Informativeness and Timeliness of Text Similarity Measures for Predicting Banks' Tail Comovement

Robert M. Bushman
Kenan-Flagler Business School
University of North Carolina-Chapel Hill

Jason V. Chen
University of Illinois at Chicago

Christopher D. Williams
Ross School of Business
University of Michigan

This Draft: May 2017

We appreciate the helpful comments of Donny Zhang and seminar participants at Chicago Booth, USC, Seoul National University and the University of Iowa. Bushman thanks Kenan-Flagler Business School, University of North Carolina at Chapel Hill, Chen thanks the University of Illinois for their support, and Williams thanks the Stephen M. Ross School of Business and Michael & Joan Sakkinen Faculty Fellowship for financial support.

Informativeness and Timeliness of Text Similarity Measures for Predicting Banks' Tail Comovement

Abstract

In this paper we examine the informativeness and timeliness of a bank's 10-K discussions for predicting its future downside tail risk comovement with other banks. We measure a bank's connectedness by constructing a measure of its text similarity with other banks based on 10-K business description and MD&A discussions. Focusing first on average similarity of a bank's textual disclosures with those of all other banks, we find that comovement between the lower tail of a given bank's future equity return distribution and the lower tail of the banking system's returns is increasing in the bank's average similarity. We also construct groups of connected peer banks with the most text similarity, finding that banks co-move significantly more in the tails with its highest similarity peers than with lower similarity banks. We then disaggregate similarity into business description and MD&A components finding that, while both predict future tail comovement, the economic significance of business description similarity is much higher than MD&A similarity. Further, footnote text similarity has no incremental explanatory power relative to business description and MD&A text similarity. Finally, we separate 10-K text into boilerplate and non-boilerplate components. We find that both boilerplate and non-boilerplate similarity have incremental information about future tail comovement. However, non-boilerplate similarity is significantly timelier than boilerplate, consistent with non-boilerplate similarity capturing commonalities across banks in currently evolving fundamentals and boilerplate similarity capturing commonalities in structural features that evolve slowly over time.

Introduction

The Financial Crisis of 2007–2009 focused significant attention on assessing and managing the downside tail-risk of banks. In addition to focusing on the standalone risks of individual banks, increasing attention has been focused on the complex web of direct and indirect interconnections between banks through which illiquidity, insolvency, and losses can spread during periods of financial distress. Strong interconnectivity can result in banks sharing similar vulnerabilities that expose them to comovement of extreme downside outcomes. In this paper, we investigate the extent to which similarity in verbal disclosures in 10-K reports across banks provides valuable information about their interconnectedness. Specifically, we examine both the informativeness and timeliness of a bank's 10-K text similarity with other banks for predicting the bank's future tail-risk comovement with these other banks.

Our objective in pursuing this line of inquiry is twofold. First, we seek to provide evidence to bank outsiders (bank regulators, investors, researchers, etc.) about the value of incorporating text-based financial analysis into assessments of bank connectedness and related systemic risk exposures. The challenges involved in constructing useful measures of the susceptibility of banks to systemic risk exposure has motivated a vibrant, growing literature which raises the possibility that multiple risk measures may be needed to capture the complex and adaptive nature of the financial system (e.g., Hansen, 2014; Bisias, 2012). A common approach to measuring bank connectedness and systemic risk exposure relies exclusively on quantitative information such as return series of traded securities and balance sheet data (e.g., Billio et al., 2012; Adrian and Brunnermeier, 2016; Cai et al., 2016; Acharya et al., 2017). However, a recent literature demonstrates that verbal discussions in mandatory financial reports comprise a rich source of valuable information that can be extracted using natural language processing techniques (e.g.,

Hoberg and Philips, 2016; Loughran and McDonald, 2016). We extend the systemic risk measurement literature by exploring measures of bank connectedness based on text similarity across banks'10-K business descriptions and MD&A discussions.

Second, in addition to extending the risk measurement literature we seek to provide new insights into the usefulness of 10-K textual disclosure by applying these disclosures in a novel decision context. Specifically, we investigate the informativeness of different sections within the 10-K (i.e., business description, MD&A, risk factors and footnotes) and the nature of the discussion (boilerplate vs non-boilerplate) for forecasting future tail comovement. While prior literature has examined many characteristics of text based measures, extending the analysis to new contexts (e.g., tail risk comovement in banks) is important to achieving a more complete understanding of the informativeness of such disclosures. As noted by Gjesdal (1981) and Dechow et al., (2010), the usefulness of an information system can vary across decision contexts. Our exploration of differences in information properties across distinct aspects of 10-K text discussions, for example boilerplate and non-boilerplate, contributes to the debate about the consequences of increasingly onerous accounting disclosures (Dyer et al., 2017; Li, 2008; SEC, 2013). In this regard, a notable feature of our paper is our use of 10-K text similarity across banks to isolate network clusters of banks that share similar vulnerabilities to downside tail risk. This contrasts with a growing literature using 10-K discussions to extract incremental information about the prospects of individual firms.[1]

Our approach to measuring bank connectedness is related in some respects to the text-based approach used by Hoberg and Philips (2016) to organize publicly traded firms into

[1] This includes findings that there is incremental information content in the tone of 10-K text (Feldman, Givindaraj, Livnat, and Segal, 2010; Loughran and McDonald, 2011), in its readability (Li, 2008; Loughran and McDonald 2014), and in year-on-year changes to the MD&A section (Brown and Tucker, 2011; Cohen, Malloy, and Nguyen, 2015).

industry groupings. Applying textual analysis to the business description section of 10-K reports, Hoberg and Philips (2016) compute measures of pairwise product similarity that reflect the extent to which firms are related to each other in terms of their product offerings. Unlike Hoberg and Philips, our objective is to construct measures of bank connectedness that capture the extent to which a group of banks share similar vulnerabilities to downside tail risk. Such risk vulnerabilities can be driven by sources of connectedness beyond product market competition, including key aspects of the bank's current and forecasted situation that spans performance, business models, credit risk, investment concentrations, funding sources, and liquidity exposures, among other issues. In light of this, we expand our 10-K text analysis to consider the business description section (similar to Hoberg and Philips), as well as the Management Discussion and Analysis disclosure (MD&A).

To measure bank connectedness using the textual discussions in the 10-K, we construct time-varying measures of cosine similarity between a bank's business description and MD&A discussions and those of all other publicly traded banks. We then use the matrix of pairwise similarities to design measures of connectedness between banks and investigate the extent to which these connectedness measures predict future tail comovement among connected banks.

We measure tail risk comovement using two measures designed to capture both the magnitude and the frequency of tail risk comovement. To capture the magnitude of comovement, our first measure is constructed as the average of the abnormal returns of a given bank over the 20 days where the portfolio of banks in a chosen index group has its lowest abnormal performance for the year. To capture the frequency of tail risk, our second measure is the number of days in which a bank experiences one its 20 lowest abnormal return days at the same time that the portfolio of banks in a chosen index group is also experiencing one its 20 lowest return days.

The portfolio of banks used to construct our index groups is either the entire banking sector or portfolios consisting of banks designated as a bank's peer or non-peer banks based on our text similarity measure of connectedness.

We begin our analyses by first examining the relation between a bank's average cosine similarity with all other publicly traded banks and its tail co-movement with an index portfolio consisting of all other banks. We find that comovement between the lower tail of a given bank's future equity return distribution and the lower tail of the banking system's return distribution is increasing in the average cosine similarity of the bank. Our results hold after controlling for the current level of tail comovement and a quantitative measure of connectedness constructed by estimating pairwise cosine similarities between all banks based on the entire, standardized vector of quantitative accounting data required to be reported in banks' regulatory financial report filings.

While our previous analyses considered a bank's average text similarity and tail comovement with all other banks, it is likely that a bank's connectedness and tail comovement is not uniform across banks. Rather, a bank's tail comovement should be significantly higher with banks with which it is most similar than it is with less similar banks. To explore this possibility, we examine the extent to which the matrix of pairwise text similarities scores allows us to effectively cluster the banking sector into subsets of banks for which future tail movement is expected to be the highest. We find that a bank co-moves significantly more in the tails within its highest similarity cluster than with lower similarity banks.

Up to now, our analyses have computed text similarity using the combined text of the business description and MD&A sections of banks' 10-K reports. To examine the relative informativeness of different aspects of 10-K textual discussions, we compute a bank's average

cosine similarity score separately for the business description and MD&A sections. We find that both business description similarity and MD&A similarity have incremental information for predicting future tail comovement. However, the economic significance of business description similarity is higher than MD&A similarity. While not part of our original similarity measure, in further analyses, we find that footnote text similarity has no incremental information relative to business description and MD&A text similarity. Finally, more recently firms have been required to disclose risk factors in a separate section. For a subset of banks, we also show that the risk factor section similarity does have incremental information, however, the coefficient on the risk factor similarity is negative, suggesting that it is useful in rebalancing the weighting across different dimensions of fundamentals aggregated together within business description and MD&A similarity.

In addition to examining the usefulness of different sections of the 10-K for predicting tail risk we also investigate the information content of the type of language found in the 10-K. Investors, preparers, regulators, and standard setters have expressed concern that the use of boilerplate language in 10-K textual disclosures has been increasing over time, potentially reducing the usefulness of these disclosures (Li, 2008; SEC, 2013). In this regard, we separate our primary measure (i.e., the combined business and MD&A text) into boilerplate language and non-boilerplate language following the methodology from Dyer et al., (2017). We then compute a separate boilerplate similarity measure and non-boilerplate similarity measure.

Using these new measures of similarity, we first document that boilerplate *similarity* across banks has not significantly increased in recent years. Second, we find that both boilerplate and non-boilerplate similarity have incremental information about future tail comovement. Finally, when we include three lags in each of the similarity measures (boilerplate and non-boilerplate),

we find that the first lag in non-boilerplate similarity is most informative, while the third lag is most informative for boilerplate similarity. This suggests non-boilerplate similarity is timelier than boilerplate, this evidence is consistent with non-boilerplate similarity capturing commonalities across banks in currently evolving fundamentals and boilerplate similarity capturing commonalities in structural features that evolve slowly over time.

This paper makes several contributions to the literature. First, we contribute to the literature on systemic risk by demonstrating that similarity across banks' 10-K textual disclosures is informative about future tail comovement. Our text similarity technique for organizing banks into high comovement groups may be useful to bank outsiders (regulators, investors, etc.) for extending oversight or monitoring to clusters of similarly vulnerable publicly traded banks, in addition to a focus on individual banks. This technique may also be valuable to outside investors and researchers who must rely on public information in assessing the implications of bank vulnerability for future tail comovement. We complement Rönnqvist and Sarlin (2016), who estimate interconnections between large European banks based on co-occurrences of bank names in news articles, and papers measuring bank connections using quantitative data ((e.g., Billio et al., 2012; Huang et al., 2011; Adrian and Brunnermeier, 2016; Acharya et al., 2016; Cai et al., 2016). We also complement Hanley and Hoberg (2016) who use computational linguistics of bank's risk disclosures in the 10-K to develop an empirical model of dynamic, interpretable emerging risks that predicts the emergence of financial instability.

Second, where Hoberg and Philip use similarity in product market descriptions to form product market industry clusters, we extend the literature by using textual similarity across banks to form clusters of banks based on common vulnerabilities to downside tail outcomes. Our multi-

firm perspective on 10-K disclosures also extends a growing body of research that uses textual analysis to assess the implications of narrative financial disclosures at the individual firm level.[2]

Finally, we contribute to the literature that examines the usefulness of textual discussions found in the 10-K by applying these disclosures to a multi-firm setting in a novel decision context. Significant increases in the length and complexity of 10-K verbal disclosures have prompted concerns about the usefulness and informativeness of these disclosures (KPMG, 2011; SEC, 2013; Dyer et al. (2017). Because the value of information depends on context, our extension to predicting tail comovement across banks expands the frontier of knowledge about the information content of 10-K disclosures. Further, our analyses of the informativeness of text similarity across banks for different sections of the 10-K extends the literature that looks at this for individual firms (e.g., Amel-Zadeh and Faasse, 2016). Also, our results provide evidence that while boilerplate similarity is not as timely as non-boilerplate similarity, boilerplate similarity still does have information content for predicting comovement over and above the non-boilerplate language. These results add a new perspective on boilerplate language by suggesting that such language reflects structural aspects of a bank that change slowly over time. It is also noteworthy that while the use of boilerplate language has increased significantly over time (Dyer et al., 2017), we find that boilerplate similarity across banks does not exhibit an increasing trend and the informative of such similarity for predicting future tail comovement has not deteriorate through time.

Our remainder of our paper proceeds as follows, section 2 discusses the sample selection and measurement of connectedness. Section 3 discusses the empirical approaches we take and the results. Section 4 concludes.

[2] Recent reviews of the literature include Loughran and McDonald (2016) and Kearney and Liu (2014).

2. Sample, Measures of Connectedness and Comovement, and Descriptive Statistics

Our sample consists of all banks with two digit historic SIC codes between 60 and 62 which are available in Compustat Annual or Compustat Annual Bank. We download each financial institutions 10-K and 10-K405 filings from the SEC EDGAR online filling system. Our sample of 10-K filings begins in 1995 and ends in 2014. The sample begins in 1995 because this is the first year in which the SEC required all publically traded companies to make their filings publically available electronically through the EDGAR filing system. The Management Discussions and Analysis (hereafter MD&A) and Business (here after BUS) sections of each 10-K are extracted using PERL. Financial information was obtained from Compustat Annual and Compustat Annual Bank. Market returns and pricing information is obtained from Eventus and CRSP where needed. In the remainder of this section, we discuss in detail how we measure qualitative similarity (section 2.1) and tail risk comovement (section 2.2). We also provide descriptive statistics for these measures (section 2.3).

2.1 Text-based Measure of Connectivity

Our initial analyses focus on cosine similarity measures of connectivity between banks constructed using the verbal discussion contained in the 10-K business description section (BUS) and the MD&A disclosures of banks' annual 10-K reports. We will construct connectivity using the combined text of BUS and MD&A, as well as constructing measures for BUS and MD&A separately. Later in the paper we will also compute text similarity scores for the footnotes and risk factor disclosures from the 10-K report, as well as disaggregating the combined BUS and MD&A into boilerplate and non-boilerplate sentences and computing separate text similarity

scores for each group of sentences.[3] For any given subset of 10-K verbal disclosures, we compute pairwise word similarity scores for each pair of banks in a given year, and then use the matrix of pairwise similarity scores to calculate measures of bank connectivity in a given year.

The business description section typically appears as Item 1 or Item 1A in bank's 10-K. To estimate cosine similarities, we first extract BUS and MD&A from each financial institution's 10-K filing for each year. As is common in the literature, stop words are eliminated from the text. Using the text from the combined BUS and MD&A disclosures or from BUS and MD&A separately, we construct a vector summarizing each bank's usage of words. The number of elements in these vectors is equal to the number of unique words used by the bank. Each element of a vector represents the number of times that a unique word is mentioned by a bank in their discussion in a given year. For each year, we then estimate the pairwise cosine similarity between a given institution's word vector and the word vectors of all other banks in the sample.

Cosine similarity is a technique from the field of textual analysis which calculates the similarity between two sets of texts (Kogan et al 1998). The technique has had wide spread use in the areas of computer science and web development (Joydeep et al 2000). Recently studies in accounting and finance have used this technique to examine changes in firm's fundamentals and similarity in product market offerings (Brown and Tucker 2011, Hoberg and Phillips 2016). The cosine similarity between two banks is the cosine of the angle between the vectors of words that comprise the combined BUS and MD&A. Specifically, the cosine similarity between two vectors of words B_1 and B_2 is calculated as follows:

$$Cosine\ Similarity = \frac{B_1 \cdot B_2}{\|B_1\| \|B_2\|}.$$

[3] We will describe our technique for defining boilerplate sentences later in the paper.

where · indicates vector dot product, and ||*B*|| is the length of vector *B*. B_1 and B_2 are the vectors of words for two distinct banks being compared. The axes of each vector are the unique words in the text and the magnitude of the axis is the number of times that the given word is mentioned in the given text. These word vectors in essence assign each bank a unique spatial location based on its word usage, and its own potential set of nearby connected banks in this space based on word overlaps. The distance between banks is defined by a cosine similarity score which is higher when banks i and j use more of the same words with similar intensity, where a cosine similarity of 1 means that the two word vectors are identical. This process allows us reduce high-dimensional word vectors to a simple matrix of bank's pairwise similarity scores.

As discussed in more detail below, we use the matrix of pairwise cosine similarities in several different ways. For some analyses, we compute the average cosine similarity between a given bank and all other banks in the sample in a given year using combined BUS and MD&A text (*AvgCos_MDABUS*). In Table 1 we report that *AvgCos_MDABUS* has a mean value of 0.70 with a standard deviation of 0.09. We also compute cosine similarity separately using either MD&A alone or BUS alone, where Table 1 reports that the mean value of *AvgCos_MDA* (*AvgCos_BUS*) is 0.68 (0.57) with a standard deviation of 0.09 (0.11).

We also use pairwise cosine similarities to form each bank's high cosine group consisting of the banks with which a given bank is most similar in a given year. Our premise is that this high cosine group represents the set of banks with which an individual bank is most connected, where we expect the bank to commove in the tails more with these banks than with banks outside the high cosine group. Our procedure for forming high cosine similarity groups of connected banks is similar to that used in Hoberg and Philips (2016) to place firms into industries based on the similarity of verbal product descriptions.

2.2 Tail Risk Comovement

Following the recent financial crisis there has been considerable interest in modeling and measuring systemic risk, the risk that many banks will simultaneously experience financial distress and impose externalities on the overall economy. There is no agreed upon approach to this measurement (e.g., Bisias et al., 2012, Hansen, 2014). One important stream of literature exploits the high frequency observability of bank's equity prices to extract measures of systemic risk, focusing on comovement in the tails of equity returns across banks (Acharya et al., 2017, Adrian and Brunnermeier, 2016). We measure tail risk comovement by constructing measures inspired by the marginal expected shortfall measure (*MES*) developed in Acharya et al. (2017). *MES* is designed to measure an individual bank's tail risk exposure to system-wide distress, and is analogous to the stress tests performed by individual institutions and regulators. It has been shown to have significant explanatory power for which firms contribute to a potential crisis (Acharya et al., 2017). The *MES* measure reflects the connection between a bank's equity returns and market equity returns on days where the market return is in the bottom 5% for the year. That is, it measures the extent to which an individual bank's returns are low when the overall (banking) market returns are low. Building on this idea, we create two measures of tail comovement that aim to capture the ideas of frequency and magnitude.

To capture frequency, our first measure, *LFM Days*, reflects the number of days in year t where bank i and a portfolio of banks included in a specified index group simultaneously experience low returns performance. An extreme low performance day occurs if it is in the set of the lowest 20 return days for year t based upon daily abnormal returns. A bank's daily abnormal return is calculated using Eventus, and is the difference between the bank's return and a value weighted market return. Eventus calculates the value weighted market return using NYSE,

AMEX, and Nasdaq stocks. We calculate a daily market return by summing the abnormal returns each day for all the banks in a specific portfolio of banks selected to represent the comparison index group, and then find the lowest 20 market performance days in a calendar year. Next, for each bank we calculate daily abnormal returns and then find their bottom 20 performance days in a given year. *LFM Days* is the number days in a given calendar year in which the bank and the selected bank index group both have low performance. This measure can vary from 0 (no overlap of low days for bank i and the index) and 20 (the low return days of bank i and the index perfectly overlap). Depending on the specific analysis, the portfolio of banks in the index group will be comprised of either all banks in the sample (excluding bank i), or a bank's high cosine group formed on the basis of high cosine similarity with bank i (highly connected peers). Table 1 shows that the mean value for *LFM Days* when the index is defined as all banks in the sample (excluding bank i) is 4.12 with a standard deviation of 2.36. This measure varies from 1 at the 5^{th} percentile to 9 days at the 95^{th} percentile.

To capture the magnitude dimension of tail-risk, our second measure, *LM AbnRet*, is measured as the average of the abnormal returns of bank i over the 20 days where the portfolio of banks in the index group has it lowest performance for the year. Table 1 shows that the mean value for *LF AbnRet* when the index is defined as all banks in the sample (excluding bank i) is –0.01, with a standard deviation of 0.01.

2.3 Univariate Correlations

In Table 2 we report univariate correlations between our main variables of interest. While our main qualitative connectedness measure of interest is *AvgCos_MDABUS*, we see that this measure has Pearson correlation with *AvgCos_MDA* of 0.84 and with *AvgCos_BUS* of 0.86.

AvgCos_MDA and *AvgCos_BUS* have a Pearson correlation of 0.65, implying that each measure contains orthogonal information.

In terms of tail comovement, Table 2 reports that *AvgCos_MDABUS* has a Pearson correlation with *LFM Days* of 0.16. This implies that banks with higher qualitative connectedness are more likely than less connected banks to have low returns days at the same time that all other banks as a group are also experiencing low returns. Similarly, *AvgCos_MDABUS* has a Pearson correlation with *LF AbnRet* of -0.15, implying that banks with higher qualitative connectedness have lower average abnormal returns than less connected banks on days when the banking sector as a whole is experiencing low returns.

3. Empirical Results

In this section, we discuss the main empirical results of our analyses of relations between our text-based bank connectedness measures and tail comovement across banks. The section is organized as follows. We begin our analyses by examining the relations between tail comovement and bank connectedness measures computed using the combined text of BUS and MD&A Section. In Section 3.1 we measure connectedness using a bank's average cosine similarity with all other banks and its comovement with all other banks, while in Section 3.2 we construct each bank's high cosine group and then examine whether a bank exhibits more future tail comovement with banks in its high cosine banks than with less connected banks. Section 3.3 uses a contagion framework (e.g., Boyson et al., 2010) to examine whether a bank's poor performance days are associated with a higher proportion of banks in its high cosine group also experiencing poor performance days than the proportion of banks with less connectedness. In section 3.4 we examine the incremental informativeness for predicting future tail comovement of connectedness measures computed separately for BUS, MD&A, risk factor disclosures and

footnote disclosures. Finally, in Section 3.5 we disaggregate combined BUS and MD&A into boilerplate and non-boilerplate sentences and examine the incremental informativeness and timeliness of cosine similarity measures computed separately for each group of sentences.

3.1 Average Cosine Analyses Using Combined BUS and MD&A Discussions

In this section, we focus on banks' text-based connectedness measured as a bank's average cosine similarity with all other banks in the market. We examine relations between this connectedness measure and future tail comovement with all other banks, Specifically, we estimate the following multivariate OLS regression:

$$RiskMeasure_{i,t} = \beta_0 + \beta_1 AvgCos\_BUSMDA_{i,t-1} + \beta_2 Size_{i,t-1} + \beta_3 Beta_{i,t-1} + \beta_4 RiskMeasure_{i,t-1} + YearFE + \varepsilon_{i,t}\,, \quad (1)$$

where the $RiskMeasure_{i,t}$ variable is defined as either *LFM Days*$_{i,t}$ or *LM AbnRet*$_{i,t}$ to proxy for tail comovement. The coefficient β_1 captures the extent to which a bank's average text-based cosine similarity with all other banks in the market is associated with a bank's susceptibility to future tail comovement. We control for bank size ($Size_{i,t}$) measured as the log of total assets, the bank's general return correlation with the banking sector ($Beta_{i,t}$), and the lagged $RiskMeasure_{i,t-1}$ (*LFM Days*$_{i,t-1}$ and *LM AbnRet*$_{i,t-1}$). We also include year fixed effects and cluster standard errors by both bank and year. Detailed descriptions of all variables are contained in the Appendix.

The results from the estimation of (1) are reported in Table 3. In Table 3, columns 1 and 3 report the results for both *LFM Days* and *LM AbnRet* respectively. We see in column 1 that the coefficient on *AvgCos_BUSMDA* is 3.76 and is significant at the 0.01 level. This result suggests that the higher the average similarity of a bank's business and MD&A discussion to all other banks' discussion in a given year, the more susceptible the bank is to systemic risk. This result is

economically significant, where a one standard deviation increase in *AvgCos_BUSMDA* results in an 8.4% increase in the number of days that bank i and the banking market overlap in their lowest return days.

In column 3 we report the result of estimating equation (1) using *LM AbnRet* as the risk measure. Because *LM AbnRet* is a returns-based measure of poor performance, negative coefficients on our measures of connectedness are consistent with greater future tail comovement. Similar to the results in column 1, we find that our qualitative measure of similarity, *AvgCos_ BUSMDA*, is statistically significantly negative. The results in column 3 suggest that for a one standard deviation increase in our qualitative measure *AvgCos_ BUSMDA* there is a 16.6% reduction in the bank's average abnormal return over the banking market's lowest return days.

Robustness - Controlling Accounting-Based Quantitative Measure of Connectivity

One potential concern is that our text-based measure might only be capturing what is already found in the quantitative accounting numbers found in the regulatory filings. To allay this concern, we construct a novel measure of banks' quantitative similarity by estimating pairwise cosine similarity between banks based on the entire, standardized vector of quantitative accounting data required to be reported in banks' mandated regulatory filings. Commercial banks subject to the FDIC prepare regulatory filings using a reporting template required by bank regulators. This regulated template structure allows us to construct vectors of accounting data for each bank based on the same standardized set of required reporting fields. Specifically, we calculate cosine similarities using financial institutions mandatory call report filings or FR Y-9C filings as appropriate, which have identical reporting fields (but different call letters). A complete set of call reports is obtained from the Federal Financial Institutions Examination

Council for the year 2001 to 2016. [4] Call reports for the years 1994 to 2000 are obtained from the Federal Reserve Bank of Chicago.[5] FR Y-9C reports were obtained from the datasets provided by the Chicago Federal Reserve Bank.[6]

To ensure consistency across all banks' regulatory reports, we prepare the quantitative data in the call reports and FR Y-9C filings by first aggregating sub-series variables to create main-series variables by summing the respective sub-series variables when necessary.[7] This allows us to create variables which are comparable across reports. A firm's quantitative information is then represented as a vector where the axis is the specific quantitative variable and the magnitude of the axis is its reported value. We then calculate the cosine similarity between each banks' quantitative information vector and those of all others banks in the same calendar year. This measure allows us to assess the similarity across the entire set of quantitative measures reported by banks in their regulatory reports (*AvgCos_Report*). While not tabulated, the average quantitative cosine similarity between a given bank and all other banks in the sample in a given year, *AvgCos_Report*, has a mean value of 0.72 with a standard deviation of 0.18. Also, *AvgCos_BUSMDA* has Pearson correlation with *AvgCos_Report* of only 0.29, suggesting that there is substantial scope for our qualitative connectedness measure to contain incremental information about tail comovement relative to quantitative similarity.

After computing the *AvgCos_Report* we have a sample of 5,499 bank years. This is smaller than those used for our qualitative connectedness measures because we require financial

[4] https://cdr.ffiec.gov/public/PWS/DownloadBulkData.aspx .
[5] https://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-data .
[6] https://www.chicagofed.org/banking/financial-institution-reports/bhc-data .
[7] Main series variables are the sum of certain sub series variables. Some banks gave main series variables while others provide the sub series variables underlying the main series variables. For a description of the main and sub series variables see https://www.federalreserve.gov/apps/mdrm/series .

institutions to have a regulatory report. We then re-estimate equation (1) for both *LFM_Days* and *LM_AbnRet* and report the result in columns (2) and (4) respectively.

In column 2 we find when both *AvgCos_BUSMDA* and *AvgCos_Report* are included, *AvgCos_Report* has a reported coefficient of 0.76 and it is statistically significant at the 0.01 level. However, the results in column (2) show that our qualitative measure *AvgCos_BUSMDA* still remains statistically significant, although the economic significance of the *AvgCos_BUSMDA* drops from about 8.4% to a 4.4% increase in the number of days that bank i and the banking market overlap in their lowest return days. In column 4 we find similar results as that found in column 2. Specifically we see that while *AvgCos_Report* is negative and statistically significant, *AvgCos_BUSMDA* remains negative and statistically significant.

The results in Table 3 show that our qualitative connectedness measure based off the bank's discussion in the business section and MD&A section of the 10-K is significantly associated with a bank's future tail comovement with other banks. However, while these analyses considered a bank's average text similarity and tail comovement with all other banks, it is likely that a bank's connectedness and tail comovement is not uniform across banks. Rather, a bank's tail comovement should be significantly higher with banks with which it is most similar than it is with less similar banks. To explore this possibility, in the next section we examine the extent to which the matrix of pairwise text similarities scores allows us to effectively cluster the banking sector into subsets of banks for which future tail movement is expected to be the highest.

3.2 Group Cosine Analysis

A benefit of our methodology in computing similarity is that it allows us to refine our definition of bank connectedness to focus on subgroups within the market that share significant similarities. We use this subgroup analysis to explore the possibility that text similarity can identify groups of banks that are particularly susceptible to tail comovement.

We construct a high cosine subgroup for each bank by matching it to other banks with which it is most similar in a given year based on cosine similarity scores. We include bank j in bank i's high cosine similarity group if their text similarity score is above certain cutoff percentiles of the distribution of bank i's similarity scores across all other banks. The percentiles that we use are 5%, 10%, 25%, and 50% of the qualitative similarity score distribution. A 5% cutoff selects the 5% of all other banks with which a given bank is most similar. We repeat this process each calendar year and so allow these groups to evolve dynamically over time. Note that this implies that in a given year, each bank will have the same number of banks in its high cosine group. An interesting property of such classification is that for each individual bank, the group of banks that are in close proximity in similarity need not be the same. For example, suppose that for Bank A the banks in the market with which it is most similar are Bank B and Bank C. However, it is possible that for Bank B the two most similar banks in the market are Bank X and Bank Z.

It is also possible that the banks in close proximity to a given bank change over time as strategies and circumstances evolve. We develop some descriptive statistics to examine the dynamic evolution of peer groups through time. For each bank, in each year we construct a vector that reflects the banks in its peer group that year. The number of elements in these vectors is equal to the number of banks minus 1 (to exclude the bank around which the peer group is

built). Then for each bank, we compute the cosine similarity between the vectors for year t and t+1. This cosine similarity score provides information on how similar the peer groups are across years. We compute this for every bank in a year and compute the average cosine similarity across banks for the year. We plot the results in Figure 1. While there is evidence of some persistence, there is also evidence of significant change over time in the banks comprising peer groups. If we use 5% (50%) as cutoffs to determine high cosine groups, there is an average change of approximately 50% (20%) in the banks comprising peer groups.

Our first test examines the difference in future tail comovement between a bank and the group of banks in its high cosine group. We conjecture that a bank will exhibit more tail comovement with those banks with which it has high cosine similarity than with other banks of lower similarity. We run the following multi-variate regression:

$$Risk\ Measure_{i,t,(HCG\ i,t-1)} = \beta_0 + \beta_1 High\ Cosine\ Group_{i,t-1} + \beta_2 Size_{i,t-1} + \beta_3 Beta_{i,t-1} + \beta_4 Risk\ Measure_{i,t-1} + FirmFE + YearFE + \varepsilon_{it}, \quad (2)$$

where *High Cosine Group*$_{i,t-1}$ is an indicator set to 1 if the risk measure is calculated using the group of banks which have a high cosine similarity with the given bank, 0 if otherwise. All other variables are defined above, except that we now include a bank fixed effect and the portfolio of banks now used to compute the risk measures *LFM Days*$_{i,t,(HCG\ i,t-1)}$ and *LM AbnRet*$_{i,t,(HCG\ i,t-1)}$ are either those banks in a bank's high cosine similarity subgroup or all the banks not in that subgroup. That is, in a given year there are two future comovement measures computed for each bank, one using its high cosine group and one using the banks not in this group. In equation (2)

our interest is in the sign and significance of β_1, the coefficient on the high cosine group indicator (*High Cosine Group*). This coefficient captures the difference in future tail comovement between banks for a bank's high cosine group relative to banks with lower connectedness. When we use the *LFM Days*$_{i,t,(HCG\ i,t-1)}$ (*LM AbnRet* $_{i,t,(HCG\ i,t-1)}$) we expect a positive (negative) coefficient β_1.

The results for the estimation of (2) are reported in Table 4 panels A and B. As reported, the results in both panels are consistent with our predictions, with the coefficient on *LFM Days*$_{i,t,(HCG\ i,t-1)}$ (*LM AbnRet*$_{i,t,(HCG\ i,t-1)}$) being positive (negative) and significant for all four high cosine cutoff thresholds. Interestingly, in both panels A and B we see that the absolute value of the coefficient monotonically decreases as we move from more the exclusive 5% cutoff to the less exclusive 50% cutoff. For example, in panel A we that *LFM Days*$_{i,t,(HCG\ i,t-1)}$ is 1.3 days higher for banks in the 5% high similarity group, where it is only 0.27 days higher in the 50% similarity group. Note that 1.3 days is a very large increase when compared to the unconditional mean for *LFM Days* of 4.16 days reported in Table 1. Similarly, the economic effects in panel B suggest for the 5% high similarity group there is a -0.0065 decrease in the return or a 54% reduction.

These results provide evidence that our text-based connectedness measure is able to identify subgroups within a bank that are more likely in the future to commove in the tails.

3.3 Examining Connectedness and Tail Comovement in a Contagion Framework

So far, our analyses of text-based connectedness examine the extent to which the returns of a bank are low when conditioning on whether similar banks are experiencing low returns. We now turn this around and examine whether the returns of the group of banks with which an

individual bank is similar are low when conditioning on the individual bank's returns being low. While this approach is clearly related to our previous analyses, it further allows us to consider how comprehensive the tail comovement effect across connected bank peers by examining the proportion of banks in a bank's high cosine group that are having a low return day when the bank is having a low return day. Our analysis builds on framework used in the Boyson et al. (2010) analysis of hedge fund contagion. Specifically, we estimate the following models by high and low cosine group for each bank:

$$Proportion\ Low_{i,t,(HCG\ i,t-1)} = \beta_0 + \beta_1 Low\ Day_{i,t} + \beta_2 High\ Cosine\ Group_{i,t-1} + \beta_3 High\ Cosine\ Group_{i,t-1} x\ Low\ Day_{i,t} + FirmFE + YearFE + \epsilon \quad (3)$$

where *Proportion Low*$_{i,t,(HCG\ i,t-1)}$ is the proportion of banks in a given bank's subgroup on day t that are have a low performance day. Recall, these subgroups are either banks in a given bank's high cosine group or banks outside this group. *Low Day* is an indicator variable set to 1 for bank i on day t if its daily abnormal return is in the bottom 5% of the entire set of its daily returns, and zero otherwise. We then interact *Low Day* with *High Cosine Group*$_{i,t-1}$ which is an indicator set to 1 if the dependent variable is calculated using the group of banks which have a high cosine similarity with the given bank, and 0 if otherwise. We predict that the coefficient on the interaction *Low Day* x *High Cosine Group*$_{i,t-1}$ will be positive, consistent with the proportion of banks also experiencing a low return when bank i is experiencing a low return day being relatively higher for banks within its high cosine similarity subgroup.

We estimate equation (3) and report the results in Table 5. Similar to Table 4, we again present the results for different group similarity cutoffs (i.e., 5%, 10%, 25%, and 50%). As

predicted, the coefficient on the interaction *Low Day* x *High Cosine Group*$_{i,t-1}$ is positive and significant for all four similarity cutoffs.

3.4 Relative Informativeness of Different Aspects of 10-K Text

While we have up to now used the combined text the BUS and MD&A sections of the 10-K report, it is possible that the informativeness of text similarity measures may vary across distinct segments of the 10-K. We explore this possibility by examining the incremental informativeness for predicting future tail comovement of connectedness measures computed separately for BUS, MD&A, risk factor disclosures and footnote disclosures.

Before looking at the multivariate results, Figure 2 plots over our sample period the average of both *AvgCos_BUS* and *AvgCos_MDA*. From the graph it is evident that over the sample period *AvgCos_BUS* as remained much flatter compared to *AvgCos_MDA*. To examine the information content of these measures, Table 6 reports the results of considering separate average cosine text similarity measures for BUS (*AvgCos_BUS*) and MD&A (*AvgCos_MDA*). Columns 1-3 report results when comovement is measured by *LFM Days* and columns 3-6 for *LM AbnRet*. We see in Table 6 that for *LFM Days* (*LM AbnRet*), the coefficients on *AvgCos_BUS* and *AvgCos_MDA* are both positive (negative) and significantly different from zero ($p<.01$) when included separately, and that each has significant incremental informativeness when they are included simultaneously. However, the economic significance of business description similarity is higher than MD&A similarity. For *LFM Days* (*LM AbnRet*), a one standard deviation increase in BUS similarity results in a 6% (11%) increase in future tail comovement, while for MD&A similarity there is only a 2.5% (3%) increase.

We next consider the incremental informativeness for predicting future tail comovement contained in the text similarity of financial statement footnotes across banks. For each bank, we compute average cosine similarity based on the text in the footnotes, *AvgCos_Notes*. In Table 7, we see that *AvgCos_Notes* is significantly associated with future tail comovement when included alone. However, when we also include *AvgCos_BUSMDA* in addition to *AvgCos_Notes*, Table 7 shows that while *AvgCos_BUSMDA* continues to be significantly associated with future tail comovement, *AvgCos_Notes* has no incremental informativeness relative to the information contained in *AvgCos_BUSMDA*.

Finally, we also consider the incremental informativeness contained in the text similarity of required 10-K risk factor disclosures across banks. The SEC only required these disclosures after 2005, and so our sample size is significantly smaller (approximately half of the original sample size) for this analysis. For each bank, we compute average cosine similarity based on the text in the risk factor disclosure section of the 10-K, *AvgCos_Risk*. In Table 7, we see that *AvgCos_Risk* is significantly associated with future tail comovement when included alone. When we also include *AvgCos_BUSMDA* in addition to *AvgCos_Risk*, Table 7 shows that *AvgCos_BUSMDA* still has incremental information over and above *AvgCos_Risk* about future tail comovement. However, note that the coefficient on *AvgCos_Risk* in columns 4 and 5 is *negative*, suggesting that it is useful in rebalancing the weighting across different dimensions of fundamentals aggregated together within business description and MD&A similarity.

3.5 Relative Information Content of Boilerplate and Non-Boilerplate Language

The previous section provides evidence on the information content of similarities in different defined sections of the 10-K. In this section, we shift our focus from the information

content in similarity within a section of the 10-K to understanding similarities in the nature of the language used in the 10-K. Specifically, we are interested in understanding more about the information content of similarities in boilerplate and non-boilerplate language for understanding future tail risk.

We begin by separating our primary measure, *AvgCos_BUSMDA*, into boilerplate and non-boilerplate similarity following the methodology from Dyer et al., (2017). Specifically, a sentence is designated as boilerplate if it contains a 4-word phrase which appears in more than 60% of the BUS or MD&A disclosures of all banks in a given year. Stop words are not included in the 4-word phrases. We then compute a separate average similarity measure for all boilerplate sentences found in the business section and MD&A, termed *AvgCos_BUSMDA_Boilerplate*. All other sentences in a bank's business and MD&A sections that are not boilerplate are classified as non-boilerplate sentences. Taking the non-boilerplate sentences we again construct a cosine similarity measure and term it *AvgCos_BUSMDA_NonBoilerplate*.

In figure 3, we plot the annual average across all banks of the average cosine similarity of boilerplate and non-boilerplate sentences from 1997 to 2014. The figure shows that non-boilerplate similarity has steadily, but slowly increased over time. In contrast, we see that while boilerplate similarity increased between 1997 and 2004, since 2004 it has very slowly decreased. Thus, while Dyer et al. (2017) show that in recent years the extent of boilerplate language has increased significantly, in banking we do not see this increase in the use of boilerplate language translating into higher boilerplate similarity across banks.

In Table 8, we report the results of including average boilerplate and non-boilerplate similarity as separate variables. When comovement is measured by either *LFM Days* or *LM AbnRet*, both average boilerplate and non-boilerplate similarity are seen to have incremental

information content. However, the economic significance of non-boilerplate similarity is higher than boilerplate similarity. For *LFM Days* (*LM AbnRet*), a one standard deviation increase in non-boilerplate similarity results in a 7.5% (12.7%) increase in future tail comovement, while for boilerplate similarity there is only a 1.6% (2%) increase.

Timeliness of Boilerplate and Non-Boilerplate Language

Given the results in table 8, we next investigate the timing aspects of boilerplate vs non-boilerplate information. From Figure 3 we can see that at the aggregate level there is lower variation in the boilerplate similarity compared to non-boilerplate. Given similarity in boilerplate discussions has information content but little variability over time, it may be possible that the nature of the boilerplate discussions is different and potentially capture the underlying structural fundamentals of the bank. This would potentially make non-boilerplate discussion timelier.

In table 9, we explore the possibility that non-boilerplate similarity is timelier than boilerplate similarity. Specifically, we include 3 lags of non-boilerplate and boilerplate similarity. The results show only the third lag (t-3) of boilerplate similarity is significantly associated with tail comovement at time t. For *LFM Days* (*LM AbnRet*), a one standard deviation increase in boilerplate similarity at t-3 results in a 2.6% (3.5%) increase in future tail comovement. In contrast, for non-boilerplate similarity the coefficient on first lag (t-1) loads significantly (the second lag also loads for *LFM Days*), while the third lag does not load significantly. These results suggest that non-boilerplate similarity is timelier than boilerplate, consistent with non-boilerplate similarity capturing commonalities across banks in currently evolving fundamentals and boilerplate similarity capturing commonalities in structural features that evolve slowly over time.

Dyer et al. (2017) documents a significant increase in boilerplate language in recent years. While we have not seen an increase in average boilerplate similarity through time, as a final analysis we perform an exploratory analysis examining whether the information content of boilerplate similarity has increased or decreased over time. To examine this question, we interact non-boilerplate and boilerplate similarity with an indicator variable, *Post-2006*, which is set equal to 1 for years after 2006, and 0 otherwise. While the average boilerplate similarity has not increased over time, the documented increase found in Dyer et al. may result in less information content in the similarity measure. If the information content was different through time, this would lead us to predict a moderating effect on the interaction term.

The results from our analysis are reported in table 10. We find no evidence that the informativeness of boilerplate language has changed in recent years. Both of the interaction terms are statistically insignificant. Interestingly, we do see that the coefficient on non-boilerplate similarity has significantly increased post 2006. These results suggest that while the information content of boilerplate language has remained constant despite the increased use of boilerplate language per Dyer et al., the information content of the non-boilerplate language has increased over time.

Summary

In this paper we examine the informativeness and timeliness of a bank's 10-K discussions for predicting its future downside tail risk comovement with other bank**s.** Our objective is twofold. First, we seek to provide evidence to bank outsiders (bank regulators, investors, researchers, etc.) about the value of incorporating text-based financial analysis into assessments of bank connectedness and related systemic risk exposures. Second, in addition to extending the

risk measurement literature we seek to provide new insights into the usefulness of 10-K textual disclosure by applying these disclosures in a novel decision context, the prediction of future tail comovement among banks.

To measure connectedness, we construct time-varying measures of cosine similarity between a bank's business description and MD&A discussions and those of all other publicly traded banks. We then use the matrix of pairwise similarities to design measures of connectedness between banks and investigate the extent to which these connectedness measures predict future tail comovement among connected banks.

Focusing first on average similarity of a bank's textual disclosures with those of all other banks, we find that comovement between the lower tail of a given bank's future equity return distribution and the lower tail of the banking system's returns is increasing in the bank's average similarity. While this analysis considers a bank's average text similarity, it is plausible that a bank's connectedness and tail comovement is not uniform across banks. Rather, a bank's tail comovement should be significantly higher with banks with which it is most similar than it is with less similar banks. To explore this possibility, we construct groups of connected peer banks with the most text similarity, finding that banks co-move significantly more in the tails with its highest similarity peers than with lower similarity banks.

To examine the relative informativeness of different aspects of 10-K textual discussions, we disaggregate similarity into business description and MD&A components finding that, while both predict future tail comovement, the economic significance of business description similarity is much higher than MD&A similarity. Further, footnote text similarity has no incremental explanatory power relative to business description and MD&A text similarity. We also show that

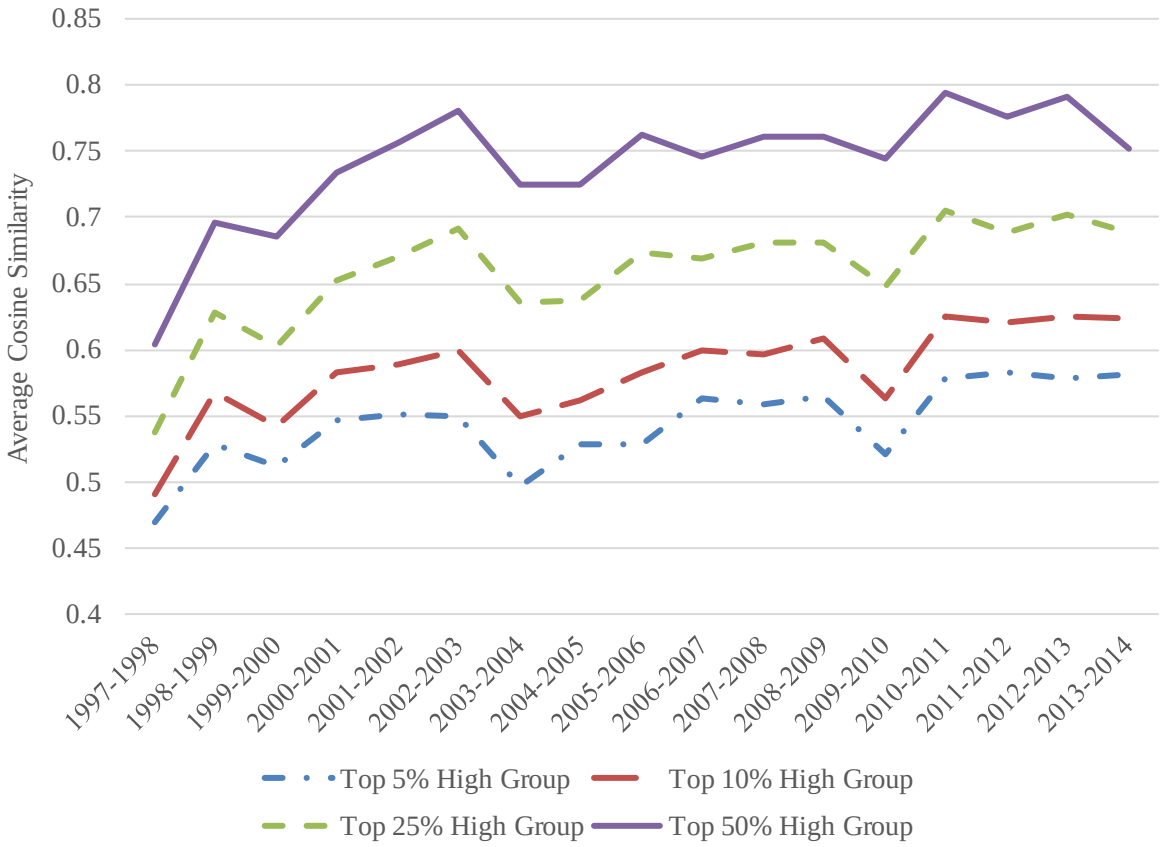
business description and MD&A similarity and risk factor disclosure both have incremental information.

Finally, we separate 10-K text into boilerplate and non-boilerplate components. We document that boilerplate similarity across banks has not significantly increased in recent years, and that both boilerplate and non-boilerplate similarity have incremental information about future tail comovement. However, non-boilerplate similarity is significantly timelier than boilerplate similarity.

Appendix 1
Variables and Grouping Descriptions

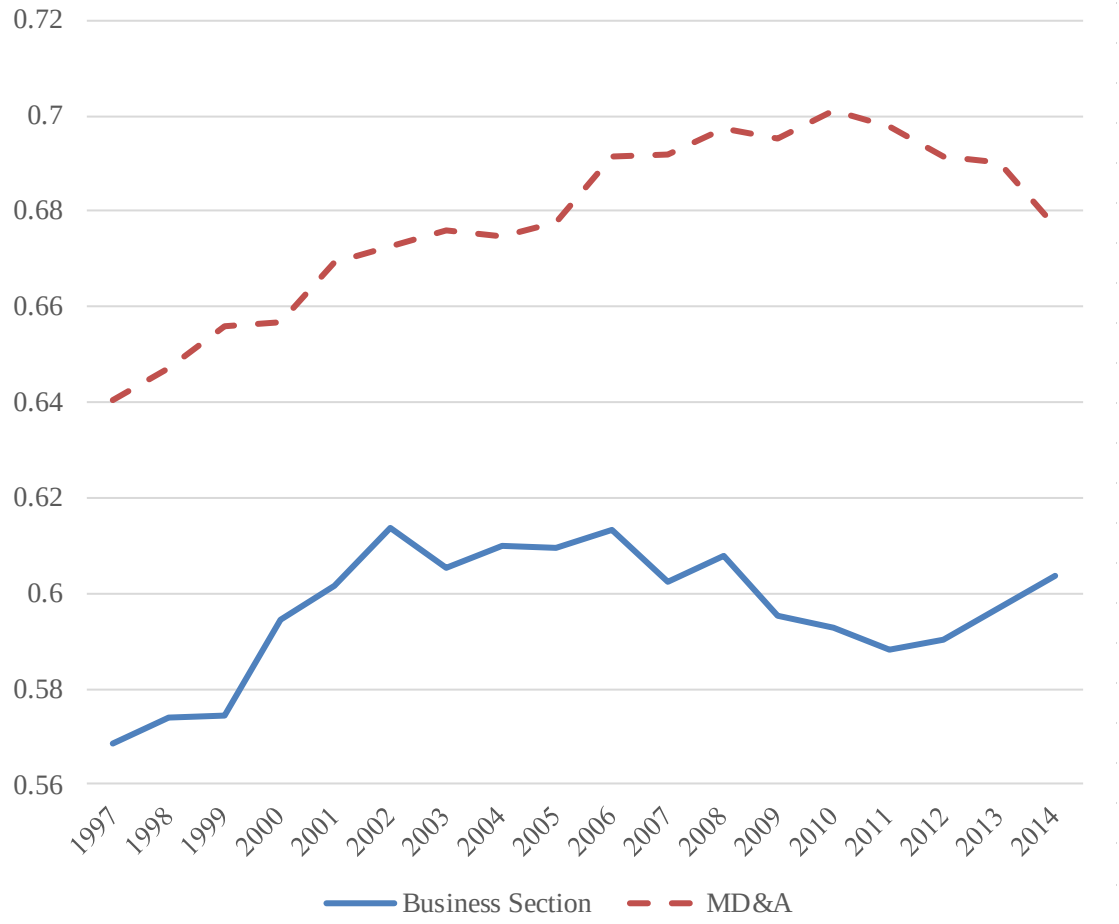
Variable	Description
AvgCos_BUSMDA $_{i,t}$	Average of the cosine similarity of the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t.
AvgCos_MDA $_{i,t}$	Average of the cosine similarity of the MD&A (Item 7) between financial institution i and all other financial institutions filed in year t-1.
AvgCos_BUS $_{i,t}$	Average of the cosine similarity of the Business Section (Item 1) between financial institution i and all other financial institution filed in year t-1.
AvgCos_BUSMDA_(Non)Boilerplate $_{i,t}$	Average of the cosine similarity of the boilerplate (non-boilerplate) sentences in the Business Section (Item 1) and MD&A (Item 7) between financial institution i and all other financial institutions filed in year t. A sentence is marked as boilerplate if it contains a tetragram which appears in more than 60% of all financial institutions Business Section or MD&A disclosures in the given year. Stopwords are not included in the tetragrams.
AvgCos_Report $_{i,t}$	Average of the cosine similarity of the items in the Y-9C or Call Report between financial institution i and all other financial institutions filed in year t-1.
LFM Days $_{i,t,(HCG\ i,t)}$	The number of days in year t where bank i and banks in the group ($HCG_{i,t}$) both have low returns performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the group). $HCG_{i,t}$ (High Cosine Group) is defined below.
LM AbnRet $_{i,t,(HCG\ i,t)}$	The average abnormal returns of bank i on days when the group ($HCG_{i,t}$) has low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the group). $HCG_{i,t}$ (High Cosine Group) is defined below.
Size $_{i,t}$	Log of total assets for financial institution i in year t.
Beta $_{i,t}$	Market Beta measured using the market model and a rolling three year window of returns.
High Cosine Group $_{i,t}$	An indicator set to 1 if the given observation is computed using banks in the high cosine group, 0 otherwise. Banks in the high cosine group are chosen based upon their cosine similarity with firm i (see High(Low) Cosine Group Cutoff).
Low Day $_{i,t}$	An indicator set to 1 for financial institution i on day t if its daily abnormal return is in the bottom 5% of all of its daily returns in our sample
Prop. Low $_{i,t,(HCG\ i,t)}$	The proportion of firms in the group ($HCG_{i,t}$) for firm i on day t which have low performance (i.e. Low $Day_{i,t}$ = 1). $HCG_{i,t}$ is an indicator which denotes if the group is the high cosine group. High cosine group is defined as above.
High(Low) Cosine Group Cutoff	High(Low) Cosine Group Cutoff denotes the cutoff for high (low) cosine similarity. For example 5% means that a pair of financial institutions have high cosine similarity if their cosine similarity is above the 5th percentile of the cosine distribution.

Figure 1
Change in the Composition of a Bank's Group Over Time



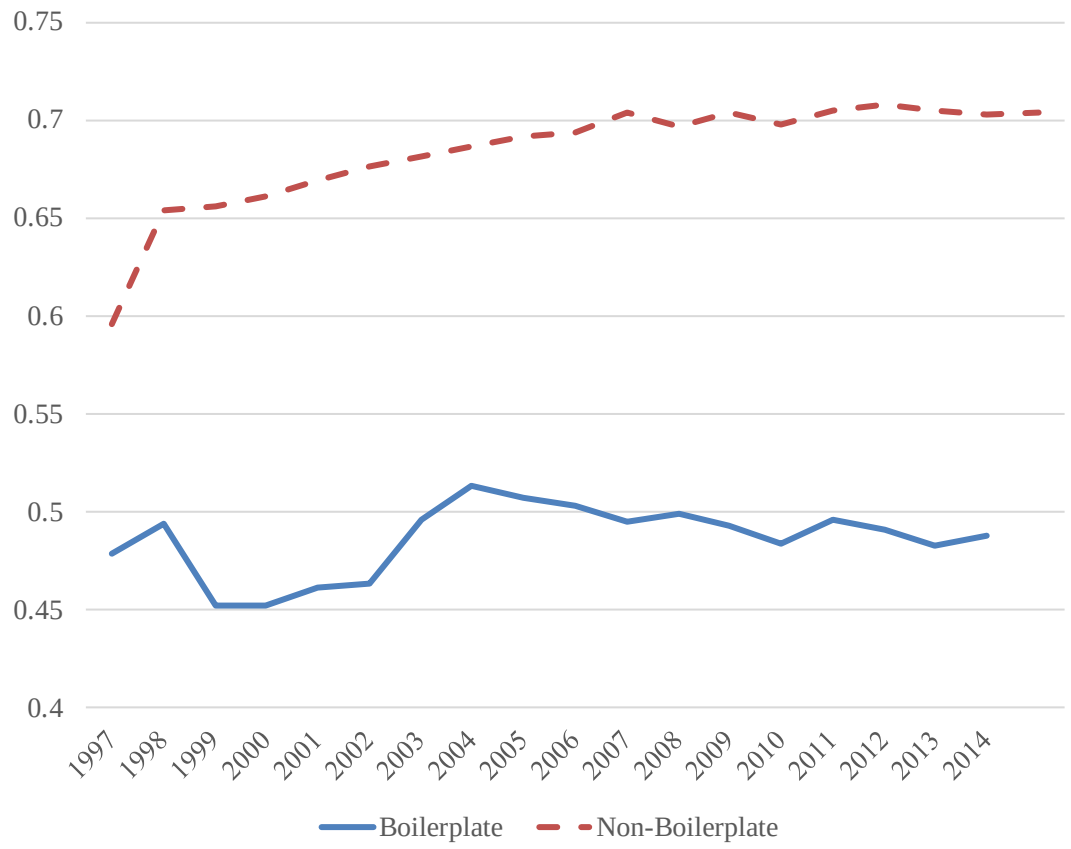
This figure presents the change in the banks which comprise a banks group over time. A banks group is constructed using the top 5%, 10%, 25%, and 50% of the given banks cosine similarity distribution. The change in the composition of a banks group is calculated using the cosine similarity between the banks in its group in year t and year t+1. The figure shows the average of these cosine similarities across all banks in a given year.

Figure 2
Average of the Average Cosine of the Business Section and MD&A



This figure presents the average of the average cosine similarity of firms Business Section (Item 1) and Management Discussion and Analysis Section (Item 7) of the 10-K filing.

Figure 3
Average of the Average Cosine of Boilerplate and Non-Boilerplate



This figure presents the average of the average cosine similarity of the boilerplate and non-boilerplate sentences in firms Business Section (Item 1) and Management Discussion and Analysis Section (Item 7) of the 10-K filing.

Table 1

Summary Statistics

Variable	N	Mean	P5	P25	Median	P75	P95	σ
$AvgCos\_BUSMDA_{i,t}$	8,784	0.694	0.485	0.659	0.727	0.759	0.787	0.093
$AvgCos\_BUS_{i,t}$	8,307	0.600	0.341	0.558	0.642	0.683	0.716	0.116
$AvgCos\_MDA_{i,t}$	8,782	0.682	0.499	0.646	0.710	0.745	0.774	0.087
$LFM\ Days_{i,t}$	8,903	4.164	1.000	2.000	4.000	6.000	9.000	2.443
$LM\ AbnRet_{i,t}$	8,747	-0.012	-0.031	-0.018	-0.010	-0.005	0.002	0.011
$Size_{i,t}$	8,785	7.294	5.355	6.276	7.004	8.047	10.265	1.524
$Beta_{i,t}$	8,783	0.626	0.008	0.164	0.459	1.049	1.639	0.542

This table presents descriptive statistics for the main variables used in our study.

$AvgCos\_BUSMDA_{i,t}$ is the average of the cosine similarity of the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. $AvgCos\_BUS_{i,t}$ is the average of the cosine similarity of the Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. $AvgCos\_MDA_{i,t}$ is the average of the cosine similarity of the MD&A (Item 7) between financial institution i and all other financial institutions filed in year t-1. $LFM\ Days_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). $LM\ AbnRet_{i,t}$ The average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. $Size_{i,t}$ is the log total assets for financial institution i in year t. $Beta_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 2

Pearson and Spearman Pair-Wise Correlations

Variable		1	2	3	4	5	6	7
$AvgCos\_BUSMDA_{i,t}$	1	-	0.75	0.85	0.11	-0.08	0.00	-0.20
$AvgCos\_BUS_{i,t}$	2	0.84	-	0.56	0.15	-0.13	-0.04	-0.26
$AvgCos\_MDA_{i,t}$	3	0.86	0.65	-	0.09	-0.08	0.05	-0.15
$LFM\ Days_{i,t}$	4	0.16	0.19	0.13	-	-0.63	0.03	-0.09
$LM\ AbnRet_{i,t}$	5	-0.15	-0.17	-0.13	-0.55	-	0.17	0.21
$Size_{i,t}$	6	0.01	-0.01	0.02	-0.01	0.16	-	0.64
$Beta_{i,t}$	7	-0.26	-0.31	-0.19	-0.12	0.19	0.57	-

This table presents the pair-wise pearson (below diagonal) and spearman correlations (above diagonal) of the main variables used in our study.

$AvgCos\_BUSMDA_{i,t}$ is the average of the cosine similarity of the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. $AvgCos\_BUS_{i,t}$ is the average of the cosine similarity of the Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. $AvgCos\_MDA_{i,t}$ is the average of the cosine similarity of the MD&A (Item 7) between financial institution i and all other financial institutions filed in year t-1. $LFM\ Days_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). $LM\ AbnRet_{i,t}$ The average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. $Size_{i,t}$ is the log total assets for financial institution i in year t. $Beta_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 3

Regression of Downside Tail Risk Comovement on Banks Average Cosine

Variable	(1) $LFM\ Days_{i,t}$	(2) $LFM\ Days_{i,t}$	(3) $LM\ AbnRet_{i,t}$	(4) $LM\ AbnRet_{i,t}$
$AvgCos\_BUSMDA_{i,t-1}$	3.7572***	1.9488***	-0.0178***	-0.0089***
	(6.65)	(3.60)	(-5.24)	(-3.09)
$AvgCos\_Report_{i,t-1}$		0.7613***		-0.0026**
		(2.72)		(-2.44)
$Size_{i,t-1}$	0.0034	0.0100	0.0008**	0.0010**
	(0.05)	(0.10)	(2.50)	(2.16)
$Beta_{i,t-1}$	0.2453	0.5730*	-0.0003	-0.0016
	(0.94)	(1.89)	(-0.21)	(-0.73)
$LFM\ Days_{i,t-1}$	0.2899***	0.2596***		
	(5.87)	(5.12)		
$LM\ AbnRet_{i,t-1}$			0.1640**	0.1030
			(2.50)	(1.31)
Constant	0.2984	0.8255	-0.0046	-0.0104***
	(0.51)	(0.95)	(-1.51)	(-2.66)
Fixed Effects	Year	Year	Year	Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	8,370	5,334	8,189	5,280
Adjusted R-Squared	0.258	0.244	0.308	0.323

This table presents the regression of downside tail risk comovement measures on the average cosine of the Business Section and MD&A of financial institutions.

$LFM\ Days_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). $LM\ AbnRet_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. $AvgCos\_BUSMDA_{i,t}$ is the average of the cosine similarity of the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t. $AvgCos\_Report_{i,t}$ is the average of the cosine similarity of the items in the Y-9C or Call Report between financial instituion i and all other financial institutions filed in year t. $Size_{i,t}$ is the log total assets for financial institution i in year t. $Beta_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 4

Downside Tail Risk Comovement in High and Low Cosine Groups

Panel A - Overlap in Bank and Market Low Performance Days

	(1) Cutoff = 5%	(2) Cutoff = 10%	(3) Cutoff = 25%	(4) Cutoff = 50%
Variable	*LFM Days* $_{i,t,(HCG\ i,t-1)}$	*LFM Days* $_{i,t,(HCG\ i,t-1)}$	*LFM Days* $_{i,t,(HCG\ i,t-1)}$	*LFM Days* $_{i,t,(HCG\ i,t-1)}$
High Cosine Group $_{i,t-1}$	1.3156***	1.1111***	0.6077***	0.2713***
	(7.70)	(5.69)	(4.53)	(3.50)
LFM Days $_{i,t-1,(HCG\ i,t-1)}$	0.1701***	0.2010***	0.2056***	0.1386***
	(5.82)	(8.85)	(7.59)	(5.09)
Size $_{i,t-1}$	0.1486***	0.0966*	0.1917***	0.1744**
	(2.84)	(1.77)	(2.89)	(2.16)
Beta $_{i,t-1}$	0.3741*	0.4280**	0.5399**	0.5740**
	(1.93)	(2.02)	(2.15)	(2.02)
Constant	1.4110***	1.9058***	1.7953***	2.5604***
	(3.64)	(4.98)	(4.24)	(4.44)
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	15,196	15,163	15,171	15,189
Adjusted R-Squared	0.293	0.302	0.322	0.370

This table presents the difference in downside tail risk comovement of a financial institution in relation to financial institutions with which it shares a high and low cosine similarity. High/Low Cosine Group Cutoff Percentage (Cutoff) denotes the cutoff for high/low cosine similarity. For example 5% means a pair of financial institutions have high cosine similarity if their cosine similarity is above the 5th percentile of the cosine distribution.

LFM Days$_{i,t,(HCG\ i,t-1)}$ is the number of days in year t where financial institution i and the financial institutions in the group (HCG i,t) both have low returns performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the abnormal returns is used for the group). *LM AbnRet*$_{i,t}$ the average of the abnormal returns of financial institution i over the days where the group (HCG i,t) has low performance. A low performance day is defined the same as above. *High Cosine Group*$_{i,t}$ is a flag set to 1 if LFM Days (LM AbnRet) is calculated using banks in the high cosine group, 0 otherwise. *Size*$_{i,t}$ is the log total assets for financial institution i in year t. *Beta*$_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 4 - Downside Tail Risk Comovement in High and Low Cosine Groups

Panel B - Abnormal Returns on Low Market Days

	(1) Cutoff = 5%	(2) Cutoff = 10%	(3) Cutoff = 25%	(4) Cutoff = 50%
Variable	*LM AbnRet* $_{i,t,(HCG\ i,t-1)}$	*LM AbnRet* $_{i,t,(HCG\ i,t-1)}$	*LM AbnRet* $_{i,t,(HCG\ i,t-1)}$	*LM AbnRet* $_{i,t,(HCG\ i,t-1)}$
High Cosine Group $_{i,t-1}$	-0.0065***	-0.0056***	-0.0036***	-0.0021***
	(-7.12)	(-7.96)	(-6.47)	(-4.34)
LM AbnRet $_{i,t-1,(HCG\ i,t-1)}$	0.2055***	0.2561***	0.2960***	0.1764***
	(4.37)	(4.27)	(4.27)	(3.87)
Size $_{i,t-1}$	-0.0009**	-0.0013***	-0.0017***	-0.0017***
	(-2.15)	(-2.90)	(-3.64)	(-3.39)
Beta $_{i,t-1}$	-0.0022*	-0.0023	-0.0023	-0.0023
	(-1.66)	(-1.47)	(-1.25)	(-1.11)
Constant	0.0036	0.0055	0.0061*	0.0033
	(1.16)	(1.64)	(1.77)	(0.90)
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	14,841	14,832	14,858	14,848
Adjusted R-Squared	0.306	0.324	0.323	0.378

Table 5
Overlap of Bank and Market Low Performance Days

	(1) Cutoff = 5%	(2) Cutoff = 10%	(3) Cuttoff = 25%	(4) Cuttoff = 50%
Variable	*Prop. Low* $_{i,t,(HCG\ i,t-1)}$	*Prop. Low* $_{i,t,(HCG\ i,t-1)}$	*Prop. Low* $_{i,t,(HCG\ i,t-1)}$	*Prop. Low* $_{i,t,(HCG\ i,t-1)}$
Low Day $_{i,t}$	0.0193***	0.0216***	0.0284***	0.0340***
	(3.82)	(4.50)	(5.74)	(6.63)
High Cosine Group $_{i,t-1}$	0.0018	0.0014	0.0005	-0.0001
	(0.72)	(0.57)	(0.31)	(-0.09)
High Cosine Group $_{i,t-1}$ *x Low Day* $_{i,t}$	0.0311***	0.0256***	0.0153***	0.0079***
	(9.08)	(7.96)	(6.78)	(6.96)
Constant	0.0294***	0.0311***	0.0292***	0.0306***
	(19.67)	(22.51)	(34.16)	(76.06)
Fixed Effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	3,900,768	3,900,768	3,900,768	3,900,768
Adjusted R-squared	0.223	0.286	0.362	0.405

This table presents the regression of the proportion of banks which have low performance on a given day on whether the given bank is also having a low performance day. High/Low Cosine Group Cutoff Percentage (Cutoff) denotes the cutoff for high/low cosine similarity. For example 5% means that two banks have high cosine similarity if their cosine similarity is above the 5th percentile of the cosine distribution.

Prop. Low$_{i,t,(HCG\ i,t-1)}$ is the proportion of financial institutions in the group (HCG i,t) on day t which are having a low performance day (i.e. Low Day$_{i,t}$ = 1). *Low Day*$_{i,t}$ is an indicator variable set to 1 for financial institution i on day t if its daily abnormal return is in the bottom 5% of all of its daily returns. *High Cosine Group*$_{i,t}$ is a flag set to 1 if Prop. Low is calculated using banks in the high cosine group, 0 otherwise.

Table 6

Regression of Downside Tail Risk Comovement on the Average Cosine of the Business and MD&A Sections

Variable	(1) $LFM\ Days_{i,t}$	(2) $LFM\ Days_{i,t}$	(3) $LFM\ Days_{i,t}$	(4) $LM\ AbnRet_{i,t}$	(5) $LM\ AbnRet_{i,t}$	(6) $LM\ AbnRet_{i,t}$
$AvgCos\_BUS_{i,t-1}$	2.7498***		2.2387***	-0.0134***		-0.0114***
	(5.95)		(3.95)	(-4.15)		(-3.38)
$AvgCos\_MDA_{i,t-1}$		2.9445***	1.1768***		-0.0137***	-0.0044***
		(7.53)	(2.87)		(-5.15)	(-2.64)
$Size_{i,t-1}$	0.0096	0.0253	0.0013	0.0007**	0.0006*	0.0007**
	(0.14)	(0.37)	(0.02)	(2.07)	(1.84)	(2.04)
$Beta_{i,t-1}$	0.2297	0.1141	0.2604	-0.0003	0.0002	-0.0003
	(0.75)	(0.40)	(0.86)	(-0.12)	(0.13)	(-0.16)
$LFM\ Days_{i,t-1}$	0.3015***	0.3047***	0.3005***			
	(5.25)	(5.44)	(5.29)			
$LM\ AbnRet_{i,t-1}$				0.1816***	0.1909***	0.1817***
				(2.79)	(3.00)	(2.82)
Constant	1.5920***	1.2191**	1.1702**	-0.0072**	-0.0055*	-0.0055*
	(2.84)	(2.35)	(2.14)	(-2.28)	(-1.84)	(-1.81)
Fixed Effects	Year	Year	Year	Year	Year	Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	7,918	8,366	7,770	7,750	8,184	7,602
Adjusted R-Squared	0.272	0.261	0.273	0.323	0.316	0.322

This table presents the regression of downside tail risk comovement on the average cosine of the Business Section (Item 1) and the MD&A Section (Item 7) separately.

$LFM\ Days_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). $LM\ AbnRet_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. $AvgCos\_BUS_{i,t}$ is the averager of the cosine similarity of the MD&A (Item 7) between financial institution i and all other financial institutions filed in year t-1. $AvgCos\_MDA_{i,t}$ *is the a*verage of the cosine similarity of the Business Section (Item 1) between financial institution i and all other financial institution filed in year t-1. $Size_{i,t}$ is the log total assets for financial institution i in year t. $Beta_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 7

Regression of Downside Tail Risk Comovement on Average Cosine Controlling for the Similarity of Other Sections of the 10-K

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variable	*LFM Days* $_{i,t}$	*LFM Days* $_{i,t}$	*LFM Days* $_{i,t}$	*LFM Days* $_{i,t}$	*LFM Days* $_{i,t}$	*LM AbnRet* $_{i,t}$	*LM AbnRet* $_{i,t}$	*LM AbnRet* $_{i,t}$	*LM AbnRet* $_{i,t}$	*LM AbnRet* $_{i,t}$
AvgCos_BUSMDA $_{i,t-1}$		3.5912***		5.1921***	5.1413***		-0.0170***		-0.0189***	-0.0189***
		(6.31)		(5.88)	(5.58)		(-5.02)		(-3.68)	(-3.80)
AvgCos_Notes $_{i,t-1}$	2.1539***	0.4640			1.0882*	-0.0112***	-0.0025			-0.0015
	(3.52)	(0.95)			(1.88)	(-4.42)	(-1.26)			(-0.52)
AvgCos_Risk $_{i,t-1}$			2.0706***	-1.3437**	-1.6566**			-0.0212***	-0.0080**	-0.0073**
			(2.85)	(-2.08)	(-2.46)			(-5.23)	(-2.54)	(-2.47)
Assets $_{i,t-1}$	0.0506	0.0032	0.1833***	0.1327*	0.1470*	0.0005*	0.0007**	-0.0004	-0.0002	-0.0003
	(0.80)	(0.05)	(2.58)	(1.78)	(1.95)	(1.65)	(2.44)	(-0.83)	(-0.42)	(-0.57)
Beta $_{i,t-1}$	0.0019	0.2403	0.3781*	0.6667***	0.6771***	0.0006	-0.0004	-0.0006	-0.0016	-0.0016
	(0.01)	(0.91)	(1.66)	(2.75)	(2.80)	(0.41)	(-0.24)	(-0.37)	(-0.85)	(-0.91)
LFM Days $_{i,t-1}$	0.3084***	0.2909***	0.2056**	0.1638*	0.1476*					
	(5.86)	(5.81)	(2.09)	(1.84)	(1.76)					
LM AbnRet $_{i,t-1}$						0.1898***	0.1630**	0.0343	0.0054	-0.0039
						(2.76)	(2.37)	(0.73)	(0.12)	(-0.08)
Constant	0.9805	0.0909	0.0692	-0.7133	-1.6979	-0.0069***	-0.0032	0.0110**	0.0142**	0.0154***
	(1.64)	(0.14)	(0.06)	(-0.64)	(-1.38)	(-2.78)	(-1.19)	(2.15)	(2.50)	(2.67)
Fixed Effects	Year	Year	Year	Year	Year	Year	Year	Year	Year	Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	7,797	7,652	3,366	3,309	3,067	7,622	7,478	3,275	3,218	2,979
Adjusted R-Squared	0.239	0.253	0.128	0.160	0.160	0.289	0.304	0.242	0.255	0.258

This table presents the regresison of downside tail risk comovement measures on the average cosine of the Business Section and MD&A of financial institutions controlling for the cosine similarity of the Notes to the Finanical Statements and the Risk Factors section of the 10-K filing.

LFM Days$_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). *LM AbnRet*$_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. *AvgCos_BUSMDA*$_{i,t}$ is the average of the cosine similarity of the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t. *AvgCos_Notes*$_{i,t}$ is the average of the cosine similarity of the notes to the financial statements of firm i and all other financial institutions filed in year t. *AvgCos_Risk*$_{i,t}$ is the average of the cosine similarity of the Risk Factors disclosure (Item 1a) of firm i and all other financial institutions filed in year t. *Size*$_{i,t}$ is the log total assets for financial institution i in year t. *Beta*$_{i,t}$ is market Beta for financial institution i in year t measured using the market model a rolling three year window of returns.

Table 8

Regression of Downrisk Tail Risk Comovement on the Average Cosine of Boilerplate and Non-Boilerplate Sentences

Variable	(1) *LFM Days* $_{i,t}$	(2) *LFM Days* $_{i,t}$	(3) *LFM Days* $_{i,t}$	(4) *LM AbnRet* $_{i,t}$	(5) *LM AbnRet* $_{i,t}$	(6) *LM AbnRet* $_{i,t}$
AvgCos_BUSMDA_Boilerplate $_{i,t-1}$	3.1400***		0.9101**	-0.0147***		-0.0032**
	(5.53)		(2.22)	(-4.19)		(-1.99)
AvgCos_BUSMDA_NonBoilerplate $_{i,t-1}$		3.7862***	3.3968***		-0.0179***	-0.0166***
		(7.12)	(6.90)		(-5.08)	(-5.06)
Size $_{i,t-1}$	0.0472	0.0027	0.0031	0.0005	0.0007**	0.0007**
	(0.67)	(0.04)	(0.04)	(1.56)	(2.31)	(2.20)
Beta $_{i,t-1}$	0.0631	0.2495	0.2665	0.0004	-0.0004	-0.0004
	(0.22)	(0.86)	(0.91)	(0.23)	(-0.20)	(-0.23)
LFM Days $_{i,t-1}$	0.3026***	0.2922***	0.2894***			
	(5.42)	(5.35)	(5.35)			
LM AbnRet $_{i,t-1}$				0.1836***	0.1696***	0.1621**
				(2.86)	(2.68)	(2.55)
Constant	1.4633***	0.7772	0.5700	-0.0067**	-0.0035	-0.0027
	(2.71)	(1.34)	(0.95)	(-2.22)	(-1.07)	(-0.80)
Fixed Effects	Year	Year	Year	Year	Year	Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Observations	8,317	8,367	8,166	8,136	8,187	7,988
Adjusted R-Squared	0.260	0.270	0.271	0.315	0.327	0.326

This table presents the regression of downside tail risk comovement on the average cosine of boilerplate and non-boilerplate disclosures.

LFM Days$_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). *LM AbnRet*$_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. *AvgCos_BUSMDA_Boilerplate(NonBoilerplate)*$_{i,t}$ is the average of the cosine similarity of the boilerplate (nonboilerplate) sentences in the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. *Size*$_{i,t}$ is the log of total assets for financial institution i in year t. *Beta*$_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 9
Regression of Downside Tail Risk Comovement on the Average Cosine of Boilerplate and Non-Boilerplate with Lags

Variable	(1) *LFM Days* $_{i,t}$	(2) *LM AbnRet* $_{i,t}$
AvgCos_BUSMDA_Boilerplate $_{i,t-1}$	0.2047	-0.0016
	(0.30)	(-0.72)
AvgCos_BUSMDA_Boilerplate $_{i,t-2}$	-0.1727	0.0025
	(-0.21)	(0.85)
AvgCos_BUSMDA_Boilerplate $_{i,t-3}$	1.4300**	-0.0057**
	(2.07)	(-2.17)
AvgCos_BUSMDA_NonBoilerplate $_{i,t-1}$	2.0156***	-0.0080**
	(2.87)	(-2.21)
AvgCos_BUSMDA_NonBoilerplate $_{i,t-2}$	2.3596***	-0.0042
	(2.69)	(-0.95)
AvgCos_BUSMDA_NonBoilerplate $_{i,t-3}$	-1.0776	-0.0043
	(-1.41)	(-1.20)
Size $_{i,t-1}$	0.0289	0.0005
	(0.33)	(1.25)
Beta $_{i,t-1}$	0.3077	-0.0007
	(1.06)	(-0.33)
LFM Days $_{i,t-1}$	0.2987***	
	(4.95)	
LM AbnRet $_{i,t-1}$		0.1657**
		(2.32)
Constant	0.7284	-0.0064
	(0.80)	(-1.34)
Fixed Effects	Year	Year
Cluster	Firm & Year	Firm & Year
Observations	5,247	5,139
Adjusted R-Squared	0.267	0.321

This table presents the regression of downside tail risk comovement measures on the average cosine of boilerplate and non-boilerplate disclosures with lags.

LFM Days$_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). *LM AbnRet*$_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. *AvgCos_BUSMDA_Boilerplate(NonBoilerplate)*$_{i,t}$ is the average of the cosine similarity of the boilerplate (nonboilerplate) sentences in the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. *Size*$_{i,t}$ is the log of total assets for financial institution i in year t. *Beta*$_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns.

Table 10
Regression of Downside Tail Risk Comovement on Boilerplate and Non-Boilerplate Post 2006

Variable	(1) $LFM\ Days_{i,t}$	(2) $LM\ AbnRet_{i,t}$
$AvgCos\_BUSMDA\_Boilerplate_{i,t-1}$	1.4814***	-0.0039*
	(2.81)	(-1.75)
$AvgCos\_BUSMDA\_Boilerplate_{i,t-1}$ x $Post\text{-}2006_t$	-1.0108	0.0015
	(-1.36)	(0.49)
$AvgCos\_BUSMDA\_NonBoilerplate_{i,t-1}$	2.0053***	-0.0092***
	(3.77)	(-4.06)
$AvgCos\_BUSMDA\_NonBoilerplate_{i,t-1}$ x $Post\text{-}2006_t$	2.2273***	-0.0125***
	(2.58)	(-2.73)
$Size_{i,t-1}$	-0.0064	0.0010**
	(-0.07)	(2.41)
$Beta_{i,t-1}$	-1.1406***	0.0054*
	(-2.74)	(1.89)
$LFM\ Days_{i,t-1}$	0.2596***	
	(6.27)	
$LM\ AbnRet_{i,t-1}$		0.1745***
		(2.72)
Constant	1.9054***	0.0120**
	(3.77)	(2.01)
Fixed Effects	Year	Year
Cluster	Firm & Year	Firm & Year
Observations	8,166	7,988
Adjusted R-Squared	0.301	0.370

This table presents the regression of downside tail risk comovement measures on the average cosine of boilerplate and non-boilerplate disclosures interacted with an indicator variable for post-2006.

$LFM\ Days_{i,t}$ is the number of days in year t where bank i and the banking market both have low performance. A low performance day is the lowest 20 days based upon daily abnormal returns in year t (the sum of the value weighted abnormal returns is used for the market). $LM\ AbnRet_{i,t}$ the average abnormal returns of bank i over days where the banking market has low performance. A low performance day is defined the same as above. $AvgCos\_BUSMDA\_Boilerplate\ (NonBoilerplate)_{i,t}$ is the average of the cosine similarity of the boilerplate (nonboilerplate) sentences in the combined MD&A (Item 7) and Business Section (Item 1) between financial institution i and all other financial institutions filed in year t-1. $Size_{i,t}$ is the log of total assets for financial institution i in year t. $Beta_{i,t}$ is market Beta for financial institution i in year t measured using the market model and a rolling three year window of returns. $Post\text{-}2006_t$ is an indicator variable set to 1 if the year is greater than or equal to 2006, 0 otherwise. Interactions between all control variables and the post-2006 indicator are included.

References

Acharya, V., Pederson, L., Philippon, T., and M. Richardson, 2017. Measuring systemic risk. *Review of Financial Studies* 30 (1), 2-47.

Adrian, T. and M. Brunnermeier, 2016. CoVaR. *American Economic Review*, 106, 7, 1705-1741.

Amel-Zadeh, Amir and Faasse, Jonathan, 2016. The Information Content of 10-K Narratives: Comparing MD&A and Footnotes Disclosures. (November 27, 2016). Working paper University of Oxford - Said Business School.

Ball, Christopher and Hoberg, Gerard and Maksimovic, Vojislav, 2015. Disclosure, Business Change and Earnings Quality. Working paper USC and Maryland.

Bisias, D., Flood, M., Lo, A. and Valavanis. S. 2012. A Survey of Systemic Risk Analytics. Working Paper 0001, Office of Financial Research.

Billio, Monica and Lo, Andrew W. and Getmansky, Mila and Pelizzon, Loriana, 2012. Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors. *Journal of Financial Economics* 104, 535-559.

Boyson, Nicole M. and Stahel, Christof W. and Stulz, René M., 2010. Hedge Fund Contagion and Liquidity Shocks. *Journal of Finance*, Vol. 55, No. 5, pp. 1789-1816

Brown, S. V., Tucker, J.W., 2011. Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. *Journal of Accounting Research* 49, 309–346.

Bushman, Robert M. and Hendricks, Bradley E. and Williams, Christopher D., 2016. Bank Competition: Measurement, Decision-Making, and Risk-Taking. *Journal of Accounting Research* Volume 54, Issue 3: 777–826.

Bushman, R., A. Smith, and R. Wittenberg-Moerman, 2010, Price discovery and dissemination of private information by loan syndicate participants, Journal of Accounting Research, 48, 921-972.

Bushman, R., and R. Indjejikian, 1993. Stewardship Value of "Distorted" Accounting Disclosures. The Accounting Review, October 1993, 765-782.

Cai, Jian and Saunders, Anthony and Steffen, Sascha, 2016. Syndication, Interconnectedness, and Systemic Risk (December 2011). NYU Working Paper No. 2451/31373.

Cohen, L., Malloy C., & Ngyuen, Q. (2015). Lazy Prices. Harvard Business School, working paper.

Dechow, P.M., Ge, W., Schrand, C., 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics* 50, 344–401.

Dyer, Travis and Lang, Mark H. and Stice-Lawrence, Lorien, 2017. The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation. Forthcoming *Journal of Accounting and Economics*.

Feldman, R., Govindaraj, S., Livnat, J., & Segel, B. (2010). Management's tone change, post earnings announcement drift and accruals. Review of Accounting Studies, 15(4), 915-953.

Gjesdal, Frøystein, 1981. Accounting for Stewardship. *Journal of Accounting Research*, Vol. 19, No. 1 (Spring, 1981), pp. 208-231

Hanley , Kathleen Weiss and Hoberg, Gerard,2016. Dynamic Interpretation of Emerging Systemic Risks. Working paper USC.

Hansen, L. 2014. Challenges in Identifying and Measuring Systemic Risk. In Risk Topography: Systemic Risk and Macro Modeling M.K. Brunnermeier and A. Krishnamurthy, Eds., University of Chicago Press:15 – 30.

Hoberg, Gerard and Phillips, Gordon M., 2016. Text-Based Network Industries and Endogenous Product Differentiation (July 3, 2015). *Journal of Political Economy*, 124, no. 5: 1423-1465.

Huang, X., Zhou, H., Zhu, H., 2011. Systemic risk contribution, 2011. Working paper 2011-08. Board of Governors of the Federal Reserve System.

Joydeep, A.S., Strehl, E., Ghosh, J., Mooney, R., Strehl, A., 2000. Impact of Similarity Measures on Web-page Clustering. In Workshop on Artificial Intelligence for Web Search (AAAI 2000) 58–64.

Kearney, C., and S. Liu. Textual Sentiment in Finance: A Survey of Methods and Models. nInternational Review of Financial Analysis 33 (2014): 171–85.

Kogan, J., Nicholas, C., Teboulle, M., 1998. Grouping multidimensional data: Recent advances in clustering, Grouping Multidimensional Data: Recent Advances in Clustering.

KPMG, 2011. Disclosure overload and complexity: hidden in plain sight. Available at: http://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/disclosureoverload-complexity.pdf.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45(2-3), 221-247.

Loughran, Tim and McDonald, Bill, 2016. Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, Vol. 54, No. 4, 1187-1229.

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. The Journal of Finance, 69(4), 1643-1671.

William J. Mayew, Mani Sethuraman, and Mohan Venkatachalam (2015) MD&A Disclosure and the Firm's Ability to Continue as a Going Concern. The Accounting Review: July 2015, Vol. 90, No. 4, pp. 1621-1651.

Muslu, Volkan and Radhakrishnan , Suresh and Subramanyam, K.R. and Lim, Dongkuk, 2015.Forward-Looking MD&A Disclosures and the Information Environment. Management Science 61(5):931-948.

Paul, J., 1992, On the Efficiency of Stock-Based Compensation, Review of Financial Studies 5, 471-502.

Rönnqvist, Samuel and Sarlin, Peter, 2016. Bank Networks from Text: Interrelations, Centrality and Determinants, ECB Working Paper No. 1876.

Securities and Exchange Commission (SEC), 2013. Report on Review of Disclosure Requirements in Regulation S-K. Available at: http://www.sec.gov/news/studies/2013/reg-skdisclosure- requirements-review.pdf. SEC Offices, Washington D.C.

Securities and Exchange Commission (1987) Concept Release on Management's Discussion and Analysis of Financial Condition and Results of Operations. Securities Act Release No. 33-6711. Washington, D.C.

Securities and Exchange Commission (1989) Financial Reporting Release No. 36. Management's Discussion and Analysis of Financial Condition and Results of Operations: Certain Investment Company Disclosures. Securities Act Release No. 33-6835. Washington, D.C.

Securities and Exchange Commission (2003) Financial Reporting Release No. 72. Commission Guidance Regarding Management's Discussion and Analysis for Financial Condition and Results of Operations. Securities Act Release No. 33-8350. Washington, D.C.