervision. The benefit of the user-defined risks approach is that it permits flexibility and can incorporate the expertise of the user.

The user-defined risk methodology simply adds a set of user-defined semantic themes to those in the static model. The importance of the user-defined themes is then determined using the same covariance model as was used by the static model. In particular, we compute marginal adjusted R^2 by comparing the explanatory power of the full model to that with a

[34] Mark Roe and Michael Troge, How to Use a Bank Tax to make the Financial System Safer, *Financial Times*(https://www.ft.com/content/468a9fe2-b2ce-11e3-8038-00144feabdc0).

[35] See http://www.federalreserve.gov/newsevents/press/bcreg/yellen-opening-statement-20160503.htm

given user-defined theme excluded. In this section, we present two applications of the user-identified risk model. The first explores sub-themes contained in the static theme "real estate" and the second explores the sovereign debt crisis in 2010 to 2011.

Panel A of Figure 7 presents z-score results for semantic sub-themes related to real estate. We chose sub-themes that were noted as important during the financial crisis to examine the timing and magnitude of their emergence. In the first graph, we see elevation of sub-themes including "subprime," "mortgage-backed," and "foreclosed" prior to the onset of the crisis (Longstaff (2010), Mian and Sufi (2011), and Adelino, Schoar, and Severino (2016)), likely when homeowners began to default on these loans.

Interestingly, both mortgage-backed and HELOC risks are elevated in recent periods. An August 2016 Wall Street Journal article notes that the interest rate on many HELOCs taken out before the financial crisis are being reset, increasing default probability.[36]

In addition to delving more deeply into the themes identified by the model, the methodology can also examine emerging risks obtained from sources outside the model. For example, we consider risks related to the sovereign debt crisis that began in earnest in 2010. Panel B of Figure 7 presents the results. The figure shows that many proposed sub-themes are indeed elevated prior to 2010. Sub-themes related to Eurozone, the IMF, and (to a lesser extent) affected countries such as Greece and Brazil all have heightened z-scores. Interestingly, we also find that risk related to Brazil spiked to a z-score just below 80 in late 2016, right at the time Brazil was facing a serious recession relating to a political crisis.

In summary, we find that many risks are visible in the trading patterns of investors and their quantitative link to risks affecting the financial sector as a whole. Our method can detect emerging risks that are known to be related to the recent financial crisis as well as potential new risks. The ability to identify specific emerging risks in real-time can alert regulators to potential issues early, and inform researchers regarding specific channels that might merit further examination.

VI Dynamic Risks Methodology

The second application of the methodology is a dynamic model that allows risks to emerge without any prior examination of topics on the part of the researcher. To determine which themes to include using automation, we start with the annual LDA model that identifies 25 topics in each year. The LDA output for each year includes 25 vocabulary lists, with each

[36] Anna Maria Andriotis, Home Equity Loans Come Back to Haunt Borrowers, Banks (https://www.wsj.com/articles/home-equity-loans-come-back-to-haunt-borrowers-banks-1470933020)

word or commongram in each list having weights indicating the importance to the given topic. We use these 25 vocabulary lists to extract the 25 most probable commongrams for each topic. This process results in ($25x25 = 625$) commongrams in each year, which we then deem to be candidate risk factors. We next reduce this list by removing duplicates (some commongrams appear in more than one LDA topic) and also by limiting our analysis only to bigrams, as bigrams are more interpretable than single words. These initial steps reduce the list of 625 candidate risks to roughly 350-400 in each year.

We then review these candidate risk factors in each year, and delete any risk factor bigrams that are boilerplate or that lack a clear economic interpretation. Unlike the static model, where the researcher chooses topics to *include*, here the researcher merely *excludes* candidate bigrams that are obviously non-economic in nature. This step reduces the number of candidate bigrams by roughly 60%, with the remaining bigrams being both interpretable and having the potential to convey economic meaning. For each of the 150 or so bigrams, we generate a vector of companion words using SVA (similar to the process used for the 31 static themes). Because this approach has the dynamic model use a rather extensive collection of bigrams, the set of emerging risks is larger and more detailed than that of the static model.

In order to determine which of these risks are emerging, we use a forward stepwise regression to maximize the R^2 of the covariance regression in Equation (3) that includes the theme with the most explanatory power first, and then holding this theme constant, includes the next most powerful theme, and so on. We calculate a z-score for each variable by computing the mean and standard deviation of the given variable's time series marginal contribution to the R^2 using the past 5-year rolling window. We then report risk factors on in Table V if it satisfies either of the following conditions: (A) its z-score exceeds ten, (B) its z-score exceeds 7.0 and it is one of the top three emerging risks for the given quarter.

As we saw in the dyanmic model, real estate, rate swap, and underwriting standards are prominent prior to the financial crisis. As the crisis unfolds, risks related to reputation, management policies, legal proceedings and regulatory approval become manifest. In the post-crisis period, the model produces risks due to mortgages, foreclosure process, commercial real estate, and deposit insurance. In the more recent period, Basel III is mentioned as well as some risks related to weather events and institution failures.

As with the static model, a user interested in a deeper understanding of a given risk can look at the given risk's SVA word vector for additional context. If for example, one

is interested in the "weather events" risk, the underlying vector contains terms such as extreme weather, hurricanes, weather patterns, ice storms, floods, storms, and droughts. Similarly, the bigram "rate swap" has a SVA vector that includes words such as notional, LIBOR, swaptions, collar, and floors.

The static and dynamic models can be used in tandem to better understand emerging risks. A casual observer who compares these models will note that although many of the risks are new, some of the dynamic themes overlap with the static model's themes. For example, the static theme "real estate" maps to terms found in the dynamic model such as "foreclosure", "property", and "mortgage". This allows the researcher to better understand specific risks that may be driving the broad static themes. In addition, the flexibility of our approach allows new risks of particular interest identified by the dynamic model to be monitored through time by adding it to the static model. In conjunction, the static and dynamic methods of identifying emerging risks can aid researchers and regulators alike in better understanding the sources of emerging risk in the financial sector.

VII Individual Bank Outcomes

While the preceding time-series analysis is important from an early warning and macroeconomic financial stability perspective, it is even more useful if overall exposure to emerging risks can further predict which banks may have adverse financial outcomes.[37] Such cross sectional tools can also be used to identify which banks might benefit most from additional stress tests or examinations given their high predicted exposures to risk, in general. In this section, we explore whether a bank's ex ante aggregate exposure to emerging risk can be used to predict bank-specific negative outcomes.

We construct bank-specific exposures to emerging risks in each period using a decomposition of the predicted values from the covariance model based on our 31 static risk factors. The result is a direct measure of which banks are most vulnerable to financial instability given the specific risks that are emerging. In these analyses, we label this measure *Emerging Risk Exposure*. This variable is computed as the average predicted covariance bank i has with all other banks j using the main covariance model in Equation (3). This is computed separately in each quarter and for each bank using the following two step procedure. First, for each bank-pair in a given quarter, we take the product of the fitted coefficients for each SVA theme (β_1 to β_{31}) from the baseline covariance model using semantic themes only,

[37] It is important to note that because each bank has a loading on a specific theme, our methodology also allows one to isolate which banks may have high exposure to certain individual emerging risks.

and multiply it by the given bank-pair's SVA theme loading ($S_{i,j,t,1}$ to $S_{i,j,t,31}$). We then winsorize the products at the 5/95% level to reduce the impact of outliers and sum the resulting 31 products for each bank-pair to get the total predicted covariance of bank i with each bank j. Finally, we average the predicted covariances over banks j to get the total *Emerging Risk Exposure* for bank i in quarter t.

Using three cross sectional tests, we test whether individual banks with greater *Emerging Risk Exposure* will be more likely to experience subsequent ex post negative outcomes. First, we examine whether ex ante exposed banks experience more negative stock returns during the later periods of the financial crisis or during a more recent period of economic uncertainty. Second, we use the FDIC's Failures and Assistance Transactions List to analyze whether more risk-exposed banks are more likely to fail ex post.[38] Our third test examines whether deeply lagged risk exposures can predict higher bank-specific volatility.

A Predicting Crisis and Current Period Stock Returns

In Table VI, we examine whether ex ante *Emerging Risk Exposure* can predict ex-post stock returns from September 2008 until December 2012 (left hand side of the table) and separately from December 2015 to February 2016 (right-hand side of the table). In Panel A, we regress raw stock returns during these periods on *Emerging Risk Exposure* measured in the specific quarter indicated in the column titled "Quarter". For example, row (9) examines whether banks exposed to emerging risks in the first quarter of 2006 experienced more negative stock returns during the crisis. In Panel B, we examine if the model is more adept at predicting larger negative components of returns (left tail events). The dependent variable in this panel is the negative portion of the raw return (computed as min[0,raw return$_t$-sample mean return$_t$]). In all models, we include (but do not display to conserve space) controls for momentum (month t-12 to t-2), log book-to-market ratio, log market capitalization and a dummy variable for negative book-to-market ratio.

Panel A indicates that an individual bank's exposure to emerging risk factors as early as the third quarter of 2007 is negatively related to its stock returns during the aftermath of the financial crisis. The greater the bank's risk exposure prior to the crisis, the more negative is its return. Focusing only on negative returns, Panel B shows that emerging risks predict outcomes during the crisis as early as the first quarter of 2007. In both panels, as we draw closer to the third quarter of 2008 and Lehman's bankruptcy, the *Emerging Risk Exposure* coefficient becomes increasingly negative and more significant.

[38] https://www5.fdic.gov/hsob/SelectRpt.asp?EntryTyp=30&Header=1.

In Figure 6, we noted a number of new emerging risk factors since 2013. Examining these more recent period returns in Table VI indicates that the seeds of the recent economic uncertainty were evident as early as 2010. This period was characterized by a market trough after Lehman's bankruptcy, the passage of the Dodd-Frank Act, and concerns regarding the European debt crisis, eventually leading up to negotiations over the U.S. government's debt ceiling. In late 2012 and early 2013, the table shows that the ex ante ability of quarterly exposures to emerging risks to predict future bank returns diminishes for a period, and then re-emerges strongly for both raw returns and negative returns in the second quarter of 2014.

Our results are consistent with Fahlenbrach, Prilmeier, and Stulz (2012) who argue that a bank's risk culture, which is long-lasting and pervasive, can contribute to its performance in future financial crises. However, unlike their methodology, which relies on prior crisis stock returns to measure risk, we are able to pinpoint the specific sources of risk that might contribute to the bank's culture and subsequent underperformance.

B Predicting Bank Failures

We next examine whether emerging risk exposures can predict which banks fail following Lehman's bankruptcy using data from the FDIC website. The first bank failure following the Lehman bankruptcy in September 2008 occurs in November of 2008. The last occurs in June of 2012. There are 41 such failures, with 2, 12, 19, 6 and 2 occurring in the years 2008, 2009, 2010, 2011, 2012, respectively. We note that results are unchanged if we limit the sample of banks to those that failed in the narrower window between 2008 and 2010. However, we believe that even later failures during this longer interval are likely related to emerging risks associated with the financial crisis and its aftermath.

We define the dependent variable as a dummy variable, *Failure*, equal to one if the given bank was assisted or failed, and zero otherwise. We then regress this variable on the ex ante *Emerging Risk Exposure* from the period specified in the first column.[39] We include controls for bank characteristics as in prior tables and include industry fixed effects based on four-digit SIC codes.

We find in Table VII that the lagged *Emerging Risk Exposure* of an individual bank strongly predicts which banks will fail. This relationship is intermittently significant as early as 2005 in predicting ex post bank failures and then becomes more reliable starting in the fourth quarter of 2006. These results are consistent with Table VI, which shows

[39] Although we present results using a linear probability model (OLS-based) due to the presence of industry fixed effects, we note that these results are robust to using a logistic model instead.

analogous results for negative stock returns during this period.

We also find results supporting the existing literature (see Sarkar and Sriram (2001)) regarding other bank characteristics. For example, banks are more likely to fail if they have more loans and non-performing assets, but are less likely to fail if they have greater capital (Berger and Bouwman (2013)) and higher liquidity (Berger and Bouwman (2009)).[40]

C Predicting Monthly Volatility

Finally, we examine whether *Emerging Risk Exposure* can predict a bank's ex post monthly volatility in unconditional tests.[41] In Table VIII, we consider monthly Fama and MacBeth (1973) regressions where the dependent variable is the ex post monthly stock return volatility computed using daily stock returns. We include, but do not display to conserve space, controls for bank characteristics, momentum (month t-12 to t-2), log book-to- market ratio, the log market capitalization, and a dummy variable for negative book-to-market ratio.

Our baseline regression in the first row lags the independent variables by just one month allowing us to examine the link between emerging risk exposures and one-month-ahead volatility. We then apply deeper lags up to 36 months. We find that even deeply lagged risk exposures, up to 36 months, can predict subsequent return volatility. Columns three and four illustrate that using longer ex ante measurement periods does not improve predictability.

We conclude that a financial institution's exposure to emerging risks predicts future volatility even in this unconditional setting. This indicates that ongoing information production and trading by investors contributes to elevated volatility levels. These results are broadly consistent with Bekaert and Hoerova (2014) who state that stock market volatility "predicts financial instability more strongly than does the variance premium."

VIII Investor versus Bank Information Production

Our methodology relies on the assumption that both banks and investors simultaneously produce information. Public banks are required by the SEC to disclose risk factors, thus confirming our assumption on information production by banks. Investor information production is not observable, but can be inferred using big data methods to examine changes in the covariance matrix. A potentially confounding issue we acknowledge is that changes

[40] In addition to our controls, the literature has documented other determinants of bank failure including exposure to commercial real estate (Cole and White (2011)) and non-traditional banking activities such as investment banking and asset securitization (DeYoung and Torna (2013)).

[41] The same quarter's exposures are used to predict ex-post months t=1 to t=3.

in covariance could be driven by changes in the relationship between bank disclosure and trading patterns (for example due to regulatory changes), and not necessarily by new information produced by investors. We note two findings that suggest that this not the case.

First, the timing of our measurement of information production differs between investors and banks. We use annual 10-Ks to construct emerging risk themes from banks for our covariance model in Equation (3). Hence the independent variables (risk exposures) are updated only annually. For any quarter ending in calendar year t, we thus use the same 10-K for each bank from its fiscal year that ended in calendar year $t-1$. The covariance, on the other hand, is measured at a higher quarterly frequency.

Because our covariance regressions are run at the end of each quarter, it follows that all independent variables are lagged at least one quarter, and up to four quarters. In the first quarter of each year, newly filed 10-Ks for the prior fiscal year are used to generate a new set of semantic theme loadings for the model. However, in the second, third, and fourth quarter of each year, the semantic theme loadings and bank characteristics on the right-hand side do not change.

As a result of this timing, if our results are driven by updated bank disclosures alone (such as any changes instigated by disclosure rule changes), we would expect a step-function pattern in Figure 1 where the statistical significance of the R^2 jumps every first quarter and remains relatively flat for the following three quarters. If instead, investors also produce and trade on information related to changes in perceptions of the importance of disclosed risks, we would expect the R^2 to change materially even in quarters where the 10-Ks are not being updated. Examination of the graph strongly rejects the step-function prediction and shows a high degree of quarterly change even when the 10-Ks are not updated.

Second, our findings are qualitatively similar if we lag all semantic theme loadings by one additional year (two fiscal years before the covariance is measured) when implementing Equation (3). This implies that the relationship between semantic themes and covariance must have a large contribution from information production by investors. We thus conclude that both bank disclosures and investor trading patterns are important.[42]

Third, we examine whether it is possible to identify dynamic emerging risks by considering only the changes in risk disclosure by banks (without incorporating information about

[42]We note that regulation requiring that risk factors be discussed specifically in Item 1A of the 10-K became effective on December 1, 2005. This means that 10-Ks filed in the first quarter of 2006 would be subject to this rule. As can be seen from Figure 1, we find that the significance of semantic themes strongly increases in the second quarter of 2005, well before the new rules took effect. Hence the rule itself cannot explain our findings (a conclusion that is further reinforced by a robustness test where we find similar results if we lag disclosures by one additional year).

investor trading and the covariance matrix). For example, Cohen, Malloy, and Nguyen (2018) find that changes in the risk disclosure of firms can predict future returns.

We thus examine which semantic themes experienced statistically significant shifts in their disclosure intensity in a given year, and Table IX reports the results. The methodology for determining changes in disclosure has the same starting point as the dynamic SVA model: with bigrams extracted from the LDA model. In particular, for all bigrams that are used in the SVA model (between 100 to 200 in each year), we use SVA vectors for each topic and score each bank based on how much of each bigram-topic the bank discloses. We then normalize all exposures for each bank such that they sum to one. Hence, we can explore relative changes in disclosure rather than nominal changes.

In each year, we then average the exposures to each topic across all banks, obtaining a single time-series vector for each of the topics. A given topic is deemed to be emerging in a given year if the average exposure to the topic in the given year is significantly higher than the exposure in the past 5 years. To make the list fit on one page, we sort z-scores across all years from high to low and take the 40 highest. This cutoff requires that a given topic must have a z-score roughly exceeding 8.0. Note that, using this methodology, a topic can emerge in more than one year, although this is infrequent. We exclude the year 2005 from this test due to the change in disclosure rules associated with risk factors in that year.

As can be seen in the table, some of the bigrams echo themes in the dynamic and static models. For example, banks increase their disclosure of "real estate" in 2006. At first glance, using only changes in disclosure might appear a viable alternative. However, we note several reasons why using bank-only disclosure may not result in sufficient power to detect risks. First, risk factor disclosures often change little over time (Cohen, Malloy, and Nguyen (2018), Figure 9A) even though their importance can change dramatically.

Second, SEC rules may change or they may place more emphasis on certain risk disclosures, leading to an increase in disclosure but not necessarily to more risk. As an example, in October of 2011, the SEC issued guidance on cyber security risks stating that "although no existing disclosure requirement explicitly refers to cyber security risks and cyber incidents, companies nonetheless may be obligated to disclose such risks and incidents." This could be the reason for the increase in the terms "cyber attacks" and "data processing" in 2012, but it is difficult to ascertain whether this is an emerging risk or a collective decision by banks to include this risk given the SEC's new guidance. In our model, however, we do not find a corresponding elevation of these or related themes in the dynamic model in 2012.

We conclude that examining the change in disclosure alone, without considering trading patterns, can produce false positives about emerging risks. Third, using the change in the covariance alone cannot discern the economic channels of a risk build-up as it is not text-based. Furthermore, Figure 4 illustrates that aggregate covariance only becomes elevated in 2008, when the crisis is already in full swing. A related measure, the standard deviation of financial firms, also does not become elevated until it is too late.

A final concern is that banks may change their disclosure behavior if they believe that regulators use this information to more closely supervise certain banks. We view this as unlikely for a number of reasons. First, banking regulators have access to more informative proprietary information (Peek, Rosengren, and Tootell (2003)) that could also be used in a model like ours to identify risks and ultimately the banks most in need of supervision. Second, our method crowdsources information to identify only those risks that broadly affect the banking sector. Therefore, a large fraction of banks would have to reduce their disclosure of these risks in order to lower detection as emerging risks. Third, there are a number of safeguards in place that lower the incentive (and increase the penalty) for non-disclosure of relevant information. For example, increased monitoring by shareholders and regulators can offset or even reduce the probability that public firms conceal information.

IX Conclusion

We propose an empirical model that crowdsources information from investor trading patterns and banks' 10-K disclosure using computational linguistics. The model is designed to identify emerging risks that might threaten financial stability. The use of text-based crowdsourcing on the banking side allows us to identify the likely channels driving financial instability many months (or even years) in advance. The ability to detect mechanisms is a core contribution of the paper.

Our methodology extracts themes from financial firm 10-K risk factor disclosures using Latent Dirichlet Allocation (LDA) and Semantic Vector Analysis (SVA) in tandem. The combination provides a framework that is dynamic, flexible, and allows each emerging risk factor to be interpretable. We make several contributions to the literature. First, the model produces an aggregate signal regarding financial instability through crowdsourcing information from both investors and banks. The model finds elevated risks as early as the second quarter of 2005.

Second, the model can be run either as a static model using researcher-selected factors,

or as a dynamic model identifying the sources of risk automatically, even if the researcher is ex ante unaware of the relevant channels. We find that the two models in tandem detect emerging risks that foreshadow the financial crisis of 2008 well before other potential indicators become elevated. Emerging risk themes that become prominent as early as 2005 include risks associated with real estate, commercial paper, and credit cards. In addition, the model reveals which factors were most relevant not only prior to the crisis, but also in more recent years. We find evidence that emerging risks since 2013 related to sources of funding, mergers and acquisitions, taxes, real estate, and lawsuits emerge during this period. Thus, our model uniquely provides detailed and interpretable information regarding the likely economic channels driving increases in risk in real-time.

Finally, our model analyzes the consequences of individual bank exposures to emerging risks. We find that banks with greater ex ante exposure to emerging risks experience significantly lower stock returns during the financial crisis. Furthermore, these same banks are subsequently more likely to fail. In unconditional tests using the entire sample from 1998 to 2015, we find that deeply lagged emerging risk exposures generally predict subsequent elevated stock return volatility for up to 36 months.

Overall, our methodology offers insights on emerging risk exposures at both the aggregate and individual bank level that can be used by regulators and researchers alike to monitor financial stability.

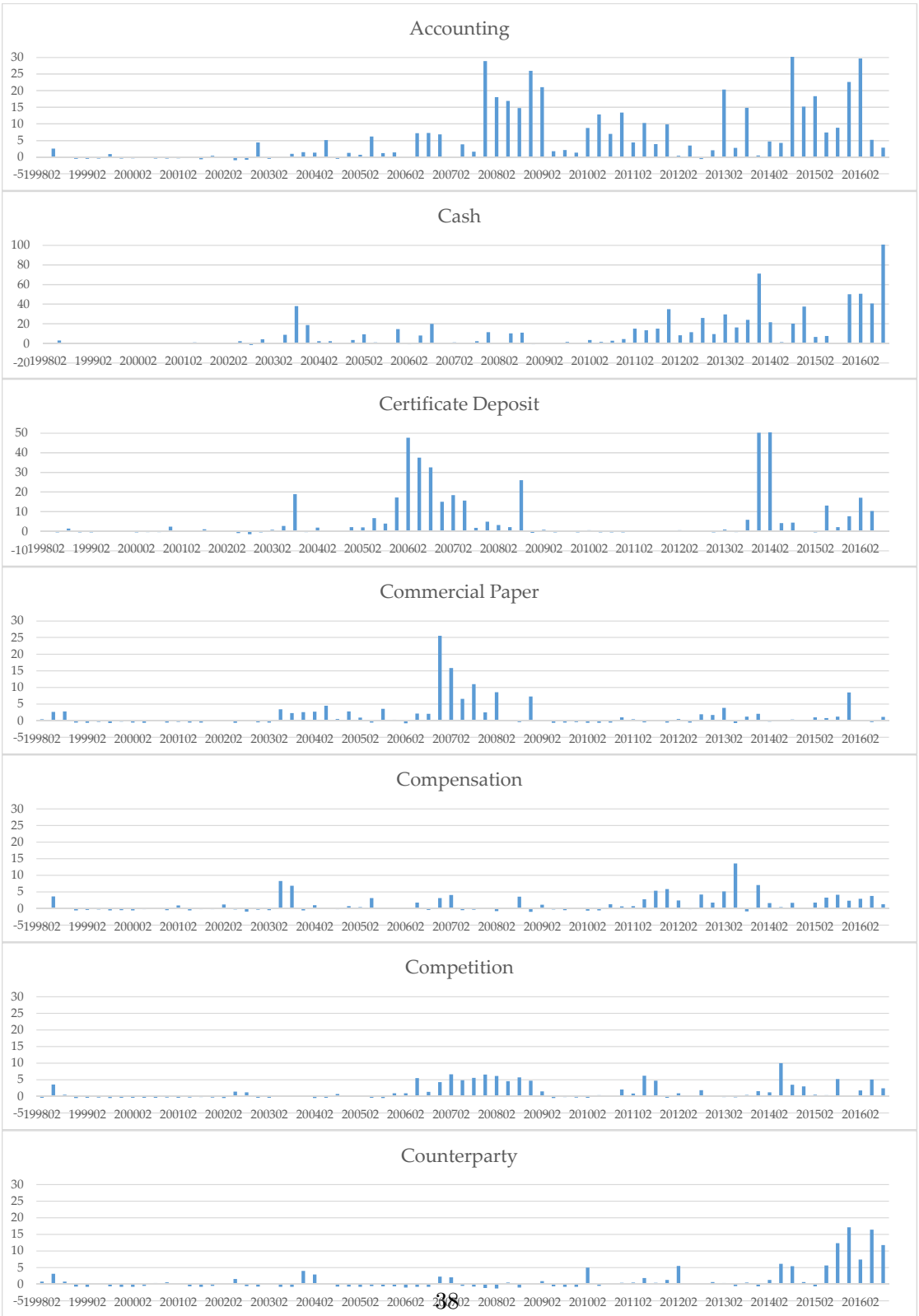
Appendix A: Candidate Static Model Themes

Candidate static themes identified from inspection of frequent topics in LDA and the academic literature on financial crises. Themes are removed because of multicollinearity.

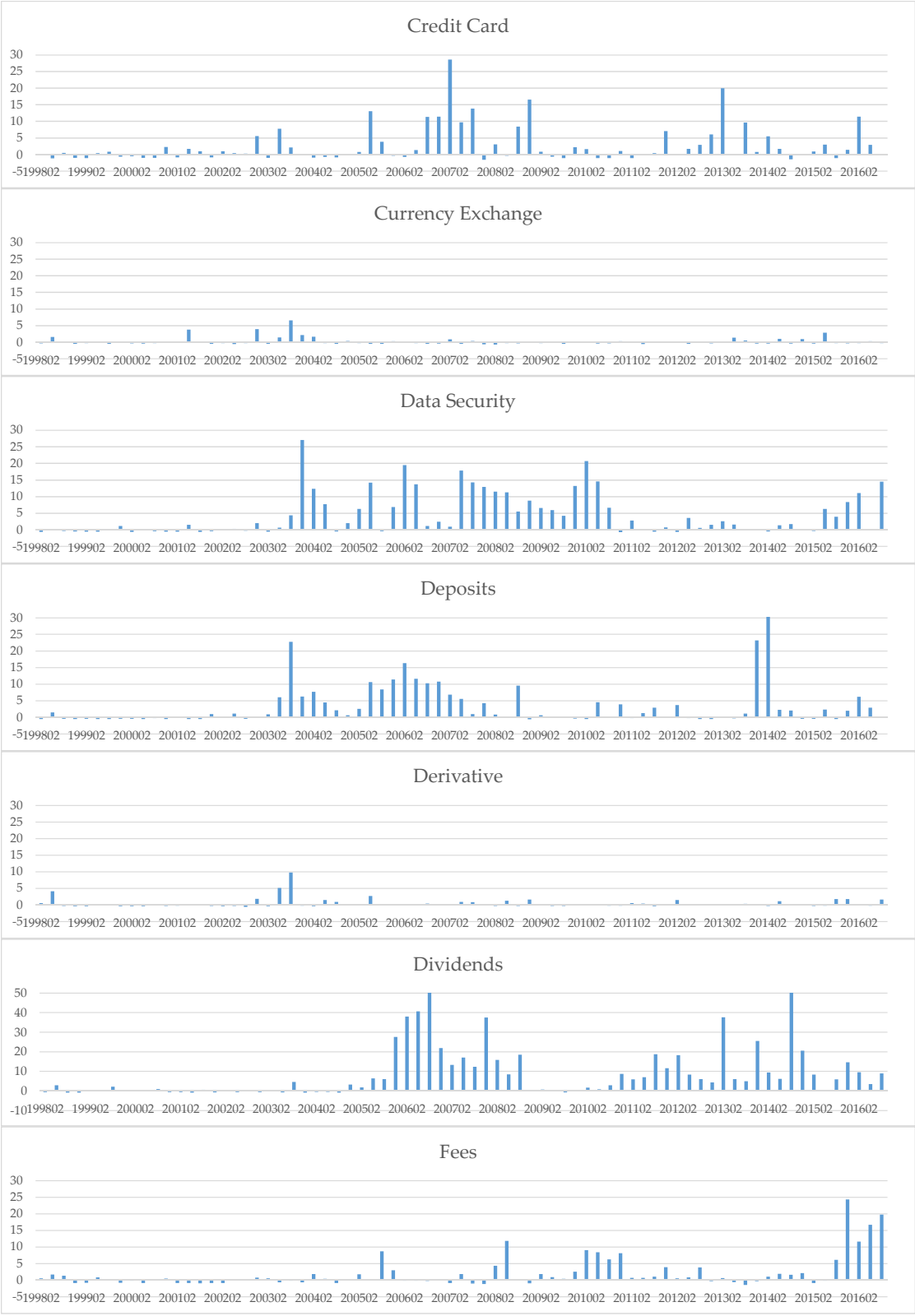
Included	Removed
Accounting	Broker Dealer
Cash	Capital Markets
Certificate Deposit	Commitments
Commercial Paper	Common Stock
Compensation	Consumer Loans
Competition	Credit Lines
Counterparty	Credit Risk
Credit Card	Economic Conditions
Currency Exchange	Federal Agency
Data Security	Federal Funds
Deposits	Foreclosure
Derivative	Gap
Dividends	Information Technology
Fees	Intangible Assets
Funding Sources	Interest Rate
Governance	Investment Securities
Growth Strategy	Letter Credit
Insurance	Liquidity
Internal Controls	Loan Losses
Lawsuit	Loans Originated
Mergers Acquisitions	Market Risk
Off Balance Sheet	Mortgage
Operational Risk	Operating Costs
Prepayment	Preferred Stock
Rating Agency	Regulation
Real Estate	Representations
Regulatory Capital	Repurchase Agreement
Reputation	Risk Management
Securitization	Servicing
Student Loans	Stress Test
Taxes	

Appendix B: Time Series of Emerging Risks

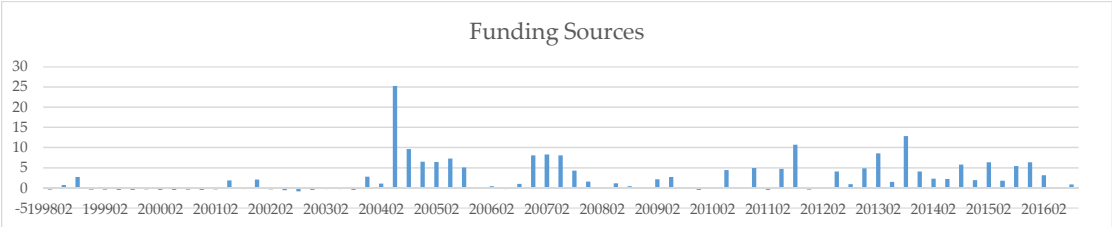
The figures report the time series of t-statistics of the R^2 from the model in Equation (3) for all 31 semantic theme emerging risks. The results are based on the time series of the contribution of individual semantic themes in explaining pairwise covariance of banks. We define the initial part of our sample (1998 to 2002 inclusive) as a calibration period, and use this period to compute each semantic themes' R^2 baseline quarterly mean and standard deviation. In each quarter, we then compute a z-score based on how many standard deviations the current value is from the baseline mean. The figure is a plot of the quarterly z-scores for each semantic theme.

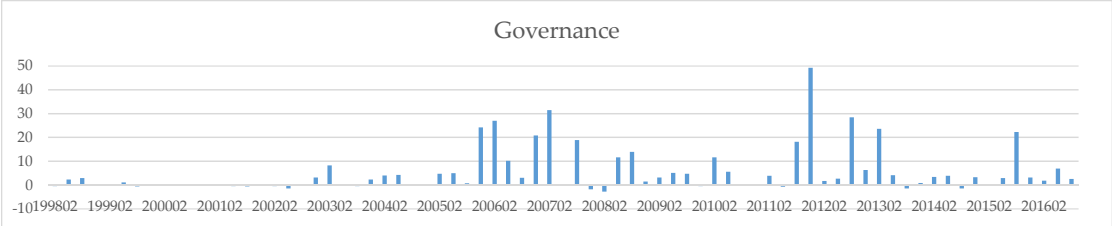


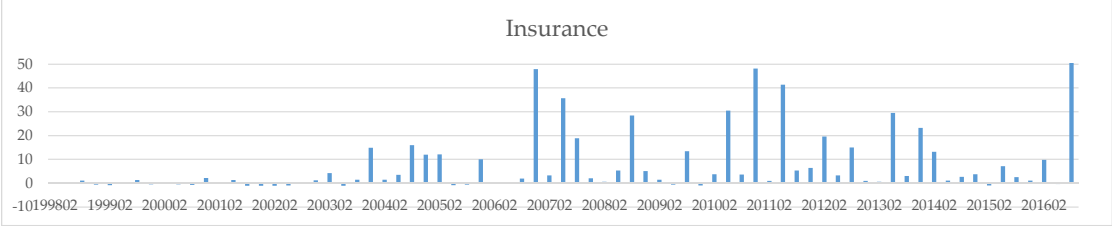
Appendix B: Time Series of Emerging Risks (continued)

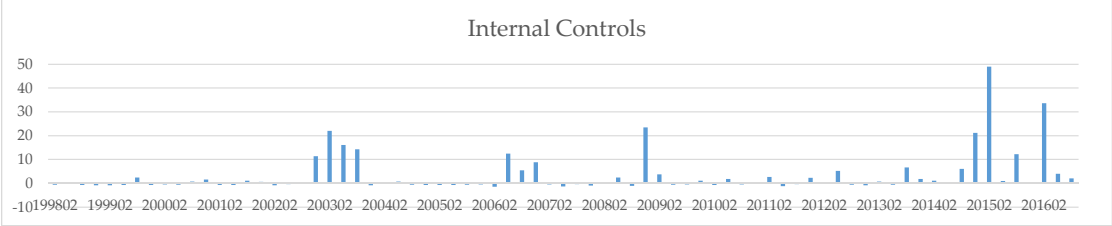


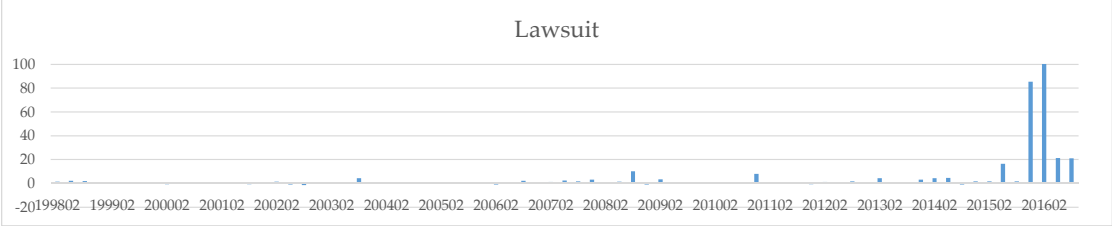
Appendix B: Time Series of Emerging Risks (continued)

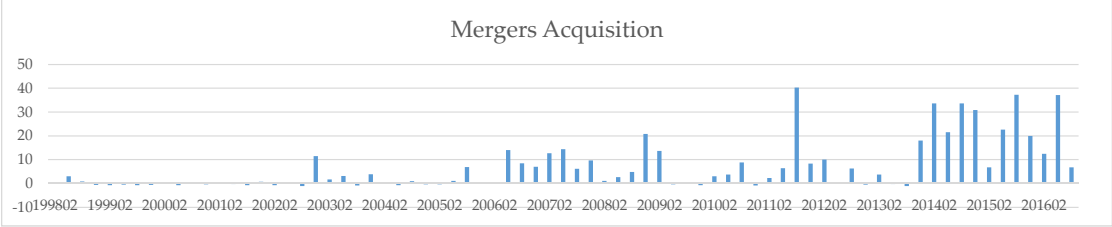




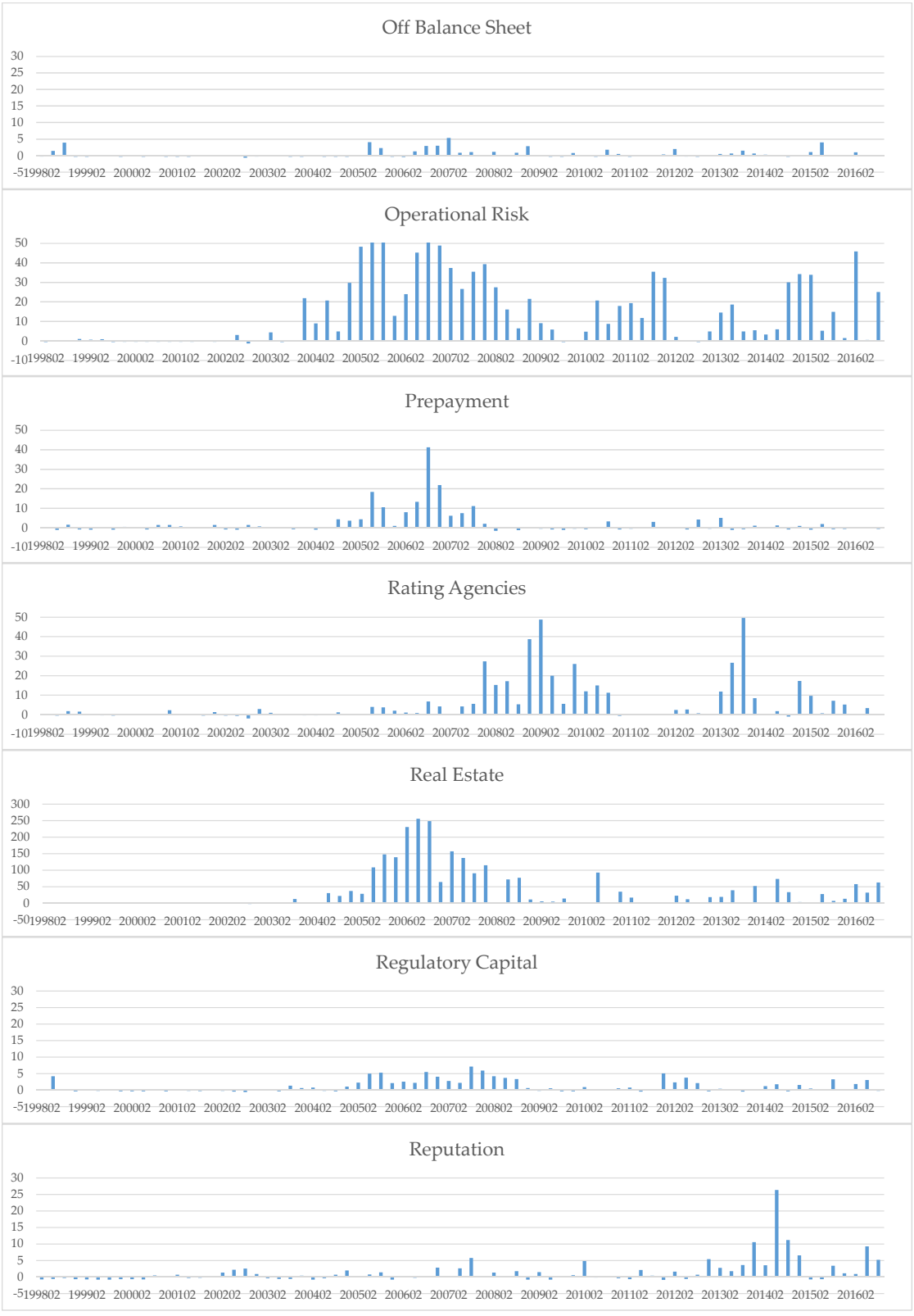




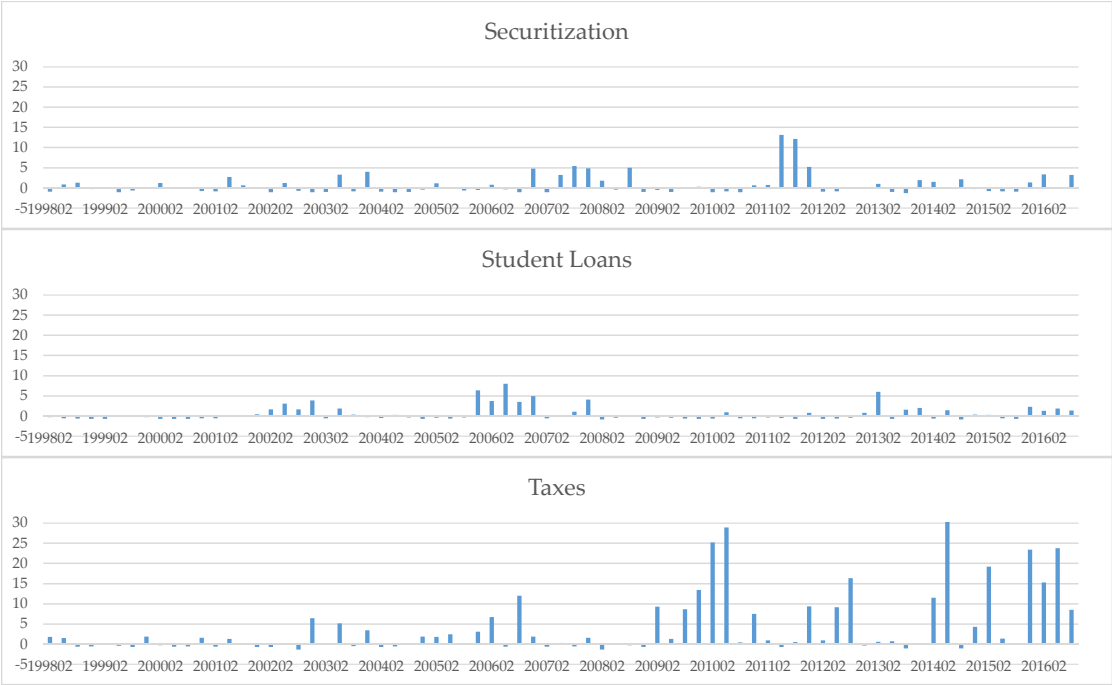




Appendix B: Time Series of Emerging Risks (continued)



Appendix B: Time Series of Emerging Risks (continued)



Appendix C: Literature Review for Candidate Risks

This table highlights candidate risk factors identified from a review of the academic literature on financial crises. The citations are meant to be representative and are thus, not inclusive.

Candidate Risk	Citation
Accounting/Fair Value	Laux and Leuz (2009), Barth and Landsman (2010), Huizinga and Laeven (2012)
Commercial Paper	Covitz, Liang, and Suarez (2013), Acharya, Schnabl, and Suarez (2013)
Compensation	Fahlenbrach and Stulz (2011)
Counterparty	Duffie (2010), Arora, Gandhi, and Longstaff (2012)
Credit Lines	Ivashina and Scharfstein (2016)
Credit Ratings	Ferri, Liu, and Stiglitz (1999), Becker and Milbourn (2011), Bolton, Freixas, and Shapiro (2012), Griffin and Tang (2012), Griffin, Nickerson, and Tang (2013), Goel and Thakor (2015)
Currency/Sovereign Debt	Kaminsky and Reinhart (1998), Goldstein (2005), Reinhart and Rogoff (2009), Reinhart and Rogoff (2011)
Deposits	Ivashina and Scharfstein (2016), Cornett, McNutt, Strahan, and Tehranian (2011)
Derivatives	Edward (1999), Acharya, Brenner, Engle, Lynch, and Richardson (2009)
Governance	Laeven and Levine (2009), Beltratti and Stulz (2012), Pena and Vahamaa (2012)
Liquidity	Chen, Goldstein, and Jiang (2010), Cornett, McNutt, Strahan, and Tehranian (2011), Bouvard, Chaigneau, and Motta (2016), Castiglionesi, Feriozzi, and Lorenzoni (2017)
Interest Rates	Hellwig, Mukherji, and Tsyvinski (2006), Reinhart and Rogoff (2011)
Loan Losses	Barth and Landsman (2010), Beatty and Liao (2011)
Real Estate	Herring and Wachter (1999), Peek and Rosengren (2000), Calomiris and Mason (2003), Reinhart and Rogoff (2009), Cole and White (2011)
Regulation (Capital)	Acharya and Richardson (2009), Fahlenbrach, Prilmeier, and Stulz (2012), Berger and Bouwman (2013), Demirguc-Kunt, Detragiache, and Merrouche (2013)
Regulation (Policy)	Acharya, Philippon, Richardson, and Roubini (2009b), Baker, Bloom, and Davis (2016)
Securitization	Benmelech and Dlugosz (2009), Longstaff (2010), DeYoung and Torna (2013)
Short-Term Funding	Diamond and Rajan (2001), Allen, Babus, and Carletti (2012), Fahlenbrach, Prilmeier, and Stulz (2012)

References

Acharya, Viral, Lasse Heje Pedersen, Thomas Philippon, and Matthew Richardson, 2012, Measuring systemic risk, CEPR Discussion Paper.

Acharya, Viral, Thomas Philippon, Matthew Richardson, and Nouriel Roubini, 2009a, The financial crisis of 2007-2009: Causes and remedies, *Financial markets, institutions & instruments* 18, 89–137.

Acharya, Viral, Thomas Philippon, Matthew Richardson, and Nouriel Roubini, 2009b, The financial crisis of 2007-2009: Causes and remedies, *Financial markets, institutions & instruments* 18, 89–137.

Acharya, Viral V, Menachem Brenner, Robert Engle, Anthony Lynch, and Matthew Richardson, 2009, Derivatives-the ultimate financial innovation, *Financial Markets, Institutions & Instruments* 18, 166–167.

Acharya, Viral V., Irvind Gujral, Nirupama Kulkarni, and Hyun Song Shin, 2011, Dividends and bank capital in the financial crisis of 2007-2009, .

Acharya, Viral V., and Matthew Richardson, 2009, Causes of the financial crisis, *Critical Review* 21, 195–210.

Acharya, Viral V., Philipp Schnabl, and Gustavo Suarez, 2013, Securitization without risk transfer, *Journal of Financial Economics* 2013, 515–536.

Adelino, Manuel, Antoinette Schoar, and Felipe Severino, 2016, Loan originations and defaults in the mortgage crisis: The role of the middle class, *The Review of Financial Studies* 29, 1635–1670.

Adrian, Tobias, and Markus Brunnermeier, 2016, CoVaR, *American Economic Review* 106, 1705–1741.

Aebia, Vincent, Gabriele Sabatob, and Markus Schmid, 2012, Risk management, corporate governance, and bank performance in the financial crisis, *Journal of Banking and Finance* 26, 32133226.

Allen, Franklin, Ana Babus, and Elena Carletti, 2012, Asset commonality, debt maturity and systemic risk, *Journal of Financial Economics* 104, 519 – 534.

Arora, Navneet, Priyank Gandhi, and Francis A Longstaff, 2012, Counterparty credit risk and the credit default swap market, *Journal of Financial Economics* 103, 280–293.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, 2016, Measuring economic policy uncertainty, *Quarterly Journal of Economics* 131, 1593–1636.

Ball, Christopher, Gerard Hoberg, and Vojislav Maksimovic, 2016, Disclosure, business change, and earnings quality, University of Maryland and University of Southern California Working Paper.

Barber, Brad M., and Terrance Odean, 2007, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21, 785–818.

Barth, Mary E, and Wayne R Landsman, 2010, How did financial reporting contribute to the financial crisis?, *European accounting review* 19, 399–423.

Beatty, Anne, and Scott Liao, 2011, Do delays in expected loss recognition affect banks willingness to lend?, *Journal of Accounting and Economics* 52, 1–20.

Becker, Bo, and Todd Milbourn, 2011, How did increased competition affect credit ratings?, *Journal of Financial Economics* 101, 493–514.

Bekaert, Geert, and Marie Hoerova, 2014, The vix, the variance premium and stock market volatility, *Journal of Econometrics* 183, 181–192.

Beltratti, Andrea, and René M Stulz, 2012, The credit crisis around the globe: Why did some banks perform better?, *Journal of Financial Economics* 105, 1–17.

Benmelech, Efraim, and Jennifer Dlugosz, 2009, The alchemy of cdo credit ratings, *Journal of Monetary Economics* 56, 617–634.

Berger, Allen, and Christa Bouwman, 2009, Bank liquidity creation, *RFS* 22, 3779–3837.

Berger, Allen, and Christa Bouwman, 2013, How does capital affect bank performance during financial crises?, *JFE* 109, 146–176.

Billio, Monica, Mila Getmansky, Andrew W. Lo, and Loriana Pelizzon, 2012, Econometric measures of connectedness and systemic risk in the finance and insurance sectors, *Journal of Financial Economics* 104, 535 – 559.

Bisias, Dimitrios, Mark Flood, Andrew W. Lo, and Stavros Valavanis, 2012, A survey of systemic risk analytics, *Annual Review of Financial Economics* 4, 255–296.

Blei, David, A Ng, and M Jordan, 2003, Latent dirichlet allocation, *Journal of Machine Learning Research* 3, 993–1002.

Bolton, Patrick, Xavier Freixas, and Joel Shapiro, 2012, The credit ratings game, *The Journal of Finance* 67, 85–111.

Bond, Philip, Alex Edmans, and Itay Goldstein, 2012, The real effects of financial markets, *Annual Review of Financial Economics* 4, 339–360.

Bouvard, Matthieu, Pierre Chaigneau, and Adolfo De Motta, 2016, Transparency in the financial system: Rollover risk and crises, *Journal of Finance* 70, 1805–1837.

Brunetti, Celso, Jeffrey H. Harris, Shawn Mankad, and George Michailidis, 2018, Interconnectedness in the interbank market, *Journal of Financial Economics* forthcoming.

Brunnermeier, Markus, Gary Gorton, and Arvind Krishnamurthy, 2014, *Risk Topography* . chap. Liquidity Mismatch Measurement (University of Chicago Press).

Bui, Dien, Chih-Yung Lin, and Tse-Chun Lin, 2016, Yesterday once more: Short selling and two banking crises, University of Hong Kong Working Paper.

Bussiere, Matthieu, and Marcel Fratzscher, 2006, Towards a new early warning system of financial crises, *Journal of International Money and Finance* 25, 953–973.

Calomiris, Charles W, and Gary Gorton, 1991, The origins of banking panics: models, facts, and bank regulation, in *Financial markets and financial crises* . pp. 109–174 (University of Chicago Press).

Calomiris, Charles W, and Joseph R Mason, 2003, Fundamentals, panics, and bank distress during the depression, *American Economic Review* 93, 1615–1647.

Campbell, John, Hsinchun Chen, Dan Dhaliwal, Hsin-Min Lu, and Logan Steele, 2014, The information content of mandatory risk factor disclosures in corporate filings, *Review of Accounting Studies* 19, 396–455.

Castiglionesi, Fabio, Fabio Feriozzi, and Guido Lorenzoni, 2017, Financial integration and liquidity crises, *Management Science*.

Chen, Qi, Itay Goldstein, and Wei Jiang, 2007, Price informativeness and investment sensitivity to stock prices, *Review of Financial Studies* 20, 619–650.

Chen, Qi, Itay Goldstein, and Wei Jiang, 2010, Payoff complementarities and financial fragility: Evidence from mutual fund outflows, *Journal of Financial Economics* 97, 239–262.

Cimiano, Peter, 2006, *Ontology Learning and Population from Text: Algorithms, Evaluation and* (Springer: New York).

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, 2018, Lazy prices, Harvard Business School working paper.

Cole, Rebel, and Lawrence White, 2011, Deja vu all over again: The causes of u.s. commercial bank failures *this* time around, *Journal of Financial Services Research* 42, 5–29.

Cornett, Marcia Millon, Jamie John McNutt, Philip E. Strahan, and Hassan Tehranian, 2011, Liquidity risk management and credit supply in the financial crisis, *Journal of Financial Economics* 101, 297–312.

Covitz, Daniel, Nellie Liang, and Gustavo Suarez, 2013, The evolution of a financial crisis: Collapse of the asset-backed commercial paper market, *Journal of Finance* 68, 815–848.

Dang, Tri Vi, Gary Gorton, Bengt Holstrom, and Guillermo Ordonez, 2016, Banks as secret keepers, Yale University Working Paper.

Debondt, Werner, and Richard Thaler, 1985, Does the stock market overreact?, *Journal of Finance* 40, 793–805.

Demirguc-Kunt, Asli, Enrica Detragiache, and Ouarda Merrouche, 2013, Bank capital: Lessons from the financial crisis, *Journal of Money, Credit and Banking* 45, 1147–1164.

Demyanyk, Yuliya, and Otto Van Hemert, 2011, Understanding the subprime mortgage crisis, *RFS* 24, 1848–1880.

DeYoung, Robert, and Gokhan Torna, 2013, *Nontraditional Banking Activities and Bank Failures During the Financial Crisis* . , vol. 22 (Journal of Financial Intermediation).

Diamond, Douglas, and Phillip Dybvig, 1983, Bank runs, deposit insurance, and liquidity, *Journal of Political Economy* 91, 401–419.

Diamond, Douglas, and Robert Verrecchia, 1987, Constraints on short-selling and asset price adjustment to new information, *JFE* 18, 277–311.

Diamond, Douglas W, and Raghuram G Rajan, 2001, Banks, short-term debt and financial crises: theory, policy implications and applications, in *Carnegie-Rochester conference series on public policy* , vol. 54 pp. 37–71. Elsevier.

Diebold, Francis X, and Kamil Yılmaz, 2014, On the network topology of variance decompositions: Measuring the connectedness of financial firms, *Journal of Econometrics* 182, 119–134.

Duca, Marco Lo, and Tuomas A. Peltonen, 2013, Assessing systemic risks and predicting systemic events, *Journal of Banking & Finance* 37, 2183–2195.

Duffie, Darrell, 2010, The failure mechanics of dealer banks, *Journal of Economic Perspectives* 24, 51–72.

Duffie, Darrell, and Haoxiang Zhu, 2011, Does a central clearing counterparty reduce counterparty risk?, *The Review of Asset Pricing Studies* 1, 74–95.

Edward, Franklin R, 1999, Hedge funds and the collapse of long-term capital management, *Journal of Economic Perspectives* 13, 189–210.

Elliot, Matthew, Benjamin Golub, and Matthew Jackson, 2014, Financial networks and contagion, *American Economic Review* 104, 3115–3153.

Estrella, Arturo, and Frederic Mishkin, 2016, Predicting u.s. recessions: Financial variables as leading indicators, *The Review of Economics and Statistics* pp. 45–61.

Fahlenbrach, Rudiger, Robert Prilmeier, and Rene Stulz, 2012, This time is the same: Using bank performance in 1998 to explain bank performance during the recent financial crisis, *Journal of Finance* 67, 2139–2185.

Fahlenbrach, Rudiger, and Rene M. Stulz, 2011, Bank ceo incentives and the credit crisis, *Journal of Financial Economics* 99, 11–26.

Fama, Eugene, and J. MacBeth, 1973, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy* 71, 607–636.

Ferri, Giovanni, L-G Liu, and Joseph E Stiglitz, 1999, The procyclical role of rating agencies: Evidence from the east asian crisis, *Economic Notes* 28, 335–355.

Firth, John Rupert, 1957, *A synopsis of linguistic theory 1930-55* (Frank Palmer) published 1968.

Flannery, Mark J., Simon H. Kwan, and Mahendrarajah Nimalendran, 2013, The 20072009 financial crisis and bank opaqueness, *Journal of Financial Intermediation* 22, 55–84.

Frankel, Jeffrey, and George Saravelos, 2012, Can leading indicators assess country vulnerability? evidence from the 2008-09 global financial crisis, *Journal of International Economics* 87, 216–231.

Giesecke, Kay, and Baeho Kim, 2011, Systemic risk: What defaults are telling us, *Management Science* 57, 1387–1405.

Giglio, Stefano, Bryan Kelly, and Seth Pruitt, 2016, Systemic risk and the macroeconomy: An empirical evaluation, *Journal of Financial Economics* 119, 457–471.

Glosten, Lawrence R, and Paul R Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of financial economics* 14, 71–100.

Goel, Anand M., and Anjan V. Thakor, 2015, Information reliability and welfare: a theory of coarse credit ratings, *Journal of Financial Economics* 115, 541–557.

Goldstein, Itay, 2005, Strategic complementarities and the twin crises, *The Economic Journal* 115, 368–390.

Goldstein, Itay, and Assaf Razin, 2015, Three branches of theories of financial crises, *Foundations and Trends® in Finance* 10, 113–180.

Gorton, Gary, and Guillermo Ordonez, 2014, Collateral crises, *American Economic Review* 104, 343–378.

Gorton, Gary, and George Pennacchi, 1990, Financial intermediaries and liquidity creation, *Journal of Finance* 45, 49–71.

Griffin, John M., Jordan Nickerson, and Dragon Yongjun Tang, 2013, Rating shopping or catering?an examination of the response to competitive pressure for cdo credit ratings, *Review of Financial Studies* 26, 2270–2310.

Griffin, John M., and Dragon Yongjun Tang, 2012, Did subjectivity play a role in cdo credit ratings?, *The Journal of Finance* 67, 1293–1328.

Grossman, Sanford J, and Joseph E Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American economic review* 70, 393–408.

Hayek, F.A., 1945, The use of knowledge in society, *American Economic Review* 35, 519–530.

Hellwig, Christian, Arijit Mukherji, and Aleh Tsyvinski, 2006, Self-fulfilling currency crises: The role of interest rates, *American Economic Review* 96, 1769–1787.

Herring, Richard, and Susan Wachter, 1999, Real estate booms and banking busts: An international perspective, in *Real estate booms and banking busts: An international perspective* (Group of Thirty).

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies in mergers and acquisitions: A text based analysis, *Review of Financial Studies* 23, 3773–3811.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-based network industry classifications and endogenous product differentiation, *Journal of Political Economy* 124, 1423–1465.

Huang, Xin, Hao Zhou, and Haibin Zhu, 2009, A framework for assessing the systemic risk of major financial institutions, *Journal of Banking and Finance* 33, 2036–2049.

Huizinga, Harry, and Luc Laeven, 2012, Bank valuation and accounting discretion during a financial crisis, *Journal of Financial Economics* 106, 614–634.

Ivashina, Victoria, and David Scharfstein, 2016, Bank lending during the financial crisis of 2008, *Journal of Financial Economics* 97, 319–338.

Kacperczyk, Marcin, and Philipp Schnabl, 2010, When safe proved risky: Commercial paper during the Financial Crisis of 20072009, *Journal of Economic Perspectives* 24, 29–50.

Kaminsky, Graciela L, and Carmen M Reinhart, 1998, Financial crises in asia and latin america: Then and now, *The American Economic Review* 88, 444–448.

Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo, 2010, Consumer credit risk models via machine-learning algorithms, *Journal of Banking and Finance* 34, 2767–2787.

Khandani, Amir E., Andrew W. Lo, and Robert C. Merton, 2012, Systemic risk and the refinancing ratchet effect, *Journal of Financial Economics* 108, 29–45.

Kravet, Todd, and Volkan Muslu, 2013, Textual risk disclosures and investors risk perceptions, *Review of Accounting Studies* 18, 1088–1122.

Kritzman, Mark, Yuanzhen Li, Sebastien Page, and Roberto Rigobon, 2011, Principal components as a measure of systemic risk, *Journal of Portfolio Management* 37, 112–126.

Kyle, Albert, 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–35.

Laeven, Luc, and Ross Levine, 2009, Bank governance, regulation and risk taking, *Journal of financial economics* 93, 259–275.

Laux, Christian, and Christian Leuz, 2009, Did fair-value accounting contribute to the financial crisis?, *Journal of Economic Perspectives* 24, 93–118.

Longstaff, Francis A, 2010, The subprime credit crisis and contagion in financial markets, *Journal of financial economics* 97, 436–450.

Loughran, Tim, and Bill McDonald, 2014, Measuring readability in financial text, *JF* 69, 1643–1671.

Merton, Robert, 1986, A simple model of capital market equilibrium with incomplete information, *JF* 42, 482–510.

Mian, Atif, and Amir Sufi, 2009, The consequences of mortgage credit expansion: Evidence from the u.s. mortgage default crisis, *The Quarterly Journal of Economics* 124, 1449–1496.

Mian, Atif, and Amir Sufi, 2011, House prices, home equity-based borrowing, and the US household leverage crisis, *American Economic Review* 101, 2132–2156.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, 2013, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean, 2013, Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems* pp. 3111–3119.

Miller, Edward M., 1977, Risk, uncertainty, and divergence of opinion, *Journal of Finance* 32, 1151–1168.

Peek, Joe, and Eric S Rosengren, 2000, Collateral damage: Effects of the japanese bank crisis on real activity in the united states, *American Economic Review* 90, 30–45.

Peek, Joe, Eric S Rosengren, and Geoffrey MB Tootell, 2003, Does the federal reserve possess an exploitable informational advantage?, *Journal of Monetary Economics* 50, 817–839.

Pena, Emilia, and Sami Vahamaa, 2012, Did good corporate governance improve bank performance during the financial crisis?, *Journal of Financial Services Research* 41, 19–35.

Peristian, Stavros, Donald P. Morgan, and Vanessa Savino, 2010, The information value of the stress test and bank opacity, .

Reinhart, Carmen M, and Kenneth S Rogoff, 2008, Is the 2007 us sub-prime financial crisis so different? an international historical comparison, *American Economic Review* 98, 339–44.

Reinhart, Carmen M, and Kenneth S Rogoff, 2009, The aftermath of financial crises, *American Economic Review* 99, 466–72.

Reinhart, Carmen M, and Kenneth S Rogoff, 2011, From financial crash to debt crisis, *American Economic Review* 101, 1676–1706.

Roberts, Michael R., and Amir Sufi, 2009, Renegotiation of financial contracts: Evidence from private credit agreements., *Journal of Financial Economics* 93, 159–184.

Sarkar, Sumit, and Ram S. Sriram, 2001, Bayesian models for early warning of bank failures., *Management Science* 47, 1457–1475.

Schwarcz, Steven L, 2008, Systemic risk, *Geo. LJ* 97, 193.

Shleifer, Andrei, and Robert Vishny, 1997, The limits of arbitrage, *Journal of Finance* 52, 35–55.

Skreta, Vasiliki, and Laura Veldkamp, 2009, Ratings shopping and asset complexity: A theory of ratings inflation, *Journal of Monetary Economics* 56, 678–695.

Stulz, Rene, 2010, Credit default swaps and the credit crisis, *Journal of Economic Perspectives* 24, 79–92.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Veldkamp, Laura, 2006, Information markets and the comovement of asset prices, *Review of Economic Studies* 73, 823–845.

Figure 1: Aggregate Emerging Risk Measure

Aggregate measure of emerging risk from the static emerging risks model. The measure is the normalized adjusted R^2 contribution to pairwise return covariance of bank stocks of all of the 31 semantic themes extracted from 10-K disclosed bank risk factors from 1998 to 2016.

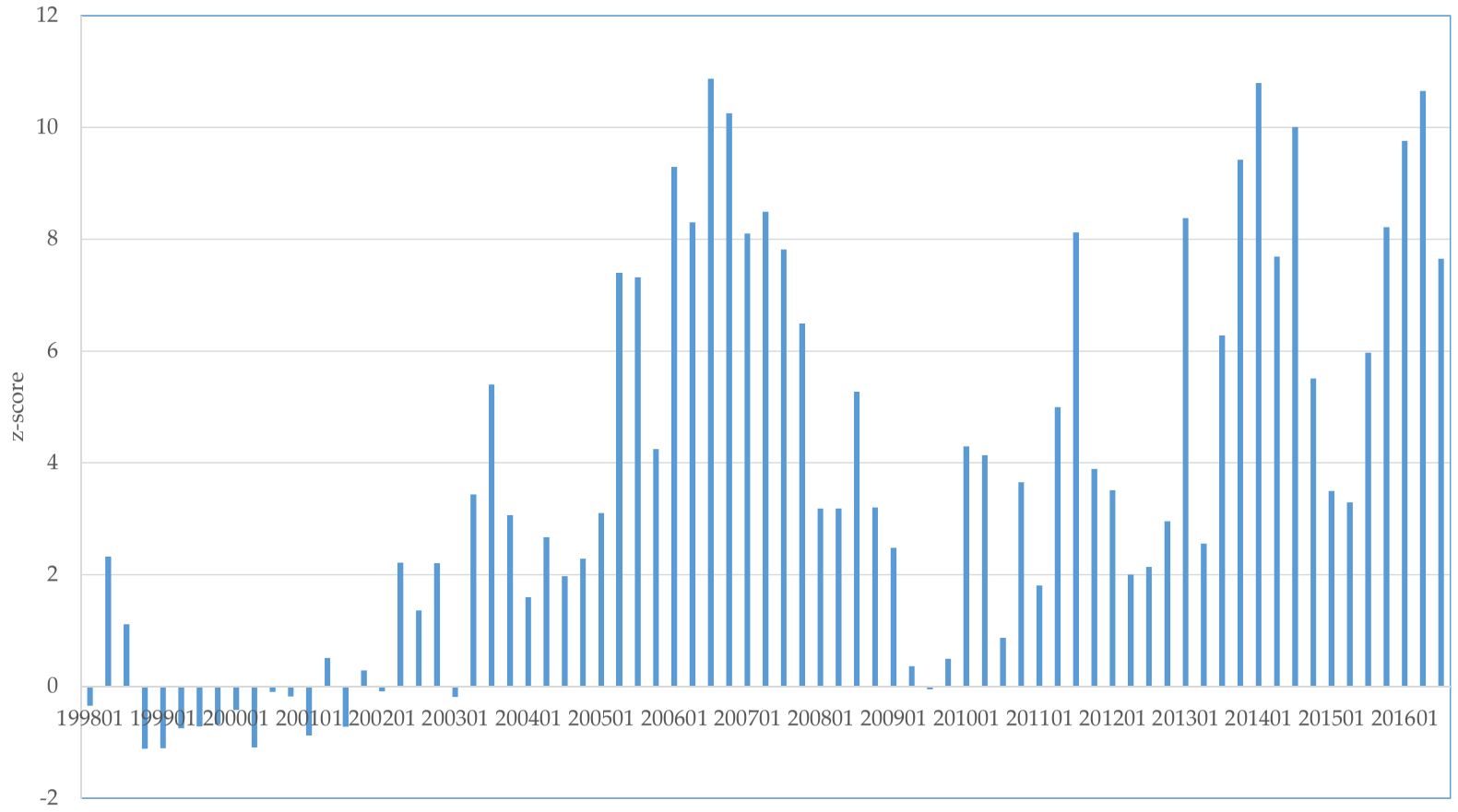


Figure 2: Emerging Risks Using LDA with 25 Topics

Overview of the 25 risk factors detected by metaHeuristica from fiscal year 2006 10-Ks of banks. Each box is ranked and sized relative to its importance in the document and contains the five most prevalent words or commongrams in the topic.

Figure 3: Sample of Banks from 1997 to 2015

Number of banks by year in our sample from 1997 to 2015. There are 10,558 bank-years total. To be included, a bank must be in the CRSP and Compustat databases, must have a SIC code in the range 6000 to 6199, and must be in the metaHeuristica database of 10-Ks with a non-zero number of paragraphs residing in a section of the 10-K that discusses risks.

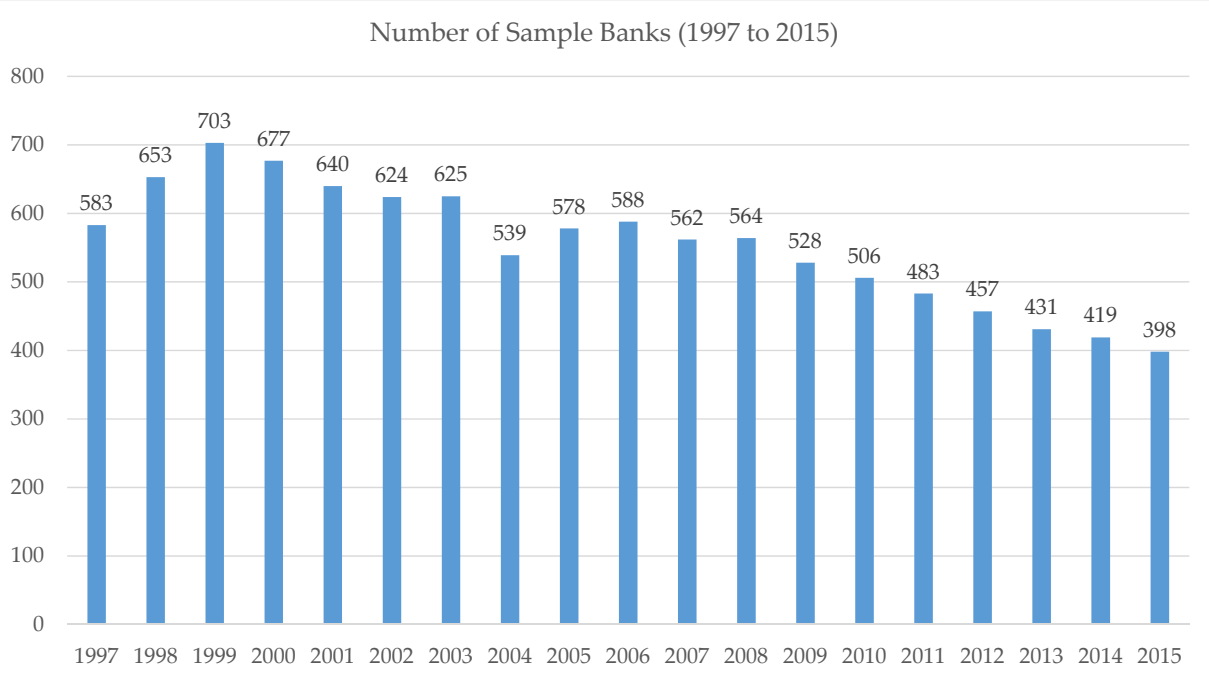


Figure 4: Emerging Risks Comparison

Time series of different risk variables. For our aggregate risk measure, we define the initial part of our sample (1998 to 2002) as a calibration period, and use this period to compute the measure's baseline quarterly mean and standard deviation. In each of the subsequent quarters from 2003 to 2016, we compute a z score based on how many standard deviations the current value is from the baseline mean. Panel A displays the time series of the levels of the VIX index, the quarterly average pairwise covariance among bank-pairs, the average quarterly standard deviation of monthly returns across all stocks in the CRSP database and for financial firms only (SIC codes from 6000 to 6199) and the Economic Policy Uncertainty (EPU) index from Baker, Bloom, and Davis (2016). Panel B reports z scores for the R^2 of the covariance model including only accounting and bank characteristics (no text), the static model (as reproduced from Figure 1), the dynamic model and an LDA-only model.

Panel A: Risk Metrics

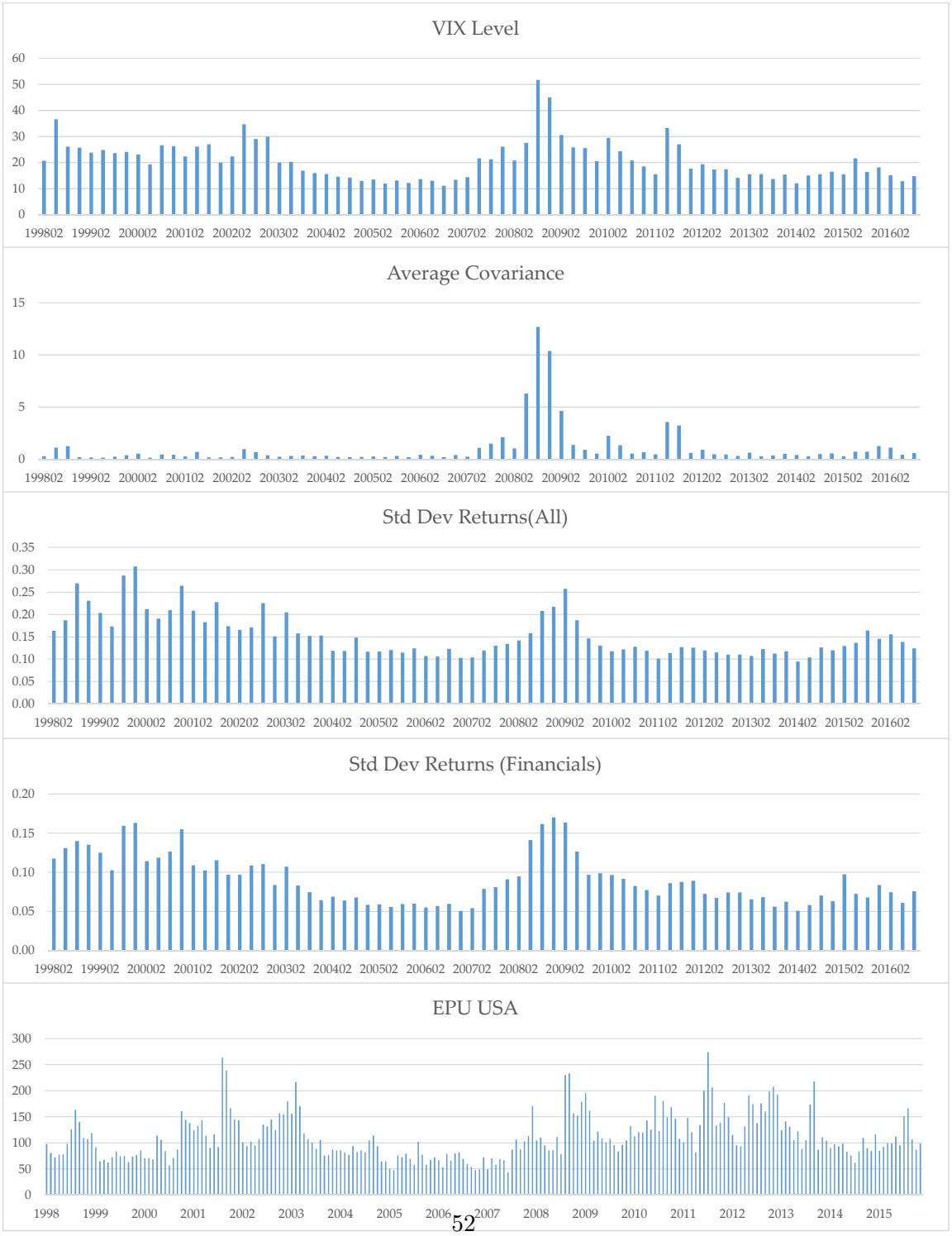


Figure 4: Emerging Risks Comparison

Panel B: Covariance Models

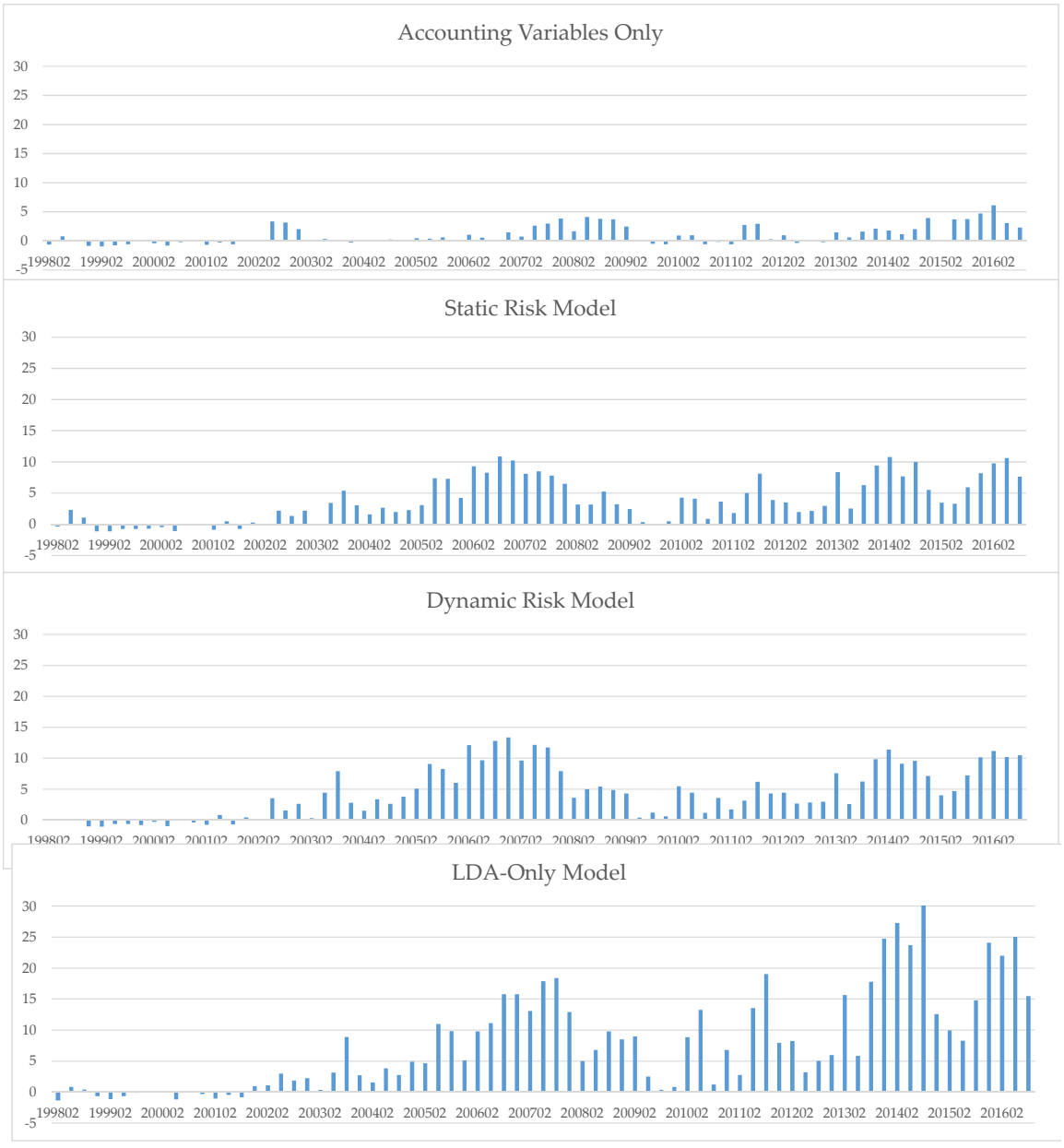


Figure 5: Crisis Period Emerging Risks

Time series of z scores of the R^2 from the model in Equation (3) for the most prominent emerging risk in 2008 (Appendix A presents all 31 semantic theme emerging risks). The results are based on the time series of the contribution of individual semantic themes in explaining pairwise covariance of banks. We define the initial part of our sample (1998 to 2002) as a calibration period, and use this period to compute each semantic themes' R^2 baseline quarterly mean and standard deviation. In each of the subsequent quarters from 2003 to 2016, we compute a z score based on how many standard deviations the current value is from the baseline mean. The figure is a plot of the quarterly z score for each semantic theme.

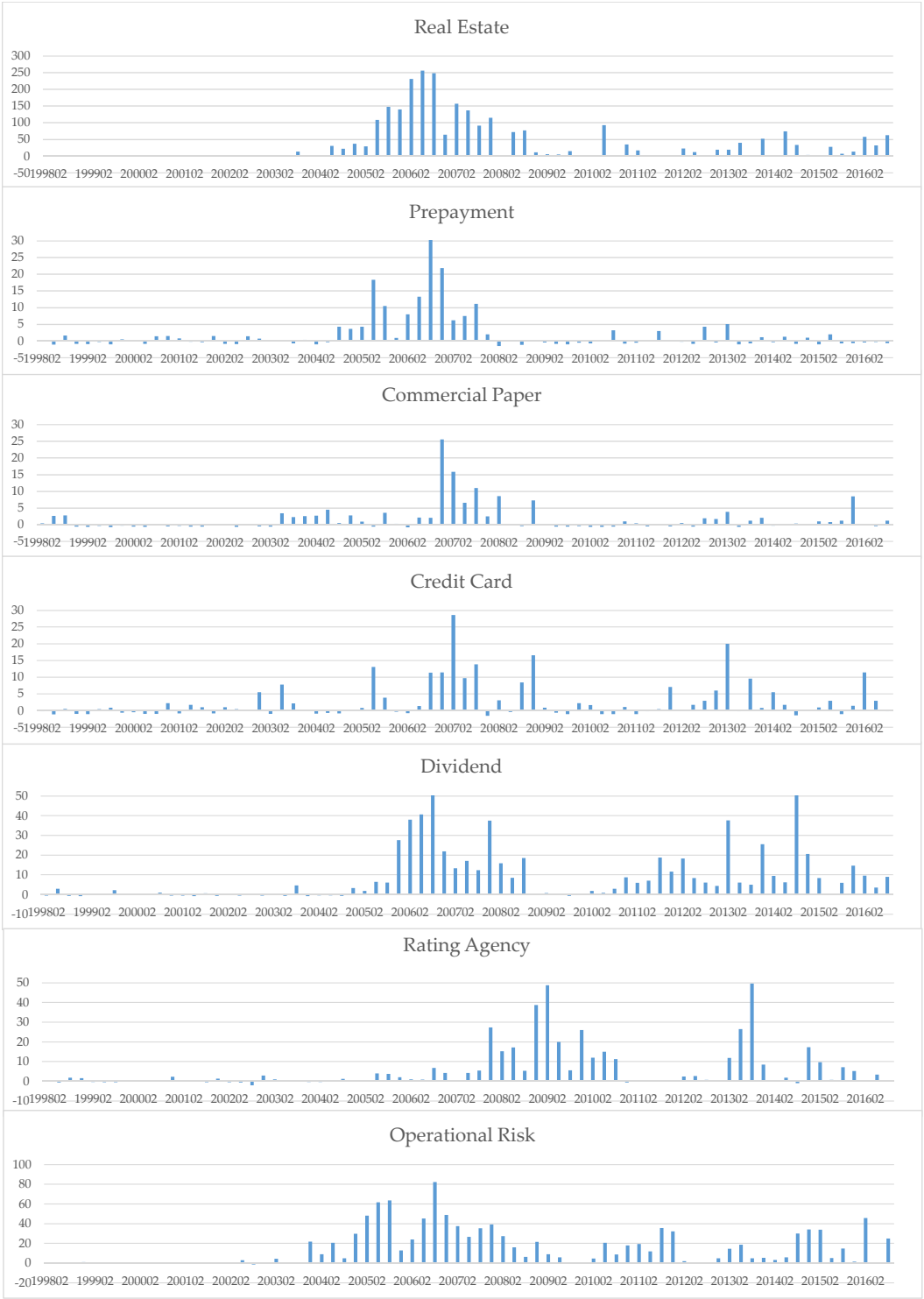


Figure 6: Current Period Emerging Risks

Time series of z scores of the R^2 from the model in Equation (3) for the most prominent emerging risks in 2014-2016 (Appendix A presents all 31 semantic theme emerging risks). The results are based on the time series of the contribution of individual semantic themes in explaining pairwise covariance of banks. We define the initial part of our sample (1998 to 2002) as a calibration period, and use this period to compute each semantic themes' R^2 baseline quarterly mean and standard deviation. In each of the subsequent quarters from 2003 to 2016, we compute a z score based on how many standard deviations the current value is from the baseline mean. The figure is a plot of the quarterly z score for each semantic theme.

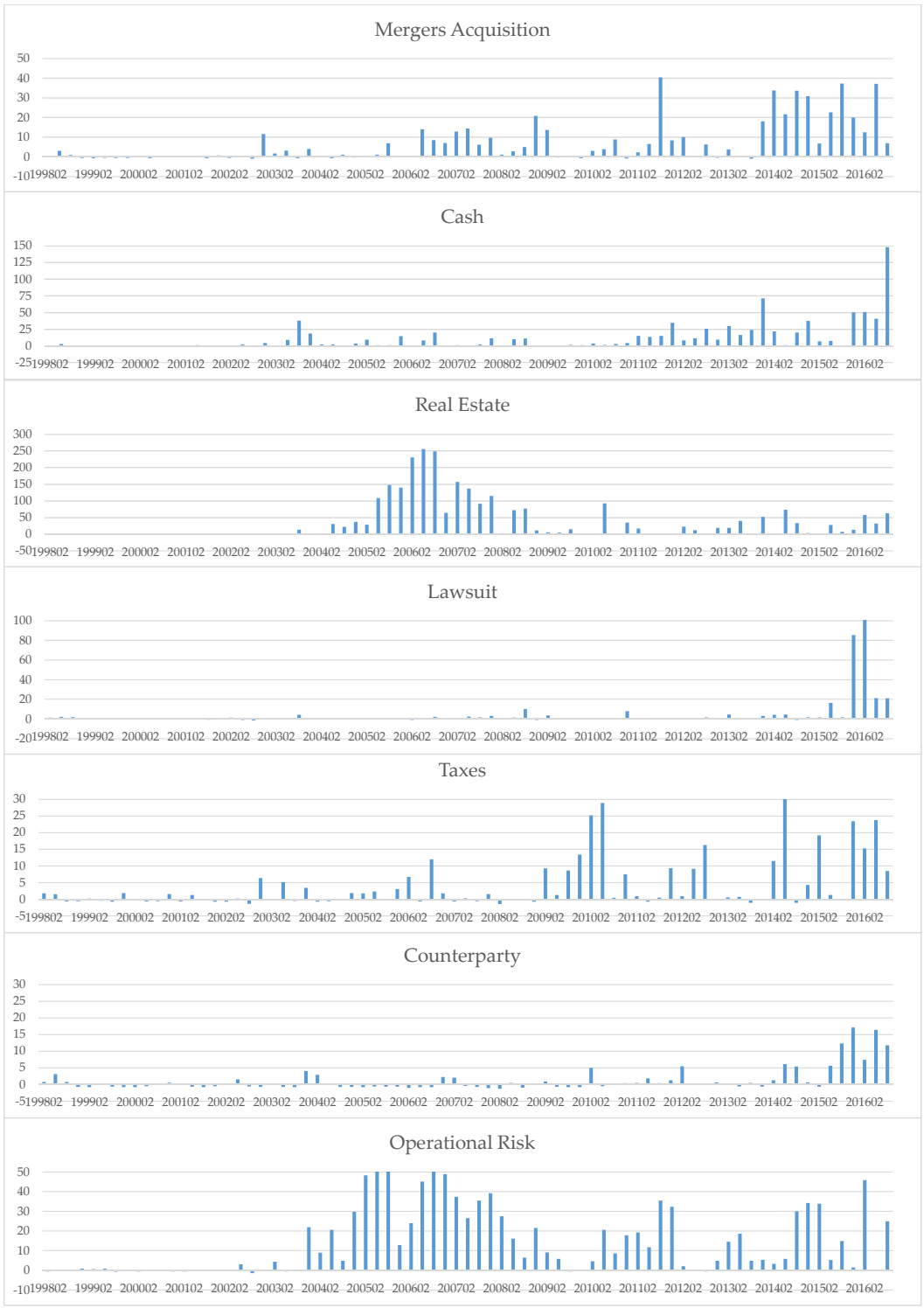


Figure 7: Drill-Down Emerging Risks

Time series of z scores for sub-themes related to the semantic theme "real estate" and to vocabulary related to the Sovereign debt crisis. We define the initial part of our sample (1998 to 2002) as a calibration period, and use this period to compute each semantic sub-themes' R^2 baseline quarterly mean and standard deviation. In each of the subsequent quarters from 2003 to 2016, we compute a z score based on how many standard deviations the current value is from the baseline mean. The figure is a plot of the quarterly z score for each semantic theme.

Panel A: Real Estate

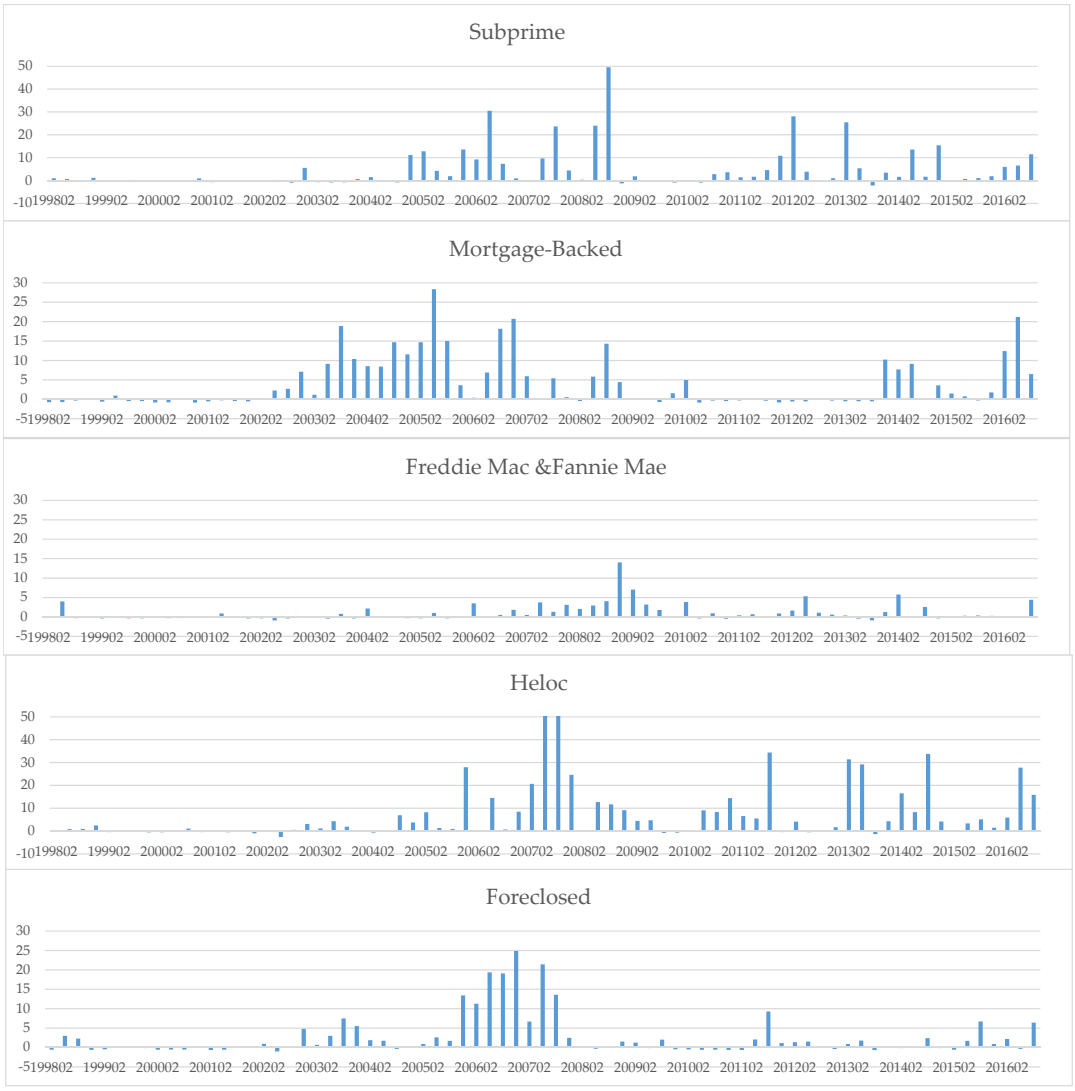


Figure 7: Drill-Down Emerging Risks (continued)

.

Panel B: Sovereign Debt Crisis

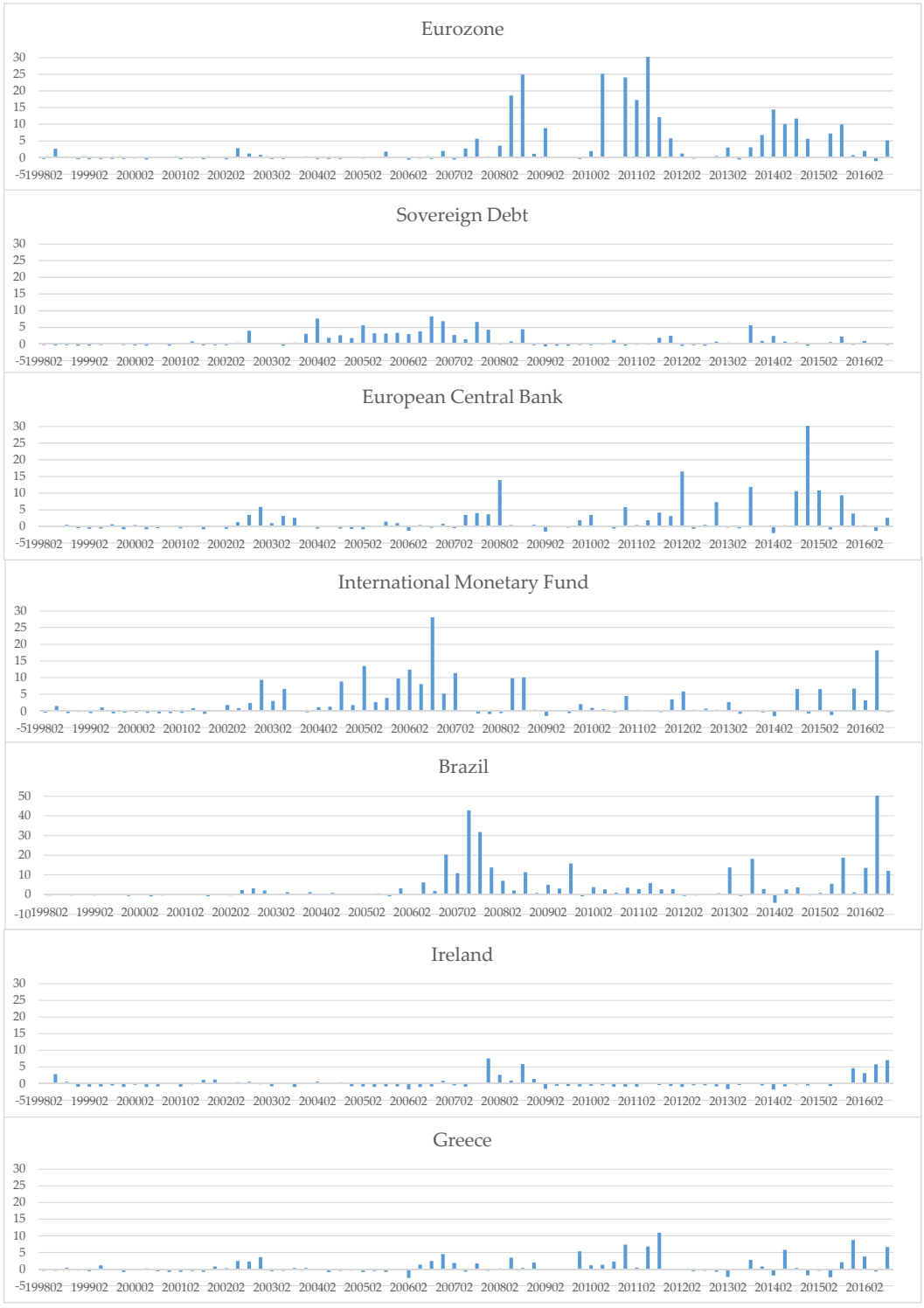


Table I: Examples of Semantic Vectors

Focal word and phrase lists for four semantic themes obtained using Latent Dirichlet Allocation and Semantic Vector Analysis of the risk factor discussion of publicly traded banks (those having SIC codes in the range 6000 to 6199). The title of each theme is the short one to two word phrase noted in the column headers. For each of the four themes, we include two columns. The first is the list of specific words or phrases identified by the Semantic Vector module in metaHeuristica as being similar to the theme's title. The second is each word's cosine similarity to the theme's title.

	Real Estate		Deposits		Mergers & Acquisitions		Regulatory Capital	
Row	Word	Cosine Dist	Word	Cosine Dist	Word	Cosine Dist	Word	Cosine Dist
1	real	0.7875	deposits	1.0000	acquisitions	0.8587	capital	0.8458
2	estate	0.7875	deposit	0.7046	mergers	0.8587	regulatory	0.8458
3	foreclosure	0.4898	brokered deposits	0.5930	mergers acquisitions	0.5876	regulators	0.5944
4	property	0.4619	cdars	0.5864	strategic	0.5671	prompt corrective	0.5518
5	personal	0.4563	account registry	0.5712	businesses	0.5636	adequacy guidelines	0.5452
6	physical possession	0.4539	brokered certificates	0.5680	consolidations	0.5426	fnbpa	0.5044
7	foreclosed real	0.4503	bearing checking	0.5657	incurrence	0.5093	additional	0.5008
8	foreclosed	0.4423	bearing deposits	0.5650	divestitures	0.4990	requirements	0.4905
9	deed	0.4323	certificates	0.5632	acquisition	0.4988	addition	0.4802
10	beneficiary	0.4283	negotiable order	0.5154	opportunities	0.4913	involve quantitative	0.4787
11	real estate	0.4262	promontory interfinancial	0.5129	complementary businesses	0.4878	maintain	0.4779
12	possession	0.4147	cdars program	0.5067	merge	0.4789	regulation	0.4750
13	oreo	0.4063	sweep ics	0.4950	strategic alliances	0.4781	banking agencies	0.4720
14	lien	0.4044	brokered	0.4943	financings	0.4764	regulatory agencies	0.4716
15	securing	0.4039	withdrawal	0.4804	integrating	0.4761	tier 1	0.4695
16	h2c	0.4014	overdrafts	0.4738	successfully integrate	0.4747	quantitative measures	0.4684
17	owned	0.3996	sweep accounts	0.4726	accretive	0.4644	bank	0.4630
18	repossessed	0.3981	bearing	0.4591	synergies	0.4618	regulatory authorities	0.4592
19	death	0.3974	cdars network	0.4547	complementary	0.4603	occ	0.4543
20	owner	0.3949	fdic insured	0.4505	engage	0.4563	qualitative judgments	0.4533
21	corporations partnerships	0.3924	esavings	0.4501	organic growth	0.4468	regulations	0.4509
22	lieu	0.3858	fhlbank advances	0.4462	identify suitable	0.4468	approval	0.4472
23	partial satisfaction	0.3847	jumbo certificates	0.4458	integration	0.4437	hampden bank's	0.4452
24	itec	0.3842	fund dif	0.4457	consummate	0.4424	liquidity	0.4442
25	liangcai zhang	0.3837	passbook	0.4433	restrictive covenants	0.4409	adequacy	0.4352
26	encumbrance	0.3763	saving deposits	0.4409	finance	0.4361	certain	0.4326
27	inheritance	0.3746	include passbook	0.4395	pursue	0.4306	must	0.4317
28	favor	0.3712	denominations	0.4333	mobilitie	0.4287	supervisory	0.4316
29	encumbrances	0.3674	accounts mmdas	0.4322	strategy	0.4286	fsgbank	0.4314
30	fee simple	0.3656	noninterest bearing	0.4260	consummating	0.4252	banking regulators	0.4297

Table II: Summary Statistics

Summary statistics for our sample of 10,558 bank-year observations from 1997 to December 2015. Bank characteristics in Panel A is based on Compustat data and includes *Ln(Assets)* and *Ln(Bank Age)*, the time since the first appearance in CRSP. Panel B is based on Call Reports and includes *Cash/Assets*, *Loans/Assets*, *Loan Loss Prov & Allow/Assets*, the sum of loan loss provision and allowances to assets, *Capital*, the ratio of equity to assets, *Negative Earnings Dummy* an indicator variable equal to one if net income is negative, zero otherwise, *Bank Holding Co. Dummy*, an indicator variable equal to one if the bank has a parent, zero otherwise, *Non-Performing Assets*, the sum of loans that are 30 days and 90 days past due, and *CatFat/Assets* from Berger and Bouwman (2009). Panel C reports summary statistics based on bank-pair-quarter observations (19.7 million observations). The bank-pair daily covariance is the quarterly covariance of daily stock returns for a pair of banks winsorized in each quarter at the 1/99% level. Bank-pair SIC variables are dummy variables equal to one if the pair of banks is in the same 2, 3 or 4 digit SIC-based industry, zero otherwise. The TNIC similarity for a pair of banks is from Hoberg and Phillips (2010). Panel D reports statistics for the time series variables. There are 76 quarterly observations in our database from 1998 to 2016. The average pair covariance is the quarterly average pairwise covariance among bank-pairs. The average quarterly standard deviation of monthly returns is calculated across all stocks in the CRSP database and for financial firms only (SIC codes from 6000 to 6199). The Economic Policy Uncertainty (EPU) index from Baker, Bloom, and Davis (2016). The accounting variable adjusted R^2 is the quarterly adjusted R^2 from a regression of bank-pairwise correlation on the bank characteristics and industry variables. The semantic theme variable adjusted R^2 is the incremental improvement to R^2 when textual information is also included in the pairwise covariance regression. Daily covariance figures are multiplied by 10,000 for ease of viewing.

Variable	Mean	Std. Dev.	Minimum	Median	Maximum	# Obs.
Panel A: Bank Characteristics (Compustat)						
Ln(Assets)	7.289	1.635	0.247	6.998	14.986	10,558
Ln(Bank Age)	2.318	0.769	0.000	2.398	4.060	10,558
Panel B: Bank Characteristics (Call Reports)						
Cash/Assets	0.044	0.039	0.001	0.033	0.442	8,167
Loans/Assets	0.507	0.177	0.000	0.513	0.908	8,167
Loss Prov & Allow/Assets	0.002	0.004	0.000	0.000	0.031	8,167
Capital	0.098	0.029	0.000	0.093	0.277	8,167
Negative Earnings Dummy	0.091	0.288	0.000	0.000	1.000	8,167
Bank Holding Co. Dummy	0.842	0.365	0.000	1.000	1.000	8,167
Non-Performing Assets/Assets	0.005	0.007	0.000	0.003	0.054	8,167
CatFat/Assets	0.347	0.249	-0.048	0.385	1.291	8,167
Panel C: Bank-Pair Characteristics						
Bank-Pair Daily Covariance	0.868	3.416	-35.415	0.265	71.993	19.7M
Bank-Pair Same 4-digit SIC	0.469	0.499	0.000	0.000	1.000	19.7M
Bank-Pair Same 3-digit SIC	0.499	0.500	0.000	0.000	1.000	19.7M
Bank-Pair Same 2-digit SIC	0.869	0.338	0.000	1.000	1.000	19.7M
Bank-Pair TNIC Similarity	0.104	0.084	0.000	0.110	0.909	19.7M
Panel D: Time-Series Variables						
VIX Level	20.914	7.522	11.190	20.047	51.723	76
Avg Pair Covariance	1.062	2.015	0.150	0.444	12.704	76
Avg Std Dev Monthly Returns	0.154	0.049	0.095	0.136	0.307	76
Avg Std Dev Monthly Returns (Fin. Only)	0.091	0.031	0.051	0.083	0.170	76
Econ Policy Uncertainty	110.595	33.609	63.118	103.840	215.891	76
Econ Policy Uncertainty (News Only)	116.749	38.424	52.089	111.348	235.084	76
Cov Model R^2 (Acct Vars Only)	0.081	0.063	0.005	0.056	0.279	76
Cov Model R^2 (Text Vars Only)	0.015	0.011	0.001	0.013	0.037	76

Table III: Pearson Correlation Coefficients (Time Series Variables)

Pearson correlation coefficients are reported for the time series variables. There are 76 quarterly observations in our database from 1998 to 2016. The average pair covariance is the quarterly average pairwise covariance among bank-pairs. We also report the average quarterly standard deviation of monthly returns across all stocks in the CRSP database and for financial firms only (SIC codes from 6000 to 6199). The Economic Policy Uncertainty (EPU) index from Baker, Bloom, and Davis (2016). The accounting variable adjusted R^2 is the quarterly adjusted R^2 from a regression of bank-pairwise correlation on the bank characteristics and industry variables. The ACCT Variable Adj R^2 is the incremental improvement to R^2 when industry variables, accounting variables and call report variables are included in the pairwise covariance regression. The Text Variable Adj R^2 is the incremental improvement to R^2 when textual information is also included in the pairwise covariance regression. We report this variable for the static model, the dynamic model and the LDA-based model.

Row	Variable	Vix Index	Avg Pairwise Covariance	Avg Std Dev Returns (All)	Avg Std Dev Returns (Financials)	Economic Policy Uncert. (Base)	Economic Policy Uncert. (News)	Accounting Variable Adj R^2	Semantic Text (Static) Adj R^2	Semantic Text (Dynamic) Adj R^2
(1)	Avg Pair Covariance	0.721								
(2)	Std Dev Monthly Returns (All)	0.557	0.197							
(3)	Std Dev Monthly Returns (Financials)	0.779	0.484	0.875						
(4)	Econ. Policy Uncertainty	0.473	0.454	-0.072	0.182					
(5)	Econ. Policy Uncertainty (News-based)	0.533	0.426	0.062	0.236	0.879				
(6)	ACCT Variable Adj R^2	0.219	0.475	-0.146	-0.072	0.223	0.329			
(7)	Text Variable Adj R^2 (Static)	-0.380	0.077	-0.602	-0.608	-0.111	-0.115	0.676		
(8)	Text Variable Adj R^2 (Dynamic)	-0.400	0.074	-0.595	-0.603	-0.176	-0.180	0.671	0.968	
(9)	Text Variable Adj R^2 (LDA-based)	-0.307	0.112	-0.511	-0.495	-0.007	-0.005	0.719	0.919	0.867

Table IV: Baseline Semantic Themes and Bank Characteristics

OLS regression of the determinants of the 31 semantic themes. The dependent variable is a bank's loading on a given theme, and the independent variables are defined in Table II. t statistics are in parentheses.

Row	Semantic Theme	Log Assets	Loans/ Assets	Loss Prov/ Assets	Cap-ital	Neg. Earn.	CatFat/ Assets	NPA/ Assets	Adj R^2
(1)	accounting	0.002 (3.91)	-0.001 (-0.44)	-0.176 (-9.21)	0.024 (2.10)	-0.001 (-1.58)	-0.004 (-3.59)	0.164 (1.63)	0.041
(2)	cash	0.008 (7.06)	-0.017 (-0.81)	-2.388 (-15.45)	0.090 (2.24)	0.018 (3.64)	0.003 (0.37)	0.330 (2.86)	0.066
(3)	certificate deposit	-0.000 (-1.41)	-0.007 (-2.58)	-0.696 (-10.81)	0.027 (5.49)	0.006 (10.43)	-0.000 (-0.07)	0.285 (15.48)	0.120
(4)	commercial paper	0.001 (2.18)	-0.002 (-1.93)	-0.290 (-6.25)	0.009 (3.63)	0.002 (6.11)	-0.000 (-0.60)	0.101 (1.48)	0.067
(5)	compensation	0.001 (3.15)	-0.002 (-0.56)	-0.040 (-1.87)	0.023 (1.51)	0.004 (6.16)	-0.002 (-2.30)	0.101 (2.71)	0.021
(6)	competition	0.002 (14.22)	-0.007 (-1.90)	-1.838 (-11.65)	0.026 (2.23)	0.010 (8.43)	0.005 (1.09)	0.236 (1.80)	0.194
(7)	counterparty	0.002 (15.83)	0.002 (1.68)	0.344 (7.93)	-0.006 (-0.75)	-0.001 (-0.96)	0.001 (1.73)	-0.045 (-1.22)	0.177
(8)	credit card	0.001 (2.15)	0.003 (2.35)	-0.052 (-0.56)	-0.006 (-0.44)	0.003 (2.12)	0.003 (2.33)	0.249 (2.44)	0.035
(9)	currency exchange	0.001 (7.51)	-0.002 (-1.64)	0.245 (9.49)	0.007 (1.04)	0.000 (0.31)	0.002 (2.63)	-0.073 (-3.09)	0.145
(10)	data security	0.002 (10.07)	-0.005 (-6.97)	-0.419 (-6.33)	0.016 (1.94)	0.003 (2.84)	-0.002 (-4.69)	0.102 (6.15)	0.115
(11)	deposits	-0.000 (-1.96)	-0.005 (-1.78)	-0.453 (-7.48)	0.029 (5.09)	0.004 (8.75)	0.000 (0.28)	0.163 (6.84)	0.101
(12)	derivative	0.003 (14.27)	0.004 (3.33)	0.615 (11.52)	-0.012 (-1.25)	-0.002 (-2.85)	0.001 (1.11)	-0.097 (-1.66)	0.152
(13)	dividends	0.002 (10.04)	-0.006 (-0.95)	-1.334 (-26.17)	0.020 (5.51)	0.017 (10.04)	-0.004 (-0.95)	0.465 (8.37)	0.117
(14)	fees	0.004 (21.48)	-0.004 (-0.59)	-0.713 (-13.58)	0.037 (2.44)	0.009 (5.14)	-0.005 (-1.78)	0.196 (3.61)	0.070
(15)	funding sources	0.004 (23.60)	-0.009 (-2.75)	-1.839 (-9.49)	0.040 (3.23)	0.016 (10.18)	0.003 (0.64)	0.419 (3.50)	0.179
(16)	governance	0.001 (4.91)	-0.001 (-1.14)	-0.042 (-14.70)	0.006 (1.04)	0.001 (5.47)	-0.001 (-2.78)	0.020 (3.55)	0.061
(17)	growth strategy	0.004 (18.54)	-0.012 (-1.66)	-2.464 (-14.32)	0.041 (4.52)	0.016 (11.68)	0.005 (0.81)	0.388 (1.69)	0.211
(18)	insurance	0.001 (7.77)	-0.004 (-3.93)	-0.483 (-10.54)	0.017 (4.99)	0.003 (8.62)	-0.000 (-0.03)	0.152 (7.34)	0.125
(19)	internal controls	0.003 (12.14)	-0.006 (-3.84)	-0.590 (-5.16)	0.021 (2.95)	0.003 (2.78)	0.004 (2.24)	0.151 (1.95)	0.181
(20)	lawsuit	0.000 (2.61)	-0.000 (-0.37)	-0.004 (-0.76)	-0.000 (-0.44)	0.000 (3.72)	-0.000 (-2.21)	0.003 (2.48)	0.034
(21)	mergers acquisition	0.002 (11.67)	-0.007 (-4.40)	-1.099 (-15.63)	0.032 (5.26)	0.006 (7.28)	0.003 (1.07)	0.118 (1.76)	0.225
(22)	off balance sheet	-0.000 (-1.93)	0.004 (7.44)	0.479 (8.24)	-0.008 (-1.00)	-0.001 (-1.28)	0.004 (3.68)	0.028 (1.37)	0.075
(23)	operational risk	0.009 (23.20)	-0.016 (-2.84)	-2.248 (-7.51)	0.065 (4.97)	0.015 (5.53)	0.009 (1.47)	0.386 (1.49)	0.250
(24)	prepayment	0.000 (0.39)	0.004 (1.37)	0.069 (3.42)	-0.009 (-0.79)	-0.004 (-1.72)	-0.001 (-0.64)	-0.168 (-3.40)	0.113
(25)	rating agencies	0.000 (4.80)	-0.001 (-8.28)	-0.023 (-6.14)	0.005 (4.30)	0.000 (4.69)	-0.000 (-0.14)	0.031 (8.58)	0.126
(26)	real estate	-0.000 (-2.26)	0.000 (0.05)	-0.552 (-13.00)	0.001 (0.33)	0.004 (8.07)	0.000 (0.63)	0.120 (1.20)	0.088
(27)	regulatory capital	0.007 (25.75)	-0.020 (-2.64)	-3.634 (-14.98)	0.062 (3.92)	0.032 (13.04)	-0.004 (-0.47)	0.827 (3.63)	0.198
(28)	reputation	0.001 (10.86)	-0.006 (-10.66)	-1.051 (-12.36)	0.020 (2.10)	0.006 (13.92)	-0.001 (-0.32)	0.092 (1.20)	0.206
(29)	securitization	0.001 (8.23)	0.002 (1.21)	-0.020 (-0.82)	-0.007 (-0.91)	0.000 (0.77)	-0.002 (-4.64)	0.028 (0.92)	0.372
(30)	student loans	0.001 (5.57)	0.005 (1.23)	-0.423 (-8.50)	-0.020 (-3.13)	0.003 (1.99)	-0.001 (-1.24)	0.185 (2.07)	0.088
(31)	taxes	0.002 (2.97)	-0.003 (-0.25)	-0.038 (-0.84)	0.042 (1.68)	-0.001 (-0.43)	-0.001 (-0.39)	-0.133 (-1.64)	0.034

Table V: Dynamic SVA Model

The table reports the emerging risk factors discovered using the dynamic model where LDA-bigrams are directly input into semantic vector calculations and then are run dynamically using stepwise regression to maximize the R^2 of the covariance regression in equation (3). The table reports the bigram identifying each emerging risk factor and the year it emerged. z scores indicate significance of each emerging risk based on a regression where quarterly adjusted R^2 from the covariance regression are regressed on a dummy indicator for the given year reported. To ensure no look ahead bias, each regression considers a four year moving window of adjusted R^2 ending in the year being tested.

Row	Emerging Risk	Year	z-score	Row	Emerging Risk	Year	z-score
1	related litigation	200401	10.80	51	real property	200803	8.93
2	deposits borrowings	200401	10.38	52	legal proceedings	200804	9.80
3	material adverse	200401	10.37	53	mergers acquisitions	200901	12.03
4	notional amount	200402	8.61	54	regulatory approval	201002	8.57
5	mortgage banking	200403	35.98	55	cost funds	201003	10.04
6	operational risk	200403	12.04	56	economic downturn	201103	11.92
7	charged off	200403	9.50	57	education loans	201103	7.69
8	origination fees	200404	12.33	58	identity theft	201103	7.39
9	backed securities	200404	11.49	59	customer deposits	201104	11.52
10	rate environment	200404	7.05	60	results operations	201201	26.75
11	off balance	200502	18.08	61	secondary mortgage	201201	9.57
12	rate environment	200502	17.18	62	actual anticipated	201202	24.45
13	human resources	200502	9.41	63	deposit insurance	201202	20.92
14	commitments extend	200503	22.73	64	foreclosure process	201202	7.75
15	return plan	200503	15.16	65	commercial real	201203	9.75
16	internal audit	200503	12.95	66	operational risk	201204	8.05
17	real estate	200503	11.79	67	stock price	201301	10.32
18	income taxes	200504	14.82	68	ability compete	201302	41.41
19	negatively impact	200504	13.63	69	trust preferred	201302	17.49
20	rate swap	200504	13.09	70	extend credit	201302	10.05
21	investment securities	200504	12.25	71	weather events	201303	33.86
22	recruiting hiring	200601	12.94	72	executive compensation	201303	9.98
23	accounting policies	200601	11.42	73	material adverse	201304	13.53
24	return plan	200601	11.11	74	supervision regulation	201304	9.57
25	board directors	200602	26.28	75	regulatory requirements	201304	7.69
26	interest bearing	200602	24.14	76	basis point	201401	17.58
27	independent auditors	200602	21.98	77	basel iii	201401	7.73
28	fasb interpretation	200602	10.18	78	negative publicity	201402	17.49
29	accounting policies	200603	8.19	79	supervision regulation	201402	9.95
30	underwriting standards	200603	7.77	80	risk exposure	201402	7.63
31	business strategy	200603	7.39	81	investment securities	201403	14.40
32	time deposits	200604	29.50	82	capital levels	201403	11.58
33	brokered deposits	200604	15.16	83	regulatory authorities	201403	9.82
34	commitments extend	200604	13.97	84	brokered deposits	201404	11.58
35	investment securities	200604	13.40	85	subsidiary bank	201501	8.11
36	certain provisions	200701	26.60	86	senior management	201501	8.06
37	federal reserve	200701	12.49	87	carrying value	201502	7.19
38	senior notes	200701	10.01	88	income taxes	201502	7.08
39	shares common	200702	53.63	89	accounting principles	201503	8.35
40	audit committee	200702	14.94	90	business strategy	201504	23.21
41	board directors	200702	12.16	91	legal proceedings	201601	12.52
42	policies procedures	200703	17.48	92	servicing rights	201601	9.13
43	prevent fraud	200703	7.78	93	institution failures	201601	7.02
44	equity return	200703	7.40	94	risk profile	201602	18.27
45	damage reputation	200704	23.84	95	emerging growth	201603	14.19
46	accounting policies	200704	11.01	96	merger agreement	201603	8.83
47	extend credit	200704	8.02	97	credit risk	201603	8.77
48	cost funds	200801	9.20	98	trading volume	201604	36.15
49	rate risk	200802	23.99	99	data processing	201604	9.37
50	management policies	200803	14.93				

Table VI: Crisis and Current Period Return Regressions

Cross-sectional OLS regressions predicting individual bank outcomes during and after the financial crisis and under current economic conditions. For the crisis period on the left hand side of the table, the dependent variable is the bank's stock return from September 2008 to December 2012. For the current period on the right hand side, the dependent variable is the bank's stock return from December of 2015 to February 2016. We display results for raw returns (Panel A) and also for the negative part of the return (Panel B) (computed as the minimum of zero and the raw return minus the average return across banks). The independent variable of interest, *Emerging Risk Exposure*, is the quarterly predicted covariance based on Equation 3 using only the portion of the predicted value attributable to the 31 static themes. We note that all regressions use ex ante data and are predictive when noted as such in the *Predictive Timing* column. We include, but do not display in order to conserve space, controls for bank characteristics, momentum, log book to market and the log market capitalization in each regression. We also include industry fixed effects based on four-digit SIC codes. t statistics are reported in parentheses.

	Crisis Period					*Current Period*			
Row	Quarter	Emerging Risk Exposure	Obs	Predictive Timing		Quarter	Emerging Risk Exposure	Obs	Predictive Timing
		Panel A: Raw Returns					***Panel A: Raw Returns***		
(1)	2004 1Q	4.878 (1.07)	352	Predictive	—	2010 1Q	-0.819 (-1.08)	334	Predictive
(2)	2004 2Q	4.915 (1.38)	352	Predictive	—	2010 2Q	-0.646 (-2.36)	334	Predictive
(3)	2004 3Q	1.338 (0.24)	368	Predictive	—	2010 3Q	-1.085 (-4.62)	341	Predictive
(4)	2004 4Q	0.712 (0.18)	368	Predictive	—	2010 4Q	-0.527 (-1.53)	341	Predictive
(5)	2005 1Q	-0.004 (-0.00)	388	Predictive	—	2011 1Q	-1.179 (-3.14)	351	Predictive
(6)	2005 2Q	0.437 (0.11)	388	Predictive	—	2011 2Q	-1.018 (-2.38)	350	Predictive
(7)	2005 3Q	-0.973 (-0.19)	418	Predictive	—	2011 3Q	-1.346 (-5.09)	356	Predictive
(8)	2005 4Q	3.604 (0.55)	418	Predictive	—	2011 4Q	-1.099 (-3.91)	356	Predictive
(9)	2006 1Q	-1.750 (-0.40)	407	Predictive	—	2012 1Q	-1.357 (-2.86)	349	Predictive
(10)	2006 2Q	-4.924 (-1.36)	407	Predictive	—	2012 2Q	-1.287 (-1.24)	349	Predictive
(11)	2006 3Q	-4.694 (-1.13)	430	Predictive	—	2012 3Q	-0.848 (-1.74)	360	Predictive
(12)	2006 4Q	-6.614 (-1.25)	430	Predictive	—	2012 4Q	-1.131 (-1.13)	360	Predictive
(13)	2007 1Q	-3.509 (-1.59)	444	Predictive	—	2013 1Q	-0.196 (-1.68)	351	Predictive
(14)	2007 2Q	-3.713 (-1.68)	444	Predictive	—	2013 2Q	-0.815 (-1.24)	351	Predictive
(15)	2007 3Q	-4.010 (-3.03)	469	Predictive	—	2013 3Q	-0.262 (-0.34)	368	Predictive
(16)	2007 4Q	-2.285 (-3.03)	469	Predictive	—	2013 4Q	-1.052 (-1.87)	368	Predictive
(17)	2008 1Q	-3.731 (-2.59)	468	Predictive	—	2014 1Q	0.047 (0.11)	356	Predictive
(18)	2008 2Q	-4.065 (-5.23)	468	Predictive	—	2014 2Q	-0.440 (-2.00)	356	Predictive
(19)	2008 3Q	-3.071 (-1.76)	489	Non Predictive	—	2014 3Q	-1.287 (-2.33)	367	Predictive
(20)	2008 4Q	-0.829 (-0.41)	491	Non Predictive	—	2014 4Q	-1.096 (-1.99)	367	Predictive
(21)	2009 1Q	-1.875 (-0.73)	518	Non Predictive	—	2015 1Q	-0.804 (-1.95)	358	Predictive
(22)	2009 2Q	-1.779 (-0.74)	518	Non Predictive	—	2015 2Q	-1.037 (-2.79)	358	Predictive
(23)	2009 3Q	-3.064 (-1.01)	529	Non Predictive	—	2015 3Q	-1.793 (-5.42)	387	Predictive
(24)	2009 4Q	-0.973 (-0.29)	522	Non Predictive	—	2015 4Q	-0.945 (-4.82)	386	Non Predictive
		Panel B: Below Mean Returns					***Panel B: Below Mean Returns***		
(1)	2004 1Q	2.410 (2.16)	352	Predictive	—	2010 1Q	-0.928 (-3.25)	334	Predictive
(2)	2004 2Q	2.489 (3.69)	352	Predictive	—	2010 2Q	-0.657 (-3.27)	334	Predictive
(3)	2004 3Q	0.319 (0.18)	368	Predictive	—	2010 3Q	-0.738 (-4.44)	341	Predictive
(4)	2004 4Q	0.415 (0.28)	368	Predictive	—	2010 4Q	-0.282 (-1.53)	341	Predictive
(5)	2005 1Q	-0.670 (-0.31)	388	Predictive	—	2011 1Q	-0.746 (-3.33)	351	Predictive
(6)	2005 2Q	-0.519 (-0.28)	388	Predictive	—	2011 2Q	-0.758 (-4.22)	350	Predictive
(7)	2005 3Q	-1.006 (-0.36)	418	Predictive	—	2011 3Q	-0.941 (-11.7)	356	Predictive
(8)	2005 4Q	1.147 (0.40)	418	Predictive	—	2011 4Q	-0.671 (-4.30)	356	Predictive
(9)	2006 1Q	0.918 (0.65)	407	Predictive	—	2012 1Q	-0.778 (-2.40)	349	Predictive
(10)	2006 2Q	-2.462 (-1.44)	407	Predictive	—	2012 2Q	-0.660 (-1.40)	349	Predictive
(11)	2006 3Q	-2.656 (-1.06)	430	Predictive	—	2012 3Q	-0.916 (-3.73)	360	Predictive
(12)	2006 4Q	-3.374 (-1.09)	430	Predictive	—	2012 4Q	-0.798 (-1.77)	360	Predictive
(13)	2007 1Q	-4.268 (-2.01)	444	Predictive	—	2013 1Q	-0.121 (-1.45)	351	Predictive
(14)	2007 2Q	-3.436 (-2.01)	444	Predictive	—	2013 2Q	-0.228 (-1.92)	351	Predictive
(15)	2007 3Q	-3.908 (-3.04)	469	Predictive	—	2013 3Q	0.198 (0.95)	368	Predictive
(16)	2007 4Q	-3.406 (-3.27)	469	Predictive	—	2013 4Q	-0.375 (-2.54)	368	Predictive
(17)	2008 1Q	-3.970 (-3.65)	468	Predictive	—	2014 1Q	-0.024 (-0.17)	356	Predictive
(18)	2008 2Q	-4.943 (-7.80)	468	Predictive	—	2014 2Q	-0.222 (-3.00)	356	Predictive
(19)	2008 3Q	-3.113 (-2.21)	489	Non Predictive	—	2014 3Q	-0.832 (-2.42)	367	Predictive
(20)	2008 4Q	-1.778 (-1.02)	491	Non Predictive	—	2014 4Q	-0.681 (-2.30)	367	Predictive
(21)	2009 1Q	-1.823 (-1.15)	518	Non Predictive	—	2015 1Q	-0.440 (-1.53)	358	Predictive
(22)	2009 2Q	-2.471 (-1.55)	518	Non Predictive	—	2015 2Q	-0.505 (-1.47)	358	Predictive
(23)	2009 3Q	-2.942 (-9.97)	529	Non Predictive	—	2015 3Q	-1.015 (-2.33)	387	Predictive
(24)	2009 4Q	-2.107 (-2.88)	522	Non Predictive	—	2015 4Q	-0.500 (-1.49)	386	Non Predictive

Table VII: Bank Failure Regressions

Cross-sectional regressions predicting which banks fail during the period after the Lehman bankruptcy from November 2008 to June 2012 as indicated on the FDIC website. The dependent variable is a dummy variable equal to one if a bank in our sample was assisted or failed during the crisis period, and zero otherwise. There are 41 such failures, with {2,12,19,6,2} occurring in the years {2008,2009,2010,2011,2012}, respectively. The independent variable of interest, *Emerging Risk Exposure*, is the quarterly predicted covariance based on Equation 3 using only the portion of the predicted value attributable to the 31 static themes. We note that all regressions use ex ante data and are predictive when noted as such in the *Predictive Timing* column. We include as independent variables bank characteristics such as *Ln(Assets) Loans/Assets*, *Loan Loss Prov & Allow*, the sum of loan loss provision and allowances, *Capital*, the ratio of equity to assets, *Neg. Earnings Dummy* an indicator variable equal to one if net income is negative, zero otherwise, *Non-Performing Assets*, the sum of loans that are 30 days and 90 days past due, and *CatFat/Assets* from Berger and Bouwman (2009). We include industry fixed effects based on four-digit SIC codes. t-statistics are reported in parentheses.

Row	Quarter	Emerging Risk Exposure	Log Assets	Loans Assets	Loss/ Assets	Cap- ital	Neg Earn.	CatFat Assets	NPA Assets	Obs	Predictive Timing
(1)	2004 1Q	0.004 (0.80)	-0.006 (-0.93)	0.155 (4.52)	27.44 (0.12)	-0.241 (-0.92)	0.028 (0.56)	-0.022 (-2.39)	2.186 (11.5)	625	Predictive
(2)	2004 2Q	0.004 (0.94)	-0.006 (-0.91)	0.156 (4.52)	47.04 (0.21)	-0.245 (-0.92)	0.029 (0.57)	-0.021 (-2.47)	2.211 (10.2)	625	Predictive
(3)	2004 3Q	-0.005 (-1.03)	-0.006 (-0.84)	0.156 (4.39)	-55.18 (-0.19)	-0.264 (-0.97)	0.029 (0.58)	-0.020 (-2.49)	2.155 (11.2)	625	Predictive
(4)	2004 4Q	-0.004 (-0.79)	-0.005 (-0.79)	0.154 (4.63)	-64.16 (-0.20)	-0.251 (-0.92)	0.028 (0.56)	-0.020 (-2.71)	2.167 (11.5)	625	Predictive
(5)	2005 1Q	-0.002 (-1.33)	-0.004 (-0.55)	0.172 (5.33)	129.94 (0.30)	-0.429 (-2.24)	0.008 (0.14)	-0.040 (-3.79)	-0.968 (-4.42)	615	Predictive
(6)	2005 2Q	-0.001 (-1.36)	-0.003 (-0.56)	0.172 (5.30)	155.81 (0.33)	-0.429 (-2.23)	0.008 (0.14)	-0.041 (-3.85)	-0.980 (-4.46)	615	Predictive
(7)	2005 3Q	0.008 (3.56)	-0.001 (-0.22)	0.172 (4.95)	493.93 (1.11)	-0.444 (-2.43)	0.009 (0.16)	-0.042 (-3.91)	-0.865 (-3.85)	615	Predictive
(8)	2005 4Q	0.006 (2.55)	-0.002 (-0.30)	0.173 (5.04)	454.13 (0.96)	-0.433 (-2.33)	0.009 (0.17)	-0.042 (-4.06)	-0.889 (-4.22)	615	Predictive
(9)	2006 1Q	-0.002 (-0.14)	-0.009 (-0.95)	0.217 (6.38)	-166.49 (-1.54)	-0.401 (-2.22)	-0.060 (-3.66)	0.062 (2.66)	-0.899 (-2.91)	578	Predictive
(10)	2006 2Q	-0.001 (-0.08)	-0.009 (-1.10)	0.217 (6.51)	-142.80 (-1.56)	-0.405 (-2.53)	-0.060 (-4.53)	0.062 (2.69)	-0.900 (-3.18)	578	Predictive
(11)	2006 3Q	0.003 (0.58)	-0.009 (-1.22)	0.218 (6.49)	-95.03 (-1.02)	-0.416 (-2.59)	-0.059 (-5.15)	0.061 (2.49)	-0.882 (-2.85)	578	Predictive
(12)	2006 4Q	0.008 (3.97)	-0.009 (-1.30)	0.219 (6.61)	-20.84 (-0.16)	-0.431 (-2.90)	-0.060 (-5.93)	0.059 (2.48)	-0.869 (-2.68)	578	Predictive
(13)	2007 1Q	0.009 (3.96)	-0.009 (-0.80)	0.255 (8.12)	-51.59 (-1.25)	-0.630 (-5.61)	-0.000 (-0.00)	0.021 (0.55)	-1.121 (-1.63)	588	Predictive
(14)	2007 2Q	0.011 (7.36)	-0.008 (-0.72)	0.253 (8.30)	-42.25 (-1.04)	-0.630 (-5.62)	-0.002 (-0.03)	0.022 (0.60)	-1.067 (-1.59)	588	Predictive
(15)	2007 3Q	0.010 (2.31)	-0.008 (-0.68)	0.254 (8.24)	-49.61 (-1.26)	-0.630 (-5.58)	0.000 (0.00)	0.020 (0.53)	-1.073 (-1.67)	588	Predictive
(16)	2007 4Q	0.014 (4.37)	-0.008 (-0.69)	0.253 (7.88)	-54.13 (-1.23)	-0.629 (-5.10)	-0.001 (-0.02)	0.020 (0.53)	-1.071 (-1.66)	588	Predictive
(17)	2008 1Q	0.014 (4.42)	-0.008 (-0.89)	0.247 (3.66)	-380.88 (-1.87)	-0.500 (-3.58)	0.042 (0.62)	-0.006 (-0.25)	2.982 (3.18)	562	Predictive
(18)	2008 2Q	0.015 (3.89)	-0.007 (-0.85)	0.249 (3.63)	-384.81 (-1.84)	-0.502 (-3.57)	0.044 (0.64)	-0.007 (-0.25)	2.987 (3.13)	562	Predictive
(19)	2008 3Q	0.015 (3.72)	-0.007 (-0.93)	0.246 (3.55)	-365.39 (-1.88)	-0.504 (-3.60)	0.043 (0.63)	-0.009 (-0.36)	2.994 (3.21)	562	Predictive
(20)	2008 4Q	0.004 (0.63)	-0.007 (-0.87)	0.250 (3.63)	-366.38 (-1.90)	-0.473 (-3.18)	0.042 (0.62)	-0.004 (-0.14)	2.798 (3.22)	562	Non Predictive
(21)	2009 1Q	0.024 (8.54)	-0.017 (-2.44)	0.127 (4.71)	-347.53 (-18.0)	-0.944 (-3.53)	0.094 (5.37)	-0.092 (-5.20)	4.268 (47.8)	564	Non Predictive
(22)	2009 2Q	0.010 (3.87)	-0.013 (-2.22)	0.136 (4.32)	-334.82 (-15.2)	-0.941 (-3.52)	0.097 (5.17)	-0.082 (-4.00)	4.321 (51.7)	564	Non Predictive
(23)	2009 3Q	-0.001 (-0.27)	-0.011 (-1.99)	0.138 (4.67)	-340.98 (-14.9)	-0.923 (-3.78)	0.098 (5.14)	-0.079 (-3.93)	4.173 (38.9)	564	Non Predictive
(24)	2009 4Q	0.007 (1.96)	-0.012 (-1.90)	0.142 (4.32)	-342.28 (-15.4)	-0.943 (-3.49)	0.098 (5.12)	-0.080 (-3.73)	4.244 (35.5)	564	Non Predictive

Table VIII: Fama-MacBeth Rolling Predictive Volatility Regressions

Fama-McBeth rolling three month cross-sectional regressions where the dependent variable is the bank's monthly volatility of daily stock returns from January 1998 to December 2016 (data from 1997 is needed to compute starting values). The independent variable of interest, *Emerging Risk Exposure*, is the quarterly predicted covariance based on Equation 3 using only the portion of the predicted value attributable to the 31 static themes. This variable is measured over the number of quarters specified in the column heading. The number of observations is based on the 1 Quarter Emerging Risk Exposure regression. We include, but do not display in order to conserve space, controls for bank characteristics, momentum (month t-12 to t-2), log book-to-market ratio, the log market capitalization and a dummy variable for negative book-to-market ratio in each regression. We also include industry fixed effects based on four-digit SIC codes. *t*-statistics are reported in parentheses.

Monthly Lag	1 Quarter Emerging Risk Exposure	2 Quarter Emerging Risk Exposure	3 Quarter Emerging Risk Exposure	4 Quarter Emerging Risk Exposure	Obs
1	0.086 (8.94)	0.105 (10.26)	0.112 (11.35)	0.113 (11.68)	52641
2	0.084 (8.72)	0.104 (10.22)	0.108 (11.13)	0.109 (11.51)	52476
3	0.086 (9.18)	0.099 (10.53)	0.104 (11.38)	0.101 (11.72)	52312
4	0.086 (9.13)	0.098 (10.81)	0.102 (11.43)	0.097 (11.29)	52148
5	0.085 (9.13)	0.093 (10.42)	0.097 (11.32)	0.092 (11.00)	51786
6	0.079 (8.96)	0.088 (10.40)	0.088 (11.09)	0.087 (10.39)	51410
7	0.076 (9.52)	0.083 (10.66)	0.081 (10.52)	0.080 (10.13)	51035
8	0.069 (8.66)	0.077 (10.04)	0.074 (9.60)	0.075 (9.25)	50660
9	0.064 (8.59)	0.069 (9.39)	0.071 (9.09)	0.072 (9.04)	50284
10	0.062 (8.65)	0.064 (8.62)	0.066 (8.82)	0.067 (8.60)	49908
11	0.058 (8.38)	0.060 (8.28)	0.063 (8.51)	0.063 (8.41)	49569
12	0.053 (7.51)	0.057 (7.74)	0.060 (8.06)	0.059 (8.09)	49230
13	0.045 (6.84)	0.049 (7.40)	0.054 (7.43)	0.053 (7.41)	48891
14	0.041 (6.29)	0.046 (6.79)	0.051 (6.95)	0.051 (6.81)	48541
15	0.037 (5.81)	0.044 (6.49)	0.047 (6.56)	0.047 (6.43)	48191
16	0.032 (5.09)	0.040 (5.54)	0.043 (5.83)	0.046 (5.99)	47841
17	0.031 (4.63)	0.040 (5.40)	0.042 (5.61)	0.044 (5.53)	47490
18	0.032 (4.73)	0.039 (5.25)	0.042 (5.60)	0.043 (5.40)	47139
19	0.030 (4.02)	0.036 (4.73)	0.041 (5.27)	0.042 (5.03)	46788
20	0.033 (4.62)	0.036 (5.00)	0.041 (5.30)	0.041 (5.16)	46438
21	0.029 (4.26)	0.035 (4.99)	0.039 (5.12)	0.039 (4.96)	46088
22	0.028 (4.16)	0.036 (5.24)	0.039 (5.25)	0.038 (4.99)	45738
23	0.024 (3.80)	0.034 (4.68)	0.036 (4.86)	0.037 (4.81)	45404
24	0.028 (4.23)	0.034 (4.59)	0.035 (4.72)	0.037 (4.69)	45071
25	0.030 (4.24)	0.035 (4.34)	0.035 (4.50)	0.036 (4.45)	44738
26	0.028 (3.60)	0.031 (3.80)	0.033 (4.14)	0.034 (4.18)	44397
27	0.027 (3.43)	0.029 (3.65)	0.033 (4.10)	0.032 (3.94)	44056
28	0.027 (3.36)	0.030 (3.85)	0.033 (4.20)	0.033 (4.12)	43716
29	0.025 (3.17)	0.030 (3.95)	0.034 (4.46)	0.032 (4.00)	43376
30	0.021 (2.65)	0.027 (3.53)	0.029 (3.78)	0.028 (3.40)	43035
31	0.019 (2.61)	0.024 (3.19)	0.026 (3.46)	0.024 (3.01)	42694
32	0.022 (3.08)	0.026 (3.54)	0.025 (3.32)	0.024 (2.93)	42354
33	0.023 (3.23)	0.024 (3.15)	0.023 (2.99)	0.023 (2.69)	42014
34	0.022 (3.12)	0.023 (3.07)	0.021 (2.81)	0.022 (2.67)	41675
35	0.024 (3.32)	0.022 (2.93)	0.022 (2.80)	0.022 (2.75)	41355
36	0.019 (2.49)	0.018 (2.40)	0.019 (2.40)	0.020 (2.48)	41035

Table IX: Bank Disclosure Changes

The table reports significant changes in the disclosure of risk factors during the sample period. The methodology has the same starting point as the dynamic SVA model with bigrams extracted from an LDA model. However, unlike either the dynamic or the static model, the results of this table are purely a function of bank 10-K disclosures and thus do not depend on covariance. In particular, for all bigrams that are used in the SVA model (between 100 to 200 in each year), we use SVA vectors for each topic and and score each firm based on how much of each bigram-topic the firm discloses. We then normalize all exposures for each bank such that they sum to one, and hence we can explore relative changes in disclosure rather than nominal changes. In each year, we then average the exposures to each topic across all banks, thus obtaining a single vector for each of the topics in each year. A given topic is deemed to be emerging in a given year if the average exposure in the given year is significantly higher than the exposure was in the past 5 years. To make the list fit on one page, we sort all z-scores across all years from high to low and take the 40 highest z-scores. The cutoff for inclusion is a given topic must have a z-score roughly exceeding 8.0 in the given year. We exclude the year 2005 from this test due to the change in disclosure rules associated with risk factors in that year.

	Year	Emerging Risk	Z score
1	2003	federal funds	12.9
2	2004	operational risk	9.9
3	2004	management process	9.0
4	2004	currently offered	8.7
5	2006	real estate	9.4
6	2007	borrowers repay	13.0
7	2007	loan review	11.0
8	2007	mortgage backed	9.7
9	2007	student loan	8.8
10	2008	institutional counterparty	49.8
11	2008	guarantee program	40.6
12	2008	borrow funds	18.8
13	2008	brokered deposits	10.5
14	2008	preferred shares	10.4
15	2008	credit enter	9.0
16	2008	options granted	8.7
17	2009	certificates deposit	8.6
18	2010	debit card	9.9
19	2011	freddie mac	9.6
20	2012	cyber attacks	10.1
21	2012	data processing	10.1
22	2012	war terrorism	9.8
23	2012	representations warranties	9.5
24	2013	business strategy	13.2
25	2013	products services	11.7
26	2013	internal control	10.7
27	2013	information systems	9.7
28	2013	disbursement partners	9.1
29	2013	judgments settlements	8.9
30	2013	data processing	8.3
31	2013	representations warranties	8.2
32	2014	information security	11.9
33	2014	dodd frank	10.9
34	2014	information technology	8.8
35	2014	financial services	8.7
36	2014	confidential information	8.6
37	2014	ability retain	8.5
38	2014	products services	8.5
39	2014	results operations	8.1
40	2015	dodd frank	13.4