EAGER: III: CIFRAM: Dynamic Identification and Interpretation of Emerging Systemic Risks Using Textual Analysis

Co-PIs: Kathleen Weiss Hanley, Lehigh University and Gerard Hoberg, University of Southern California

Award Number: 1449578

Awarded Amount: $299,683

Abstract at Time of Award

This project will employ linguistic tools to determine how and why financial crises, such as the 2008 crisis following the Lehman Brothers bankruptcy, form and grow in magnitude. By examining textual information gleaned by processing large volumes of verbal data from 10-K filings to the Securities and Exchange Commission, the principal investigators will use techniques developed by computer scientists to assess verbal themes and link them to market data to assess whether future crises are forming. The techniques employed enable the identification of actual risks allowing regulators and market participants the ability to respond appropriately in advance of a major development. The free provision of data and computer code on the internet will lower the cost for future researchers to also examine these issues. The principal investigators will also present the research at conferences, submit the work for publication, and will work with and train graduate students. The material will be taught to future business leaders in the classroom, where MBA students and undergraduate students can openly discuss the results and their implications. The work will also be submitted to conferences attended by regulators to share insights on how they can be used to manage potential crises before they can cause extensive damage.

The principal investigators will use methods from computational linguistics, including Latent Dirichlet Allocation (LDA) and document similarity analysis, to identify a set of verbal topics that are common among financial firms, non-financial firms with exposure to the finance industry, and then all firms in the economy. The investigators will then use clustering and network methods to assess and categorize the business links among firms in the economy and to examine how they evolve over time. The resulting firm-relatedness network will then be compared to market data during various time intervals to understand how and why stock prices comove differently in neighboring periods, especially periods leading up to major crises. The verbal factors will be interpretable, and hence this technique will provide a fully automated description of why firms comove in different ways in different time periods. This method will be replicable and not subjected to researcher prejudice, allowing the data to inform researchers regarding the most salient issues affecting markets, even if the researcher is ex-ante unfamiliar with the true drivers of a specific systemic risk event. Once the textual drivers of comovement are understood, these factors can be used to back-test how the dynamic topic structure evolves during other systemic events. If successful, this research could create an early warning system for potential future crises and serve as a risk management tool by addressing the drivers of crisis before they occur, thereby reducing the cost of resolution.

For further information see the project web site:

http://www-bcf.usc.edu/~hoberg/HanleyHobergDataSite/index.html

Project Outcomes Report

Disclaimer: This Project Outcomes Report for the General Public is displayed verbatim as submitted by the Principal Investigator (PI) for this award. Any opinions, findings, and conclusions or recommendations expressed in this Report are those of the PI and do not necessarily reflect the views of the National Science Foundation; NSF has not approved or endorsed its content.

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 investors, we use daily stock returns and estimate the bank-pair covariance matrix. For banks, we use the disclosure of material risks in their annual 10-K. 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) 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. 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 relevant to many banks. Second, we use word-to-vec, 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.

Our first main contribution is to construct an aggregate index of emerging risks in each quarter. The level of the index becomes highly elevated in 2005Q2, 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.

Our second major contribution is the implementation of a model that uses natural language processing to identify specific interpretable risks that explain elevations in our aggregate risk index. The first model, the ``static'' model, considers 31 semantic risk themes identified from the manual evaluation of LDA output and correspond to the types of financial sector risk established in the academic literature. The static model is informative because many determinants of financial instability are similar from one event to another. We find that themes related to real estate, prepayment risk, commercial paper, dividends, operational risk, and credit cards 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. For example, a researcher may be interested in more granular sub-themes of the real estate topic such as subprime, mortgage-backed, HELOC, and foreclosure. These sub-themes can be included in the static model to assess their relative importance. When we run the extended static model on the real estate sub-themes, we find that these specific risks became elevated before the crisis period.

We recognize, however, that the financial sector is both complex and constantly changing, posing a challenge for those who monitor risk. We, therefore, propose a second ``dynamic model'' that automates the identification of candidate risks in each year. The benefit of automation is that it can detect emerging risks even if the researcher is entirely unaware of its potential relevance.

Our dynamic model reassuringly finds many of the same emerging risks identified by our static model. However, the model also reveals new risks 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 types of 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 impact of individual banks' exposure to emerging risks in each period. We find that banks with higher ex ante exposures to static risks experience three ex post negative outcomes: more negative stock returns during the financial crisis, higher bank failure rates, and higher stock price volatility lasting up to 36 months.

Our methodology can also be used in real time. Examining 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 or the build-up of risk within banks that could contribute to a systemic event (as occurred in the recent financial crisis).