NARRATIVE RISK DISCLOSURES: IMPLICATIONS FOR SYSTEMATIC RISK EXPOSURES AND ASSET PRICES

John (Jianqiu) Bai, Miao Liu, Chi Wan, and Zafer Yuksel[1]

July 2024

Abstract

Utilizing unsupervised natural language processing, we assess firms' systematic risk exposures by examining the semantic relatedness of their narrative risk disclosures (NRD) in annual reports to those of other firms. Our analysis reveals a strong, positive correlation between NRD exposure and expected stock returns in a manner consistent with the Intertemporal Capital Asset Pricing Model (ICAPM). We also identify significant return predictability among firms with shared risk exposure, where the current performance of one firm forecasts the future returns of peers with similar NRD. This lead-lag relationship is more pronounced in firms receiving less investor attention and those with higher costs of arbitrage. These findings underscore that, while mandated risk factor disclosures are instrumental in facilitating risk sharing and augmenting investor welfare, there remains a critical need to enhance the clarity and accessibility of these disclosures.

Keywords: risk factor disclosure, ICAPM, stock returns, limited attention, unsupervised machine learning.

---

[1] John (Jianqiu) Bai is from the D'Amore-McKim School of Business, Northeastern University, Boston, MA, 02115 and can be reached at j.bai@northeastern.edu. Miao Liu is from the Carroll School of Management at Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467-3809 and can be reached at miao.liu@bc.edu. Chi Wan is from the College of Management, University of Massachusetts Boston, Boston, MA 02125 and can be reached at Chi.Wan@umb.edu. Zafer Yuksel is from the College of Business at the University of Rhode Island and can be reached at hzyuksel@uri.edu. We thank Mark Clatworthy (discussant), Atif Ellahie, Mac Gaulin, Xiaoxia Peng, Steve Monahan, Steven Young, workshop participants at University of Utah, WHU, Financial Reporting and Business Communication Conference, and Balyasny Asset Management (New York) for helpful feedback.

NARRATIVE RISK DISCLOSURES: IMPLICATIONS FOR SYSTEMATIC RISK EXPOSURES AND ASSET PRICES

Abstract

Utilizing unsupervised natural language processing, we assess firms' systematic risk exposures by examining the semantic relatedness of their narrative risk disclosures (NRD) in annual reports to those of other firms. Our analysis reveals a strong, positive correlation between NRD exposure and expected stock returns in a manner consistent with the Intertemporal Capital Asset Pricing Model (ICAPM). We also identify significant return predictability among firms with shared risk exposure, where the current performance of one firm forecasts the future returns of peers with similar NRD. This lead-lag relationship is more pronounced in firms receiving less investor attention and those with higher costs of arbitrage. These findings underscore that, while mandated risk factor disclosures are instrumental in facilitating risk sharing and augmenting investor welfare, there remains a critical need to enhance the clarity and accessibility of these disclosures.

Keywords: risk factor disclosure, stock returns, limited attention, unsupervised machine learning.

1. Introduction

In assessing a firm's value, investors must evaluate not only expected future cash flows but also the inherent systematic risk associated with these cash flows. While estimates of systematic risk derived from historical stock returns, such as past market betas, reveal a firm's past risk exposures relative to the market portfolio, such approach is deficient in two critical aspects. Firstly, it is backward-looking and unable to capture the firm's current and prospective exposure to systematic risks. Secondly, systematic risks are inherently complex and multifaceted, with emerging risks such as those related to climate change and pandemics either previously non-existent or not systematically evident. As a result, these emerging systematic risks are not reflected in historical stock returns or past market betas. Hence, investors stand to gain from sources that provide insights into both current and future systematic risks and the varied dimensions of these risks. Access to such information could enable investors to more accurately price in systematic risks, thereby enhancing welfare through improved risk sharing. This study investigates whether the SEC-mandated risk factor disclosures in firms' 10-K filings serve as a conduit for such critical information.

Merton's (1973) seminal work on Intertemporal Capital Asset Pricing Model (ICAPM) posits that within a multi-period economic framework, investors are motivated to hedge against future consumption and investment opportunity set risks. This suggests that capital markets should price fundamental risks correlated with future consumption and investment opportunity changes, with an asset's covariance with these risks influencing its expected returns. However, identifying concrete ICAPM fundamental risks ("state variables" in ICAPM) has remained largely conceptual, hampered by empirical challenges in pinpointing clear, interpretable risk factors. Predominantly, empirical efforts have relied on ad-hoc statistical factor models using characteristic-sorted stock portfolios as per Fama and French (2015), though this method's significant limitation is its departure from the interpretable fundamentals envisioned by Merton (1973). Alternatively, macroeconomic indicators such as industrial production, investment, and inflation have been employed as ICAPM state variable proxies (e.g., Chen, Roll, and Ross, 1986; Cochrane, 1996). Despite providing a framework closer to interpretable fundamentals, this methodology falls short of the explanatory power of ad-hoc statistical models regarding cross-sectional asset returns and risk premiums. Can firms' risk factor disclosures provide actionable information for investors to hedge systematic risks effectively while aligning with the interpretable fundamentals inherent in the ICAPM framework?

Acknowledging the unpredictability introduced by dynamic business environments on a firm's future cash flows, the Securities and Exchange Commission (SEC) mandated in 2005 that companies disclose "the most significant factors that make the company speculative or risky." This requirement holds substantial value for investors in two primary ways. Firstly, when a large number of firms disclose a specific risk within their 10-K filings simultaneously, it signals to investors the systemic nature of such risk. Secondly, it reveals the degree to which individual firms are susceptible to various systemic risks. In essence, if a firm reports a particular risk that is also disclosed by numerous others, it indicates the firm's exposure to that systemic risk. Therefore, within the ICAPM framework, these risk factor disclosures serve a dual purpose: they not only unveil the "state variables" relevant to market dynamics but also shed light on each firm's exposure to these variables. This paper aims to empirically investigate whether investors act on this disclosed information and incorporate it into stock prices. Should this be the case, risk factor disclosures create value to investors by enabling investors to hedge systemic risks more effectively, thereby promoting risk sharing and enhancing market efficiency.

This objective, however, poses some significant challenges. Firstly, risk factor disclosures are predominantly unstructured qualitative texts, complicating their direct translation into "state variables" within the ICAPM framework. Secondly, the large variety of risks disclosed by firms, often described in non-standardized language, poses a challenge for their concise representation in a parsimonious model. To address these issues, we utilize Latent Dirichlet Allocation (LDA), a topic modeling technique from natural language processing that excels in analyzing the content of extensive textual documents. It offers two key benefits: First, LDA analyzes risk factor disclosures in aggregate each year and identifies a set of "topics" discussed without requiring a pre-determined classification of topics. Second, LDA gauges the relative prevalence (or lack thereof) of a topic in a firm's risk factor disclosure, accommodating discussion of a particular topic that may be dispersed.

We employ Latent Dirichlet Allocation (LDA) to create two distinct sets of vectors representing distributions of risk topics. The first set of vectors represents the distribution of risk topics at the market-year level, aggregating the risk factor disclosures from all firms within a given year. This aggregation highlights risk topics that are widely disclosed across the market, indicating prevalent concerns among a large number of firms. The second set of vectors focuses on the individual firm-year level, tracking the evolution of each firm's risk exposure across different risk categories over time.[2] To quantify the overlap between a firm's risk disclosure and that of the broader market, we calculate the cosine similarity between their respective risk topic distributions. This metric provides a quantitative measure of the extent to which a firm's risk narrative aligns with the aggregate market risk narrative, effectively addressing the earlier outlined challenges of unstructured qualitative texts and the presence of a large variety of disclosed risks. We term this metric the firm's Narrative Risk Disclosure (NRD) exposure. Conceptually, a firm is considered to have high NRD exposure when its risk disclosures demonstrate significant overlap with risk topics commonly reported by a broad spectrum of firms, suggesting a shared vulnerability to systematic risks.

The central prediction of the ICAPM is that investors price firms' exposure to systematic risks such that firms with large exposures earn high expected returns. We initiate our analysis by evaluating if NRD exposure forecast variations in future stock returns. Utilizing a dataset of U.S. stock returns spanning from 2005 to 2018, our findings initially demonstrate that NRD exposure is a strong and positive predictor of future returns in a cross-sectional framework. When categorizing stocks into quintiles based on their NRD

[2] Please refer to Section 3 for an example of NRD topics identified in an annual report, and for an example of NRD similarity between firms.

exposures, we observe a notable average return differential between firms categorized within the lowest and highest exposure quintiles, amounting to 5.4-8.9% annually, depending on the specific asset pricing models. This differentiation cannot be solely explained by variations in loadings on established risk factors, such as those identified within the Capital Asset Pricing Model (CAPM), the Fama-French (1993) three-factor model, or the Fama-French-Carhart six-factor model. We emphasize that this disparity in returns is unlikely due to mispricing. Rather, the observed return discrepancy emerges rationally within the framework of Merton's ICAPM as a consequence of differing firm-level exposures to systematic risks.

To ensure that the correlation between NRD exposures and subsequent stock returns is not merely a reflection of firm characteristics known to relate to future returns, we run Fama-MacBeth (1973) regressions of stock returns on lagged NRD exposures alongside various firm-level attributes. Our regression incorporates conventional control variables such as a firm's market capitalization and its book-to-market ratio. Additionally, we account for recently identified predictors of stock return cross-sections that might overlap with NRD exposures, including asset growth (Cooper, Gulen, and Schill, 2008) and profitability (Novy-Marx, 2013). We also control for short-term reversal and medium-term momentum effects, as documented by Jegadeesh and Titman (1993), and Chan, Jegadeesh, and Lakonishok (1996), respectively. The outcomes of the Fama-MacBeth (1973) regressions lend further credence to the robustness of our findings from portfolio sorting, particularly highlighting that NRD exposure coefficients are consistently positive and statistically significant across all regression models. Notably, the size of these coefficients reveals the economic significance of this relationship: a one-standard-deviation increase in NRD exposures correlates with an approximate 1.94% increase in future annual returns.

Next, we develop an "NRD factor" by acquiring stocks within the highest tercile of NRD exposure and divesting those in the lowest tercile. Our examination extends to whether this theoretically grounded NRD factor can account for variations in the five "ad-hoc" statistical factors identified by Fama and French (2015), in addition to the market portfolio. Our analysis yields two notable findings. Firstly, the NRD factor exhibits negligible correlation with the market portfolio, indicating that NRD identifies exposures to present and future systematic risks that are not reflected in the covariance between firms' historical stock returns and the market. Secondly, the NRD factor significantly elucidates the variation observed in the Small Minus Big (SMB) and the Robust Minus Weak (RMW) factors, implying that the fluctuation in "ad-hoc" statistical factors may be substantially explained through our theory-grounded NRD factor.

Our analyses up to this point support the premise that our NRD exposure metric, grounded in risk factor disclosures, successfully captures the systematic risk exposures of firms and that investors take this information into account when pricing stocks. In the subsequent phase of our analysis, we explore whether investors are able to integrate such information into stock prices promptly. Two inherent characteristics of risk factor disclosures pose significant challenges for investors attempting to process this information effectively. Firstly, the unstructured nature of textual data complicates its analysis, especially when attempting to discern how a specific firm's narrative risks align with those prevalent across the market. Secondly, the lack of standardization in such disclosures, compounded by the presence of generic, redundant language, impedes investors' ability to distinguish between meaningful information and background noise. These attributes of risk factor disclosure, in conjunction with the limitations on investors' information processing capacity, likely represent a considerable impediment to the thorough and timely assimilation of the data presented in a firm's NRD. In such scenarios, we hypothesize that the returns of peer companies with similar NRD profiles may offer predictive insights into the focal firm's future stock performance. We refer to this phenomenon as the NRD momentum effect.

To examine this hypothesis, we again leverage LDA to identify shared risk exposures between a firm and its peers. Initially, we quantify the extent of similarity in risk factor disclosures between two firms through the cosine similarity of their topic distributions, a measure we designate as pairwise NRD similarity. Subsequently, for each focal firm, we compute the aggregate of its peers' returns, each weighted by the pairwise NRD similarity (this aggregate is denoted as NRDret). The construction of NRDret captures the intuition that firms with closely aligned NRD profiles are likely susceptible to the same economic, political, or geographical risks.

Our main finding is that strong return predictability exists across NRD-linked firms. That is, we document a robust lead-lag relationship between the stock returns of a focal firm and the portfolio of its peers with overlapping NRD. Focal firms whose NRD-peers earn higher (lower) returns will themselves earn higher (lower) returns in the subsequent months. A zero-cost strategy that goes long the top 20% of firms whose NRD-peers did best in the prior month while shorting 20% firms whose NRD-peers did worst in the prior month, yields an average monthly Fama-French-Carhart 6-factor alpha of 70 basis points (8.40% per annum) for equal-weighted portfolios, and 58 basis points (6.96% per annum) for value-weighted portfolios.

While our portfolio analysis shows the existence of a strong NRD momentum effect, it is important to examine whether such predictability has incremental predictive power after controlling for several known determinants of stock returns. To this end, we run Fama-MacBeth (1973) regressions, and find that even after controlling for short-term reversal (Jegadeesh and Titman, 1993), Fama-French 48 industry returns, medium term momentum (Chan, Jegadeesh, and Lakonishok, 1996), annual stock turnover (Lee and Swminathan, 2000), profitability, R&D, book-to-market, firm size, as well as asset growth, the NRD momentum effect remains robust, statistically significant and economically sizeable. Further, to alleviate the concern regarding the possibility that the overlaps in risk exposures are driven by common industry shocks, we begin by controlling for the industry momentum effect defined by conventional classifications (i.e., Fama-French 48 industries). Further, we include return momentum of the text-based industry classification (Hoberg and Phillips, 2016; 2018). Overall, these results suggest that the NRD momentum captures risks above and beyond common industry exposures.

Our next set of results examines whether the NRD momentum effect is a mere manifestation of several previously documented economic linkages that cause predictable patterns in stock returns. First, Cohen and Lou (2012) find that the returns of complex firms, i.e., firms that operate in multiple industries, lag those of their easy-to-analyze single-segment peers. To disentangle the complexity in NRD from the complexity in firm scope as the driver of our momentum effect, we re-estimate our analysis on a subsample of standalone firms (i.e. firms that operate in a single industry), which are unlikely to be subjected to the slow incorporation of industry-specific shocks compared to multi-segment firms. Our results show that *NRDret* remains a significant predictor of focal firm returns within this subsample of standalone firms. Second, NRD similarity could be due to closeness in the product market space, as opposed to overlap in multidimensional risk exposures. Following Bloom, Shankerman and Reenan (2013), we measure pairwise product market similarity based on firm sales distributions across business segments (i.e., 4-digit SIC). We find that including sales-similarity weighted peer returns in our main specification does not affect the predictive power of *NRDret*. Finally, we study whether slow information diffusion from large to small firms is a leading cause of NRD return predictability (Hou, 2007; Lo and MacKinlay, 1990; Moskowitz and Grinblatt, 1999). To account for this size-based return predictability, we exclude the largest 1/3rd firms in each industry and recalculate the NRD-similarity weighted returns. The NRD momentum remains strong in this subsample. Taken together, our results suggest that NRD similarity is distinctive from some of the previously documented economic linkages that drive return predictability.

We then conduct cross-sectional tests to shed light on the potential mechanisms of the NRD momentum effect. We focus on two specific dimensions: investors' limited attention and cost of arbitrage. The choice of investors' limited attention is motivated by our conjecture that NRD momentum is caused by investors' inability to promptly process value-relevant information embedded in complex NRD documents. Meanwhile, high arbitrage costs may also prevent investors from trading to eliminate any lead-lag NRD momentum. We conjecture that, *ceteris paribus*, firms with lower investor attention and higher arbitrage costs should exhibit a strong NRD momentum effect. Consistent with this, we find that NRD momentum is more pronounced for firms presumably with higher investor inattention and higher costs of arbitrage[3]. This suggests that the NRD momentum is most likely attributable to firms' complex disclosure of risks, which creates a high hurdle for information processing.

Our paper contributes to several strands of literature. First, we extend the literature on the informativeness of narrative risk disclosures. To the best of our knowledge, our study is the first to document the usefulness of risk factor disclosures for investors to detect systematic risks, which may be complex, nascent, and rapidly evolving, and incorporate information on firms' exposure to such systematic risks into prices. This stands in contrast to previous research, which primarily examines the impact of specific attributes of a firm's risk factor disclosures on its valuation and the reactions of investors and analysts (e.g., Campbell et al., 2014; Bao and Datta, 2014; Hope et al., 2016). Diverging from these studies, our research adopts a general equilibrium perspective within the framework of ICAPM, concentrating on the influence of focal firm's shared systematic risk exposures with **all** other firms on the focal firm's value. Our findings indicate that narrative risk disclosures can significantly enhance market efficiency and promote risk sharing across the entire investor spectrum.

Our study also enhances the extensive literature on investors with limited attention, such as the seminal works of Merton (1987), Hong and Stein (1999), Hirshleifer and Teoh (2003), and Peng and Xiong (2006). These models suggest that the delayed processing of information, attributable to investors' limited attention spans, can lead to predictable return patterns. This theoretical foundation has spurred a burgeoning number of empirical investigations, particularly those recent studies highlighting the lead-lag relationship in returns among firms with close economic connections of various types.[4] Within this context, our research offers new insights by demonstrating that investors exhibit an underreaction to information emanating from firms linked through common risk exposures, an economic connection with deep theoretical underpinnings.

[3] Higher investor inattention is measured with lower Google search intensity, lower institutional ownership, or fewer analyst followings, while high arbitrage costs are assessed by stock illiquidity. We discuss the construction of these measures in detail in Section 4.

[4] See, for example, Cohen and Frazzini (2008); Cohen and Lou (2012); Hou et al. (2012); Aobdia et al. (2014); Huang (2015); Lee et al. (2015); Lee et al. (2019); Parsons et al. (2020); and Ali and Hirshleifer (2020).

Lastly, our framework is automated to allow dynamic identification of risk topics. Besides the number of topics, the procedure requires no additional human input to detect the overlap in firms' NRD. The full set of risk topics is also updated annually. In contrast, several studies estimate static topic models using the full sample and rely on manual interpretation to classify and group topic keywords (e.g., Lopez-Lira, 2020; Ross, 2019). However, certain risks embedded in NRD are difficult, if not impossible, to identify ex ante (Hanley and Hoberg, 2019). Hence, our method provides a capable alternative to detect and capture the relatedness of firms' ever-changing risk environments.

Our research highlights important policy implications, emphasizing the value of mandated risk factor disclosures. These disclosures not only inform investors about a firm's specific risk exposures but also reveal systematic risks shared across many firms, enhancing the efficiency of pricing systematic risks. This underscores the necessity for mandated disclosures due to the benefits they extend beyond individual firms to the broader investment community. Furthermore, the complex nature and processing demands of these documents delay their influence on price discovery. Our findings are timely, considering the SEC's recent amendments to improve the readability and relevance of risk factor disclosures. They suggest that while such disclosures aid in risk sharing and investor welfare, there is a clear need for making this information more accessible and comprehensible.

The rest of this paper is organized as follows: Section 2 briefly discusses the background of narrative risk disclosure and LDA as a topic modelling tool. Section 3 describes the data and empirical methodology. Section 4 reports the empirical results. We conduct supplementary and robustness tests in Section 5. Section 6 concludes.

2. Background

2.1 Narrative Risk Disclosures (NRD)

Transparent disclosure of a firm's risk environment in annual reports serves as an important venue of communication between a publicly traded company and its investor base. It provides comprehensive information to the investing public about uncertainties associated with the firm's future cash flows. However, despite the importance of this information to investors and regulators alike, SEC did not enforce such disclosure prior to 2005 (Brown et al., 2018), with registration statements for equity and debt offerings being the only exceptions. In 2005, through Item 305 of Regulation S-K (the Market Risk Rule), SEC mandated both qualitative and quantitative disclosure within the financial statements about a registrant's exposures to market risks, to the extent that those exposures are material.

Specifically, the SEC mandates that firms include discussions related to risks in Item 1A of their Annual Report on Form 10-K. However, given the discretion firms have over such disclosure, Item 1A is often subject to criticism: firms frequently copy and paste previous disclosures using boilerplate language and fail to capture the underlying risk environment (Brown et al., 2018; Cohen et al., 2020); alternatively, Schrand and Elliot (1998) also report either the lack of relevant risk disclosure or ambiguity in these descriptions. In a follow-up in 2010[5], the SEC therefore warned firms against generic risk disclosure that is uninformative about firm-specific operations. In fact, enhancing the informativeness of narrative risk disclosure has become one of the central focuses of the SEC comment letter process and the latest SEC regulation S-K rules.

This mandated narrative risk disclosure has led to several recent studies that examine both the content and format of Item 1A in firms' annual reports. For instance, Campbell et al. (2014) analyze the contents of Item 1A to determine if they reflect useful information for capital market participants, and find that firms that are subject to riskier environments tend to disclose more risk factors and managers provide pertinent firm-specific risks in this section. Additionally, Brown et al. (2018) report that a focal firm significantly alters its own risk disclosures following SEC issuance of comment letters to its peers questioning the informativeness of their Item 1A. Campbell, Cecchini, Cianci, Ehiger and Werner (2019)

[5] Form 10-K instructions, found at https://www.sec.gov/about/forms/form10-k.pdf.

report that investors are aware of the correlation between tax-related risk factor disclosures and the firm's future cash flow levels. Chiu, Guan and Kim (2017) find that credit market participants incorporate risk factors disclosures in the pricing of credit default swap spreads.

2.2 Topic Modeling and LDA

Analyzing the content in firms' NRD poses a significant empirical challenge. Firms operate in distinct industries, have unique products and services, face different clienteles, and likely operate in different geographic areas. As a result, a systematic analysis of the content included in different firms' NRD seems to be a daunting task. To circumvent this issue, we employ the latent Dirichlet allocation (LDA), a topic modelling technique in the field of natural language processing. An unsupervised Bayesian linguistic tool, LDA is conceptually similar to factor analysis but is applicable to text (Blei, Ng, and Jordan, 2003). It can be viewed as a dimension reduction tool that extracts topics present in a collection of documents, and in each document infers the proportion of discussion dedicated to each topic. Since LDA does not require researchers' intervention for topic discovery, it is well suited for identifying key risk exposures from a firm's Item 1A.

Recent literature has employed LDA on multiple firm disclosures. Hanley and Hoberg (2019) employ LDA on Item 1A, and investigate the trends in risk factor disclosures by banks to detect buildup of latent risks that could lead to systematic failures. While quantitative disclosures were not indicative of emerging risks in the financial industry during the early 2000s, they find that banks' narrative disclosures closely reflected the volatility that firms were exposed to. Lopez-Lira (2020) applies LDA to ascertain the top 25 risks that firms recognize. He then manually labels each risk related topic to develop a factor model to explain focal firm stock returns. Since the risk-related topics remain constant in his sample, the model only captures firm exposure to several forms of well-defined risks (e.g., technological and innovation risk, production risk, international risk, and demand risk), and ignores information and evolution of other risk disclosures.[6] In comparison, we automate the topic modelling process and explore a focal firm's return predictability stemming from peer firms of overlapping risk factor disclosures.

In addition, Campbell et al (2014) use a joint methodology of key words identification and LDA to assess the usefulness of Item 1A to investors. They report consistency between firm's exposure to risks and their level of disclosure, affirming the usefulness of these narrative disclosures to the capital market. Bao and Datta (2014) apply a variant of LDA to assess the importance of risk disclosures to investors. They report that investors' risk perception is increased by disclosure related to systematic and liquidity risks while unsystematic risk disclosures decrease their risk perception.[7]

Overall, LDA is particularly suited to the task of analyzing the textual content of financial reporting. Prior literature largely follows the path of ascertaining the relevance of narrative risk disclosures, their relevance to predicting focal firm returns and market participants' acknowledgement of this relationship. We exploit the versatility and suitability of LDA to investigate collective exposure in risk environments across firms and the asset pricing implications of such overlaps.

3. Data and Key Variables

This section outlines the implementation of LDA to extract topics from Item 1A and the use of cosine similarity to gauge semantic overlaps between firms' risk disclosures. Then, for a focal firm, we construct two NRD based measures. The first measure captures the focal firm's NRD similarity to *all* firms on the market, which we term *NRD exposure*. The second measure is NRD-similarity weighted returns of the focal firm's peers to capture information embedded in returns of firms with similar NRD, which we term *NRDret*.

---

[6] Related to Lopez-Lira (2020), Ross (2019) is also interested in the relationship between risk factor disclosure and focal firm stock returns. He extracts 50 risk-related topics in Item 1A, finding that only 19 of those consistently provide pertinent information regarding stock returns.

[7] LDA has also been applied to analyze other financial documents. For example, Huang, Lehavy, Zang, and Zheng (2018) analyze conference call transcripts and analyst reports using LDA to investigate the financial intermediary role of analysts in capital markets; Hoberg and Lewis (2017) and Ball, Hoberg and Maksimovic (2015) apply this technique to Management Disclosure and Analysis (MD&A), while Dyer et al. (2017) apply LDA to the entire 10-K to analyze trends in annual reports.

3.1 Implementation of LDA

Item 1A (i.e., NRD) is a narrative disclosure where a firm provides a comprehensive review of various risks pertinent to its operations. We employ LDA to extract the semantic structure of a firm's NRD. Essentially, LDA reduces the dimensionality of each document from thousands of words to a distribution of topics; each topic is then mapped to a cluster of keywords. We calibrate LDA to identify a large spectrum of risks that captures firms' risk profiles collectively. We accomplish this by 1) focusing on bigrams (i.e., two adjacent words), the meaning of which is often less ambiguous than that of individual words (i.e., unigrams); 2) assigning the number of NRD topics conservatively; and 3) annually updating topics to gauge the dynamic evolution of a firm's risk profile.

3.1.1. Use of Bigrams to Identify NRD Topics

Conventional unigram LDA is a "bag of words model" that ignores the order of words in a given corpus. To better discern the topics discussed in the text, we extract topics based on bigrams (pairs of adjacent words) rather than unigrams. To illustrate the usefulness of bigrams in extracting topics, consider "capital expenditures". Taken separately, the context of "capital" and "expenditures" can generate spurious topics, especially in the financial context. While some words may suffice individually (such as "bankruptcy"), combining two words would generate clearer and more coherent topics.

3.1.2. Selecting the Number of Topics

The key manual input in the implementation of LDA is the total number of topics in the corpus, which depends on the researchers' objective: for example, Lopez-Lira (2020) chooses 25 in his analysis with the goal of manually identifying several key topics that systematically affect many firms. Dyer et al (2017) suggest that the corpus of 10-Ks, including Item 1A and all other sections, can be classified into 150 topics.

Formally, we employ the perplexity score to guide us in determining the total number of topics. Given a pre-specified number of topics, perplexity diagnoses the performance of LDA estimated using the training data (a subset of the documents) to predict the topic mixtures of the remaining documents.[8] As an example, Figure 1 presents the perplexity score of bigram LDA with the number of topics varying from 25 to 150 estimated using all firms' NRD in 2009 and 2013 (Panels A and B, respectively). The perplexity score decreases as the number of topics increases, indicating better generalization of topics obtained from the training data to the testing data. However, as the number of topics increases further, the improvement in model fit diminishes, which is often at the expense of loss of topic interpretability (Chang et al, 2009; Dyer et al, 2017). We therefore choose 100 topics (often referred as the "elbow" point as the rate of perplexity change begins leveling off) to implement our baseline analysis. This choice assumes that there are 100 pertinent risk-related topics firms discuss annually in the aggregate, and is also consistent with our goal to accommodate the spectrum of topics in the collection of NRD that may span various firm-specific, geographic, or technological risks. As shown in Section 4.6.1, our results stay qualitatively unchanged if we vary the number of topics as an input to the LDA from 25 to 150.

[Figure 1]

3.1.3. Annual Update of NRD Topics and Firm-level Topic Distributions

SEC mandates firms to revise their Item 1A on an annual basis to provide up-to-date risk-related information to capital market participants. Hence, to ensure that our extracted risk topics capture a firm's dynamically evolving risk profile, we separately estimate bigram LDA on all firms' Item 1A in each year, and thus, in effect, update the topics annually to track the dynamic risk environment that a firm faces. Therefore, the topics extracted by LDA from Item 1A reflect the risks faced by the firms for that year.

[8] Following Dyer et al. (2017), we train the model on 90% of the data and use a random hold-out sample of 10% as the testing data to calculate perplexity.

This procedure produces two outputs for each individual firm annually: the first data set (i.e., topics) produces a set of bigrams and the associated frequencies that make up each topic. The second data set (i.e., topic loadings) describes the distribution of topics discussed in a firm's NRD. In essence, topic loadings indicate the relative importance of each topic in a firm's NRD. When the loading on a topic is closer to one, it reflects the high relevance of that particular risk for a firm. Conversely, a smaller topic loading implies low prevalence of the corresponding topic discussed by the firm in its Item 1A.

3.1.4. An Illustrative Example of NRD Topics

To provide more information on the extracted risk topics using bigram LDA, in Figure 2, we graph the word clouds that visualize four topics (out of 100). The figure contains the set of associated key words that are identified from the collection of NRDs in 2016, where the font size of bigrams represents the relative frequency and thus the relevance of these to the topics.

[Figure 2]

Word Cloud 1 is related to discussion about exchange rates and foreign currencies. Given that nearly half of total sales of S&P 500 firms are generated in foreign markets,[9] it is not surprising that this risk affects a large number of firms. Word Cloud 2 is an example of topics that concern only a small number of firms that, in this particular case, specialize in manufacturing of medical devices. In Word Cloud 3, the topic – climate change risk – would be difficult to quantify using firm fundamentals but is well captured by our methodology.

Similar to topics extracted from entire 10-K filings (Dyer et al, 2017), some topics are difficult to interpret; for example, Word Cloud 4 depicts a topic of which the mix of bigrams seems to have no clear unifying theme. As we have used the perplexity score in guiding us for setting the total number of topics to a relatively large number (i.e., 100), LDA might yield "noise", i.e., a topic that does not correspond to a

[9] See, for instance, https://www.prnewswire.com/news-releases/sp-500-foreign-sales-for-2017-total-43-6-300698039.html.

clear textual theme. There are two possible implications of this: as highlighted by Hanley and Hoberg (2019), such topics can potentially reflect emerging risk exposures that have not fully materialized and thus go beyond researchers' information set. Second, such noise could potentially lead to spurious NRD relatedness across firms, and therefore should not drive risk premium between a focal firm and the market or return predictability between a focal firm and its NRD-peers.

Our method, therefore, aims to identify various risks that collectively span the space of corporate risk environment. At the firm-year level, the output from LDA delineates the risk environment faced by a firm and its evolution over time.

3.1.5. T_i: firm-level distribution of risk topics

For each year, LDA identifies 100 topics that best represent the distribution of topics that firms have discussed in their Item 1A. To identify the relevant risk exposure of a particular firm in a given year, we extract the probabilistic distribution of those topics present in the firm's Item 1A. As we are only concerned with mapping the possible risks a firm is exposed to and their relevance to the firm, our analysis does not require economic interpretation or further categorization of keywords extracted.

Specifically, the distribution of topics in firm i's NRD is presented as $T_i = (T_{i,1}, T_{i,2}, \dots, T_{i,k}, \dots, T_{i,100})'$, where its k^{th} element, $T_{i,k}$ gauges the relative importance of topic k in the firm's risk factor disclosure in a particular year. Elements of T_i sum up to one. Both the value of T_i and the corresponding topic are updated annually, capturing the dynamic evolution of firms' risk environments. The year subscript is omitted for brevity.

3.2 NRD-based Market Risk Exposure

We propose a novel measure of market risk exposure that provides a holistic assessment of a firm's risk exposure in the relation to the market. In a year, we calculate the value-weighted firm-level distributions of NRD topics (T_i) and denote this vector as T_M, where the year subscript is omitted for brevity. T_M is the distribution of NRD topics in the market portfolio; and measures the time-varying importance of each topic in the risk spectrum of the market portfolio.

Then, we compare the similarity of firm *i's* distribution of NRD topics with that of the market portfolio by calculating the dot product of T_i and T_M (i.e., $T_i T_M{}' / \sqrt{T_i T_i{}' \times T_M T_M{}'}$, denoted as ω_{iM}). ω_{iM} evaluates the similarity between firm *i* and the market regarding exposures to key risk-related topics collectively identified in firms' NRD. The small value of ω_{iM} suggests a low degree of relatedness between the firm's risk profile and the common risk exposure across all firms. In contrast, ω_{iM} increases as firm *i* NRD becomes more aligned with common NRD topics in the market portfolio.

3.3 NRD-linked Stock Returns

We then turn to gauge the degree of semantic overlap between firms' NRD. The relatedness between firm i and j's NRD is appraised by the dot product of T_i and T_j and is computed as $T_i T_j{}' \Big/ \sqrt{T_i T_i{}' \times T_j T_j{}'}$ (denoted as ω_{ij}). ω_{ij}, also referred as the "cosine" similarity between T_i and T_j, is bounded from 0 to 1 with a close-to-unity value indicative of the presence of great overlap among the two firms' NRDs.

For firm *i*, the NRD-linked returns (*NRDret*) is the NRD-similarity (ω) weighted sum of all other firms' monthly returns. Specifically, $NRDret$ is defined as follows.

$$NRDret_{i,t} \equiv \sum_{j \neq i} \omega_{ij} Ret_{j,t} \Big/ \sum_{j \neq i} \omega_{ij}, \tag{1}$$

where $Ret_{j,t}$ is the return of firm j in month t and the summation is across all firms in that year. Intuitively, $NRDRet$ captures the information content embedded in the returns of all other firms in that year.

Our baseline analyses employ bigram LDA with 100 topics to characterize a firm's risk profile and to subsequently assess semantic similarity between firms' risk factor disclosures. In Section 4.6, we conduct several robustness tests to show that our main findings are robust to 1) different choices of the number of topics varying from 25 to 150 when implementing LDA and 2) the use of alternative topic model, Latent Semantic Analysis (LSA).

3.3.1 An Example: Gauging NRD Similarity between Firms

LDA allows us to identify firm relatedness that may otherwise go unrecognized. Consider two firms Fresh Del Monte (GVKEY=30443, SIC=0100), a fresh fruit company, and General Moly Inc. (GVKEY=122212, SIC=1000), a mining company that specializes in advanced stage mineral deposits. The two firms belong to different (SIC and TNIC) industries and have no supply-chain relationship.

Despite the apparent lack of relatedness, the two firms' NRD similarity (i.e., ω) is 0.54 in 2016. Upon closer inspection of their topic distributions, we find that the sizable NRD similarity is driven by similar disclosures that focus on common topics related with capital expenditures and future acquisitions. In addition, these two firms' NRD significantly overlap in topics about environmental risks. For example, in 2016, General Moly stated in its Item 1A that legislative bodies are implementing changes in response to climate change, which would impose new compliance costs for the mining company. Meanwhile, Fresh Del Monte reports that adverse weather conditions and climate change can greatly impact its major fruits suppliers. The fruit company further notes that it faces potential cost increase as regulatory bodies are considering tightening the use of fertilizers due to climate concerns. Therefore, while it is difficult to quantify a firm's ex-ante exposure to environmental risks, LDA allows us to construct a firm's comprehensive yet nuanced risk exposures and to capture the degree of contextual overlap between firms' risk factor disclosures with minimal human intervention. Additionally, by independently running LDA annually on firms' NRD, we are able to capture the evolution of a firm's risk exposure without either restricting the model to focus on risk factors that may lose relevance over time or new factors that are unforeseen.

3.4 Sample Construction

To make our final analytical sample, we require firms to have non-missing book equity, market equity and SIC classification at the end of the previous fiscal year for inclusion in the sample. We also restrict the sample to firms listed on NYSE/Amex/Nasdaq and with a share codes 10 or 11. We exclude financial firms (SIC 6000-6999) and firms with stock prices lower than $1.

Our sample of firm-level NRD begins in 2005 and ends in 2018. To ensure that accounting information is publicly available to capital market participants, we impose a minimum six-month gap between a firm's fiscal year end and monthly stock returns. Specifically, in the calculation of NRD-linked returns (defined in Equation 1), NRD similarity in year t is matched with peers' stock returns from July year *t+1* to June year *t+2*. Our final analytical sample has 316,004 firm-month observations and spans the period between July 2006 and December 2018. All continuous variables are winsorized at the 1% level to minimize the impact of outliers.

[Insert Table 1 Here]

Summary statistics are presented in Table 1. *NRD_EXP* is approximately normally distributed with a standard deviation close to 1. Average monthly $NRDret$ equals 1.08%, which is approximately as volatile as industry return ($FF48\ Ret$) but fluctuates less than the firm's own past one-month return (Ret_{t-1}).

4. Empirical Results

Our first main hypothesis posits that Narrative Risk Disclosures (NRD) enable investors to accurately assess systematic risks in alignment with the ICAPM, serving a dual purpose: they offer detailed insights into individual firm's risk exposures and also shed light on common risks across multiple firms, thereby pinpointing which risks are systematic within the ICAPM framework. We present evidence supporting this hypothesis in Section 4.1. Following this, Section 4.2 elaborates on our second hypothesis, which suggests that the inherent complexity of firms' NRDs may hinder investors' ability to fully and promptly assimilate this information. Consequently, analyzing returns of firms with comparable NRDs could provide valuable insights for predicting the future stock performance of a specific firm.

4.1. NRD Exposure and Systematic Risk Premium

4.1.1. Portfolio Analysis

At the beginning of each July, we rank stocks into deciles by their NRD exposure (*NRD_EXP*). We allow a six-month gap between a firm's fiscal year end and monthly stock returns to allow information on NRD to become publicly available and hold the resulting 5 portfolios for a year. Table 2 presents average firm characteristics for the resulting quintile portfolios. Average values of *NRD_EXP* range from 4.06 for the bottom decile to 6.96 for the top decile.

[Insert Table 2 Here]

For each *NRD_EXP* portfolio, we report group-averaged CAPM Beta (β^{M}). While betas of individual stocks are estimated with noise, the noise is smaller at the portfolio level. Table 2 shows that *NRD_EXP* has a negative correlation with β^{M}, indicating that NRD identifies exposures to present and future systematic risks that are not reflected in the covariance between firms' historical stock returns and the market. High *NRD_EXP* firms are larger, more profitable, and have lower investment rate (IK). No strong relation emerges between *NRD_EXP* and other considered characteristics: book-to-market ratio (B/M), momentum, asset growth rate, and leverage (LEV).

For each quintile portfolio, we obtain monthly time series of returns from June 2006 to December 2018. Table3 summarizes excess returns and alphas for each quintile and for the portfolio that is long the quintile with high *NRD_EXP* and short the quintile with low *NRD_EXP*. To control for differences in risk across quintiles, we present alphas from the CAPM, Fama and French (1993) three-factor model, Carhart (1997) four-factor model, the Fama and French (2015) five-factor model, and the Fama and French (2015) five-factor model plus the momentum factor of Carhart (1997).

[Insert Table 3 Here]

Both raw and risk-adjusted returns of the 5 portfolios indicate a strong positive relation between *NRD_EXP* and future stock performance. Firms in the high- *NRD_EXP* quintile earn the highest average return, whereas the low-*NRD_EXP* quintile performs most poorly. The difference in the performance of the two quintiles, at 0.64%, is economically large and statistically significant (t-statistic of 2.79). The corresponding differences in alphas are similarly striking, ranging from 0.45% (t-statistic of 2.15) for value-weighted Fama-French five-factor alphas to 0.74% (t-statistic of 3.44) for equal-weighted Fama-French three-factor alphas. Results of portfolio sorts thus strongly suggest that NRD exposures are an important cross-sectional predictor of returns.

4.1.2. Fama-MacBeth Regressions

The empirical evidence from portfolio sorts provides a strong indication of a positive relation between *NRD_EXP* and subsequent equity returns. However, such univariate analysis does not account for other firm-level characteristics previously shown to relate to future returns. We now compare *NRD_EXP* with other well-established determinants of the cross-section of stock returns. Our goal is to evaluate whether the ability of *NRD_EXP* to forecast returns is subsumed by other firm-level characteristics. To this end, we run monthly Fama-MacBeth (1973) regressions of monthly stock excess returns on *NRD_EXP* and on control variables.

We include as controls commonly considered firm characteristics such as log market capitalization (ME) and log book-to-market (BM). We also control for short-term reversal and medium-term momentum effects, as documented by Jegadeesh and Titman (1993), and Chan, Jegadeesh, and Lakonishok (1996), respectively, asset growth (Cooper, Gulen, and Schill, 2008) and profitability (Novy-Marx, 2013). The timing of the variables' measurement follows the widely accepted convention of Fama and French (1992). In some specifications, we include Fama-French 48 industry fixed effects to ensure that we are not capturing industry-level risk premium.

[Insert Table 4 Here]

Table 4 summarizes the results of the Fama-MacBeth (1973) regressions. The coefficient on *NRD_EXP* is positive and statistically significant in each considered specification, even after accounting for other predictors of the cross-section of equity returns. The magnitude of the coefficient implies that, for a one-standard-deviation increase in *NRD_EXP* (0.98), subsequent monthly returns increase by at least 0.162%, and by 1.94% annually.

4.1.3. Spanning Analysis

An advantage of our methodology is its interpretability: our systematic risk exposure, inspired by theoretical considerations, is directly derived from firms' disclosures about their risk exposures. Hence, *NRD_EXP* can be theoretically understood as exposure to "state variables" within the framework of the ICAPM. In this section, we construct an "NRD factor" by acquiring stocks in the highest tercile of NRD exposure and divesting those in the lowest tercile. We then explore whether this NRD factor, which is rooted in theory, can explain variations in the five "ad-hoc" statistical factors in Fama and French (2015), including the market portfolio.

[Insert Table 5 Here]

In Table 5, we conduct regressions of each of the Fama-French five factors and our NRD factor against the remaining factors, with each row representing a separate regression. This analysis highlights two significant insights. First, the NRD factor demonstrates minimal correlation with the market portfolio. This suggests that NRD captures exposures to both current and forthcoming systematic risks not captured by the covariance of firms' historical stock returns with the market. Second, the NRD factor provides a significant explanation for the variations in the Small Minus Big (SMB) and Robust Minus Weak (RMW) factors. This suggests that the variations in "ad-hoc" statistical factors can be meaningfully attributed to our theoretically motivated NRD factor. These findings align with the univariate correlations observed in Table 2 between NRD_EXP and factors such as size and profitability.

4.2. NRD Exposure and Lead-Lag Effects

We transition to our second hypothesis, which posits that the complexity within firms' NRDs may obstruct investors' capacity to completely and quickly process this information. As a result, examining the returns of firms with similar NRDs may offer insights for forecasting the future stock performance of a particular firm.

4.2.1. Portfolio Analysis

We begin by employing a portfolio strategy that exploits the NRD momentum to gauge the potential profitability due to firms' overlapping risk exposures. We first sort firms into two size groups: small and large firms in a month. Then the top NRDret portfolio combines 20% of the small and 20% of the large firms, both of which have the highest NRDret in the previous month. Similarly, the bottom NRDret portfolio contains 20% firms from each size group that have lowest NRDret. This procedure alleviates the leading confounding effect of firm size in return predictability (Hirshleifer, Hsu, and Li, 2013). All stocks are equal (value) weighted within the given portfolio and rebalanced monthly. Our zero-cost strategy takes a long position in the top 20% firms whose NRD-linked peers performed best in the prior month and goes short the bottom 20% of firms with the worst NRD-linked peer performance in the same period.

[Insert Table 6 Here]

The results of this portfolio strategy are presented in Table 6. Panel A represents the returns to an equal weighted strategy, with Panel B reporting the results of value-weighted portfolios. The return difference between two extreme portfolios shown in Column (1) is the excess return from our strategy: the equal weighted portfolio generates 70 basis points monthly (t-statistic=2.96), or 8.40% returns annually, with the value weighted portfolio generating 58 basis points monthly (t-statistic=1.79) or 6.96% returns annually. We further control for other risk exposures as in Table 3 and find that the results remain largely unchanged: abnormal returns generated by NRD momentum persist. To conclude, investors' inability to timely incorporate overlapping NRD generates economically significant mispricing that cannot be explained away by common risk factors.

4.2.2. Cross-sectional Regressions

We now turn our attention to the regression framework to assess the incremental effect of NRD-related returns. Specifically, we employ Fama-MacBeth (1973) regressions where the dependent variable is the focal firm return in month *t* (Ret_t) and the independent variable of interest is the NRD-linked firms' weighted average returns in the previous month ($NRDret_{t-1}$).

Following prior literature on lead lag effects, we control for several previously documented determinants of stock returns: first, we control for the focal firm's lagged returns (Ret_{t-1}) to account for the short-term reversal effect documented by Jegadeesh and Titman (1993). Second, we include value-weighted industry returns based on Fama French 48 industries in the prior month ($IndRet_{t-1}$). Third, following Chan, Jegadeesh, and Lakonishok (1996), we control for medium term momentum (MOM) which equals the focal firm stock returns from *t-6* to *t-2* and annual stock turnover (*Turnover*) following Lee and Swaminathan (2000). Finally, we also incorporate a concentration index of risk exposures (HHI_NRD_i), which is calculated by squaring the proportion of each firm's risk exposure in the vector depicting its risk disclosure (i.e. *NRD* defined in Section 3) and then summing the resulting numbers. In addition, our regressions include other common determinants of stock returns documented in literature: gross profitability, R&D, book-to-market, firm size, and asset growth.

[Insert Table 7 Here]

The results are presented in Table 7. Consistent with our conjecture, $NRDret$ is a significant predictor of focal firm stock returns. Even after controlling for other determinants, average monthly return increases by 2.04% when $NRDret$ moves from the median to the third quartile.[10, 11]

[10] This is calculated as the coefficient from the regression multiplied by the difference between the median and 75th percentile: 0.748 × (4.31 - 1.58) = 2.04%.

[11] As an alternative measure of the cosine similarity between topic distributions (Section 3.2), we use the Hellinger distance to quantify the proximity between firms' topic distributions in their NRD. In particular, the Hellinger distance between firm *i* and *j*'s NRD topic distributions is defined as $\frac{1}{\sqrt{2}}\left[\sum_{k=1}^{100}\left(\sqrt{T_{i,k}} - \sqrt{T_{j,k}}\right)^2\right]^{1/2}$, where *T* is the topic distribution in NRD as detailed in Section 3.2. The Hellinger distance between two discrete distributions (our case) is directly related to the Euclidean norm of the difference of the square root vectors of T_i and T_j. For more details about the Hellinger distance, please refer to Bogachev (2007). Our results

While our baseline results include value-weighted industry returns based on Fama French 48 industries, it is possible that the documented momentum effect is a rediscovery of the lead-lag effect generated from slow investor processing of other narrative disclosures in 10-K. Specifically, standard industry classifications are based on similarity of processes of production and not updated to reflect changes in product offerings. To overcome this, Hoberg and Phillips (2016) use product descriptions provided by firms in Section 1 of their 10-K to reclassify firms annually based on their outputs. They generate a text-based industry classifications (TNIC) to reclassify firms annually based on their product offerings. In a subsequent paper (Hoberg and Phillips, 2019), they find that peers in the same TNIC industry generate return predictability of a focal firm.

To ensure that relatedness in risk environments captured by *NRDret* is distinguishable from information derived from TNIC peers, in Table 7 Column 2, we further include the TNIC momentum as a control variable. Specifically, this is calculated as the equal weighted average of TNIC-2 industry peers' returns and lagged by a month. In line with the findings of Hoberg and Phillips (2019), the TNIC momentum is a significant predictor of focal firm returns and susbumes some explanatory power of *NRDret*. Our results therefore indicate that *NRDret* provides valuation relevant information to investors that is distinct from industry mementum constructed based on either standardized or text-based industry classifications.

In Table 7 column 3, we follow Lee, Sun, Wang, and Zhang (2018) and investigate the average monthly spread generated from the NRD momentum by assigning $NRDret_{t-1}$ to deciles ranging from 0 to 1. The statistically significant coefficient on $NRDret\ decile$ (0.775, t-statistics=4.54) corresponds to an average monthly spread of 77.5 basis points between the top and bottom decile of NRD-linked firms, translating to a 9.3% annual return.

4.2.3. Controlling for Several Other Economic Linkages

continue to hold when using this alternative similarity measure to calculate NRD similarity weighted peer firms' returns (Equation 1).

Our regression results suggest that information content of firms' NRD, while economically significant, is not timely incorporated into focal firm returns. This delay in processing relevant risk information results in return predictability of focal firms from NRD-linked firms. In this section, we conduct additional tests to assess other possible explanations of this sluggish price adjustment and control for them to substantiate the distinct nature of the NRD momentum effect.

4.2.3.1 Complicated Firms

While we posit that the difficulty in interpreting and identifying overlapping risk exposures drive the return predictability we document, Cohen and Lou (2012) find that structural complexity of firms, i.e., sales to multiple industries, also leads to slower incorporation of firm-relevant news into stock prices. As stand-alone firms (firms with sales to one industry) require relatively straightforward information processing, they find that these firms can predict the returns of conglomerates, i.e. firms that have sales to multiple industries and are therefore more complicated to price. Although we conjecture that information processing complexity of risk factor disclosures drives the NRD momentum, our results may be a manifestation of complex firm structures, rather than difficulty in extracting and interpreting overlaps across Item 1A. We therefore restrict our sample to stand-alone firms and recalculate NRD-weighted returns on the truncated sample. The results are presented in Table 8 Column 1. Our findings are consistent with our conjecture: we find *NRDret* to be a powerful indicator of focal firms' returns even in the easier-to-process sample, reaffirming the distinctiveness of narrative risk disclosure complexity from structural complexity in product market.

[Insert Table 8 Here]

4.2.3.2 Overlapping Segment Sales

Intuitively, competition in the product market space, through closeness in final product and services offered by two firms, can result in their collective exposure to certain risks. For example, consider Fresh Del Monte (GVKEY=30443) and Limoneira (GVKEY=29962). Both these firms provide fresh fruits and are in the same SIC2-digit industry. Their cosine similarity has an average of about 0.80 over 2010 to 2017. Therefore, it is possible that the NRD momentum we document is a manifestation of proximity between two firms in product market space.

To assess product market proximity, we follow Bloom et al. (2013) and use sales distribution across SIC 4-digit industries (i.e., segments) to calculate firm-level pairwise sales similarity. Following the construction of *NRDret*, we generate *SaleRet*, which is the segment sales similarity weighted peers' return. In our next regression, we incorporate both lagged *SaleRet* and *NRDret*, thereby controlling for proximity in product market space. Results are presented in Table 8 Column 2. Considering that proximity in product market offerings can generate common risk exposure across firms, *SaleRet* partly subsumes the explanatory power of the NRD momentum. However, given the statistically significant coefficient of $NRDret$, our results indicate that overlap in NRD across firms provides incremental information relative to product market proximity.

4.2.3.3. Size Based Information Diffusion

Lo and MacKinlay (1990) document size-dependent diffusion in capital markets, such that larger firms incorporate relevant information in their stock price faster than relatively smaller firms. To ensure that size-based information processing does not explain the sluggish incorporation of NRD-related information, we drop the $1/3^{rd}$ largest firms in every year and recalculate NRD-similarity weighted returns based on the limited sample. Results for the Fama-MacBeth regression are presented in Table 8 Column 3. The coefficient on $NRDRet$ remains positive and statistically significant in this sample of relatively smaller firms. This result indicates that, while information often diffuses slowly from large to small firms, this phenomenon cannot explain away the return predictability generated by investors overlooking firms' common exposure to certain risks identified in their NRD.

4.2.4. Cross-sectional Tests

In the previous section, we find lead-lag effects among firms with shared NRDs. Given that NRD is inherently complex, we contend that investors' information processing constraint is an underlying driver of the documented NRD momentum effect.

4.2.4.1 Investor Inattention

First, we hypothesize that, if investors are more likely to overlook common themes in a firm and its peers' risk factor disclosures, the predictive power of peers' NRD similarity weighted return would be stronger. To this end, we employ several well-known measures of investor attention to gauge the impact of this phenomenon on NRD. We begin our investigation with the basic measure, *Market Cap*, which is a dummy variable that equals 1 for firms whose market capitalization is higher than the median of all firms in the year, and zero otherwise. Next, given the diverging level of resources and incentives of investors (Schnatterly, Shaw, and Jennings, 2008), we employ *SVI* (Google search volume index)[12] and *Institutional Holdings* to measure individual and institutional investor attention, respectively. The original *SVI* is an index ranging from 0 to 100 that captures the popularity of a phrase relative to total Google search queries submitted during a period, with higher values corresponding to higher relative searches. In our regressions, both *SVI* and *Institutional Holdings* are dummy variables that equal 1 for firms with higher than the median of all firms for the year, for the respective measures, and zero otherwise.

Financial analysts are important information intermediaries in capital markets. While studies are inconclusive on whether their discovery or interpretation role is superior (see Huang, Lehavy, Zang, and Zheng, 2018), literature documents that returns of firms with greater analyst following reflect pertinent information relatively faster (Brennan, Jegadeesh, and Swaminathan, 1993). Considering the analysts' critical role in discovering, interpreting and relaying relevant information to investors to timely update target prices, we use the number of analysts as a proxy of investor attention, as firms with higher analyst following would reflect faster information incorporation in their prices. Following this intuition, we

[12] See Da, Engelberg, and Gao (2011) for information on variable construction and advantages over prior investor attention proxies.

generate *Number of Analysts*, which is a dummy variable that equals 1 for firms with greater analyst following relative to median of all firms in the year and zero otherwise.

We then interact the different measures of investor (in)attention with $NRDRet$ to disentangle the impact of investor attention on the NRD momentum effect. We expect focal firms that are subjected to greater investor attention will have lower return predictability due to the NRD effect, as their investors are more likely to timely digest information in these firms' risk factor disclosures.

[Insert Table 9 Here]

Results are presented in Table 9 Columns 1-4. Consistent with our hypothesis, the NRD momentum is less pronounced for firms that receive greater investor attention, indicated by negative and statistically significant coefficients on the interaction terms of interest. Our results concur the impact of investor attention on return predictability generated across firms, with higher attention attenuating such a lead-lag return relationship.

4.2.4.2 Stock Illiquidity

While investor inattention test results indicate that capital market participants may be slow to price the impact of overlapping NRD due to their cognitive and information processing capabilities, Easley, Kiefer, O'hara and Paperman (1996) show that illiquidity of stocks can be a major factor that impedes timely reflection of investor information into stock prices. To test this hypothesis, we employ the illiquidity measure of Amihud (2002) (the daily price impact to the order flow, and is computed as the average ratio of the daily absolute return to the trading volume of the day) . We then generate dummy variable $illiq$, that equals 1 for firms with higher illiquidity relative to the median of all firms in that year as measured by the respective proxy, and zero otherwise. Table 9 Column 5 presents the results of the interaction term between $illiq$ and $NRDret$, with its negative and statistically significant coefficient aligned with our conjecture: illiquidity hinders investors from timely incorporating the pricing implications of NRD.[13]

[13] In untabulated tests, we find consistent results with the relative bid-ask spread measure.

5. Supplementary and Robustness Tests

5.1. Common Risk Exposures and Firm Operations

Our prior findings demonstrate the capability of returns from NRD peers to forecast the future returns of focal firms. We also expect that shared risk exposures captures similarities in firms' fundamental operations, as common future realized risks are likely to impact NRD peers' operations in comparable ways.

To test this, we construct four operation metrics: the ratio of gross profit to sales (*GP)*, the return on total assets *(ROA)*, sales as a proportion of total assets (*Sales*) and *Sales Growth*. For each of these variables, we then calculate its NRD similarity weighted ratio of its peers, which are denoted as *NRD GP*, *NRD ROA*, *NRD Sales* and *NRD Sales Growth* respectively. We then regress focal firms' operational metrics on corresponding lagged NRD-similarity weighted ratios of the firms' peers and its own lagged concurrent measure. Other control variables include the lagged value of the firm size (*log(Mkt Cap)*, book to market ratio *(B/M*) and its research and development intensity *(R&D)*. Industry and year fixed effects are also included. As these are annual regressions, our observations fall to roughly 30,000.

[Insert Table 10 Here]

The results to these regressions are presented in Table 10. The positive and statistically significant coefficients on lagged NRD-similarity weighted operational metrics are indicative of their power to predict focal firm's future operations. The results are aligned with our intuition: firms with common risk exposures not only face correlated future returns, but also have strong predictive power over future operational matters. Beyond our return-based regression analyses, these tests offer corroborative evidence that our measures of common risk exposure effectively identify the degree to which firms' business fundamentals are influenced by the same systematic risks, or the "state variables" within the ICAPM framework.

5.2. Robustness tests

As a final step of our empirical analysis, we examine whether our main results are driven by the empirical choices we make. First, we make sure that the choice of bigrams as opposed to unigrams as well as the ex-ante defined number of topics do not impact our findings. Second, we investigate whether our results still hold when we use an alternative machine learning process, latent semantic analysis (LSA), to extract risk topics from Item 1A.

5.2.1 Alternating the number of topics and unigram LDA

An important discretion provided by the researcher in applying LDA is the number of topics to detect in the corpus (i.e., the collection of documents). As explained in Section 3.1, we employ the perplexity score to determine the number of topics that would best fit our data. In this section, we also repeat the procedure for extracting risk exposures by varying the number of topics specified to in a bigram LDA. Following studies that aim to capture systematic risk (Hanley and Hoberg, 2016; Lopez-Lira, 2020), we begin with 25 topics, and increase the number gradually to 150, which is used in Dyer et al. (2017) to discern topic in the entire 10-K. We then recalculate NRD similarity weighted lagged returns of peers based on topic distributions yielded in each of these scenarios. In another robustness test, we employ unigrams instead of bigrams to capture risk environments described in Item 1A. We then recalculate NRD similarity weighted lagged returns of peers based on varied number of topics generated by unigrams. We present results using bigrams with varied number of topics and the same tests with unigrams in Online Appendix Table 1. Overall, our results remain robust to this alternative specification.

5.2.2 Alternative topic modelling method: latent semantic analysis (LSA)

In this robustness test, we use LSA to assess NRD relatedness and to ensure that our choice of NLP techniques does not drive our results. LSA derives underlying themes in documents through associations (more precisely, term frequency-inverse document frequency, or TF-IDF) between words, words and passages and among passages. This property aligns with the way human beings think and derive meaning from texts. Similar with LDA, given a pre-specified number of dimensionality, LSA assigns coordinates to a document that pin point its position in high dimensional semantic space spanned by hidden semantic themes (i.e., topics).[14]

Using the topic distribution generated by bigram LSA, we reconstruct NRD similarity weighted peer returns [denoted as *NRDret*(LSA)] and repeat our baseline analyses presented in Tables 6 and 7. Shown in Online Appendix Table 2, we first employ Fama-MacBeth regressions to examine the predictive power of *NRDret*(LSA). We also vary the number of specified themes when implementing bigram LSA. Results are presented in Panel A. In Panels B and C, we repeat the zero-cost portfolio strategy using *NRDret*(LSA) calculated based on 100 semantic themes. Overall, these results are qualitatively similar to those produced by bigram LDA.

6. Conclusion

In this study, we employ LDA, an unsupervised machine learning method, to examine investors' ability to evaluate and efficiently incorporate risk factor information into market prices as a firm's environment evolves. Within the context of ICAPM, we argue that risk factor disclosures fulfill a twofold role: they provide insights into a firm's specific risk exposures and highlight shared risks across numerous firms, thus identifying systematic risks under ICAPM. Our analysis reveals that NRD exposure is a strong predictor of future returns across different securities, evidenced by both portfolio sorting and Fama-MacBeth regressions. These results affirm that our NRD exposure measure, based on risk factor disclosures, effectively captures firms' systematic risk exposures, and investors consider this information when pricing stocks.

Further, we posit that due to various information processing costs involved in understanding and interpreting these complicated texts, investors face challenges to promptly process the value relevant information of common risk exposure across firms. As a result, we hypothesize that this delayed

[14] For a more detailed summary of LSA, please refer to Landauer, Foltz, and Laham (1998).

information incorporation results in stock return predictability across firms with overlapping NRDs. We find results consistent with this hypothesis: a long-short portfolio trading strategy exploiting the NRD momentum generates monthly equal (value) weighted returns of 70 (58) basis points. Fama-MacBeth regressions show that NRD is an economically significant predictor of focal firm returns, even after controlling for other possible explanations of slow information diffusion, such as industry diversification, size-based explanations, and overlapping sales segments. We also find that this effect is stronger for firms facing higher investor inattention and stock illiquidity.

Our findings have significant policy implications. Firstly, they bolster the argument for mandating risk factor disclosures, as such disclosures contain systematic elements that enable investors across all firms to price systematic risks more effectively. Specifically, risk factor disclosures serve a dual purpose: they not only inform investors about their own firm's risk exposures but also illuminate common risks prevalent across numerous firms, thereby identifying systematic risks. This revelation of information spillover underscores the necessity for mandated disclosures, as individual firms may not recognize the broader benefits their disclosures provide to the investor community at large.

Additionally, our research suggests that while firms are meeting SEC requirements by including pertinent risk information in their 10-K reports, the textual complexity and high processing demands of these documents slow their impact on stock returns. After SEC's mandating of Item 1A in 2005, it has subsequently issued Concept Releases in 2013 and 2016 and amendments to Regulation S-K in 2020 to improve the informativeness and relevance of the section. Our study is particularly relevant in light of the SEC's amendments on August 26, 2020, aimed at modernizing risk factor disclosures (as well as business descriptions and legal proceedings sections) to enhance their readability and relevance.15 Consistent with current regulations that aim to decrease the opaqueness and complexity of firms' NRD, our paper

[15]SEC Chairman Jay Clayton noted that the new rules (amendments to Regulation S-K) "are rooted in materiality and seek to elicit information that will allow today's investors to make more informed investment decisions." The SEC's amendments to Regulation S-K came into effect on November 9, 2020 and apply to 10-Qs, 10-Ks and registration statements filed on or after that date as applicable.

underscores that, while mandated risk factor disclosures are instrumental in facilitating risk sharing and augmenting investor welfare, there remains a critical need to enhance the clarity and accessibility of these disclosures.

References

Ali, U., Hirshleifer, D., 2020. Shared analyst coverage: unifying momentum spillover effects. *Journal of Financial Economics*, 136(3), 649-675.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, *5(1),* 31-56.

Aobdia, D., Caskey, J., Ozel, N. B., 2014. Predictability of earnings and stock returns. *Review of Accounting Studies*, 19(3), 1191–1224

Ball, C., Hoberg, G., and Maksimovic, V. (2015). Disclosure and Firm Separation: A Text-Based Examination. *Working Paper*.

Bao, Y., and Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371-1391.

Blei, D., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Bloom, N., Schankerman, M., and Reenan, J. (2013). Identifying Technology Spillovers and Product Market Rivalry. *Econometrica*, *81(4),* 1347-1393.

Bogachev, V. I. (2007). Measure Theory (Vol. 1). *Springer Science & Business Media*.

Brennan, M., Jegadeesh, N., and Swaminathan, B. (1993). Investment Analysis and the Adjustment of Stock Prices to Common Information. *Review of Financial Studies*, 6(4), 799-824.

Brown, S., Shaolee, T., and Tucker, J. (2018). The Spillover Effect of SEC Comment Letters on Qualitative Corporate Disclosure: Evidence from the Risk Factor Disclosure. *Contemporary Accounting Research, 35(2),* 622-656.

Campbell, J., Chen, H., Dhaliwal, D., Lu, H., and Steele, L. (2014). The Information Content of Mandatory Risk Factor Disclosures in Corporate Filings. *Review of Accounting Studies, 19(1)*, 396-445.

Campbell, J. L., Cecchini, M., Cianci, A. M., Ehinger, A. C., and Werner, E. M. (2019). Tax-Related Mandatory Risk Factor Disclosures, Future Profitability, and Stock Returns. *Review of Accounting Studies*, *24*(1), 264-308.

Carhart, M. M. (1997). On the Persistence in Mutual Fund Performance. *Journal of Finance*, 52(1), 57-82.

Chan, L., Jegadeesh, N., and Lakonishok, J. (1996). Momentum Strategies. *Journal of Finance, 51(5)*, 1681–1713.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*, (pp. 288-296). New York.

Chiu, T. T., Guan, Y., and Kim, J. B. (2018). The Effect of Risk Factor Disclosures On The Pricing Of Credit Default Swaps. *Contemporary Accounting Research*, *35*(4), 2191-2224.

Cohen, L., Frazzini, A., 2008. Economic links and predictable returns. *Journal of Finance*, 63(4), 1977–2011.

Cohen, L., and Lou, D. (2012). Complicated Firms. *Journal of Financial Economics, 104(2),* 383-400.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, *75*(3), 1371-1415.

Da, Z., Engelberg, J., and Gao, P. (2011). In Search of Attention. *Journal of Finance, 66(5),* 1461-1499.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The Evolution of 10-K Disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics, 64(2-3),* 221-245.

Easley, D., Kiefer, N. M., O'hara, M., and Paperman, J. B. (1996). Liquidity, information, and infrequently traded stocks. *Journal of Finance*, 51(4), 1405-1436.

Fama, E., and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33(1),* 3-56.

Fama, E., and French, K. (2015). A five-factor asset pricing model. *Journal of Financial Economics, 116(1),* 1-22.

Fama, E., and MacBeth, J. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy, 81(3),* 607-636.

Hanley, K., and Hoberg, G. (2019). Dynamic Interpretation of Emerging Risks in the Financial Sector. *Review of Financial Studies*, *32*(12), 4543-4603.

Hirshleifer, D., Teoh, S. H., 2003. Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36, 337–386.

Hirshleifer, D., Hsu, P., and Dongmei, L. (2013). Innovative Efficiency and Stock Returns. *Journal of Financial Economics, 107(3),* 632-654.

Hoberg, G., and Lewis, C. (2017). Do Fraudulent Firms Produce Abnormal Disclosure? *Journal of Corporate Finance, 43*, 58-85.

Hoberg, G., and Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy, 124(5),* 1423–1465.

Hoberg, G., and Phillips, G. (2019). Text-Based Industry Momentum. *Journal of Financial and Quantitative Analysis, 53(6),* 2355–2388.

Hong, H., Stein, J. C., 1999. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *Journal of Finance*, 54(6), 2143–2184.

Hou, K. (2007). Industry Information Diffusion and the Lead-Lag Effect in Stock Returns. *Review of Financial Studies*, *20*(4), 1113-1138.

Hou, K., van Dijk, M. A., Zhang, Y., 2012. The implied cost of capital: a new approach. *Journal of Accounting and Economics*, 53(3), 504–526.

Huang, X., 2015. Thinking outside the borders: investors' underreaction to foreign operations information. *Review of Financial Studies*, 28(11), 3109–3152.

Huang, A., Lehavy, R., Zhang, A., and Zheng, R. (2017). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science, 64(6),* 2833-2855.

Jegadeesh, N., and Titman, S. (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance, 48(1),* 65-91.

Landauer, T., Foltz, P., and Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25(2-3),* 259-284.

Lee, C., and Swaminathan, B. (2000). Price Momentum and Trading Volume. *Journal of Finance, 55(5)*, 2017-2069.

Lee, C. M. C., Ma, P., Wang, C. C. Y., 2015. Search-based peer firms: aggregating investor perceptions through internet co-searches. *Journal of Financial Economics*, 116(2), 410–431.

Lee, C., Sun, S., Wang, R., and Zhang, R. (2019). Technological links and predictable returns. *Journal of Financial Economics, 132(3)*, 76-96.

Li, N., Richardson, S., Tuna, I., 2014. Macro to micro: country exposures, firm fundamentals and stock returns. *Journal of Accounting and Economics*, 58(1), 1–20.

Lo, A., and MacKinlay, A. (1990). When Are Contrarian Profits Due to Stock Market Overreaction? *Review of Financial Studies, 3(2),* 175-205.

Lopez-Lira, A. (2020). Risk Factors that Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns. *Working Paper*.

Merton, R. C., 1973. An intertemporal capital asset pricing model. *Econometrica* 41, 867–887.

Merton, R., 1987. A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, 42(3), 483–510

Moskowitz, T. J., and Grinblatt, M. (1999). Do Industries Explain Momentum?. *Journal of Finance*, *54*(4), 1249-1290.

Parsons, C. A., Sabbatucci, R., Titman, S., 2020. Geographic lead-lag effects. The *Review of Financial Studies,* 33(10), 4721-4770.

Peng, L., Xiong, W., 2006. Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80(3), 563–602.

Ross, L. (2019). Are Item 1A Risk Factors Priced? *Working Papers*.

Schnatterley, K., Shaw, K., and Jennings, W. (2008). Information Advantages of Large Institutional Owners. *Strategic Management Journal, 29,* 219-227.

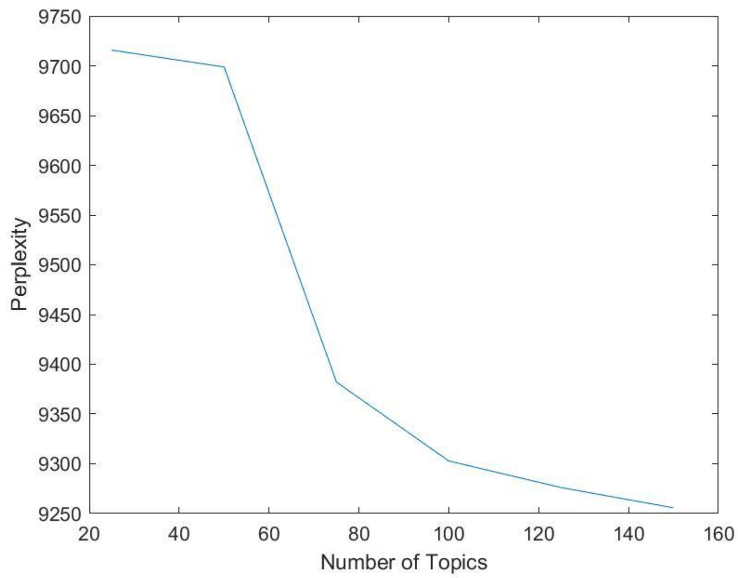
Schrand, C., and Elliot, J. (1998). Risk and Financial Reporting: A summary of the discussion at the 1997 AAA/FASB Conference. *Accounting Horizons, 12(3),* 271-288.

Wallach, H. M. (2006). Topic Modelling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (pp. 977-984). Pittsburgh: ACM New York.

Figure 1: Perplexity by number of topics

These figures map the perplexity score of applying bigram LDA on all Item 1As in 2009 (Panel A) and 2013 (Panel B). We vary the number of topics from 25 to 150.

Panel A. Perplexity of bigram LDA estimated using all NRD in 2009



Panel B. Perplexity of bigram LDA estimated using all NRD in 2013

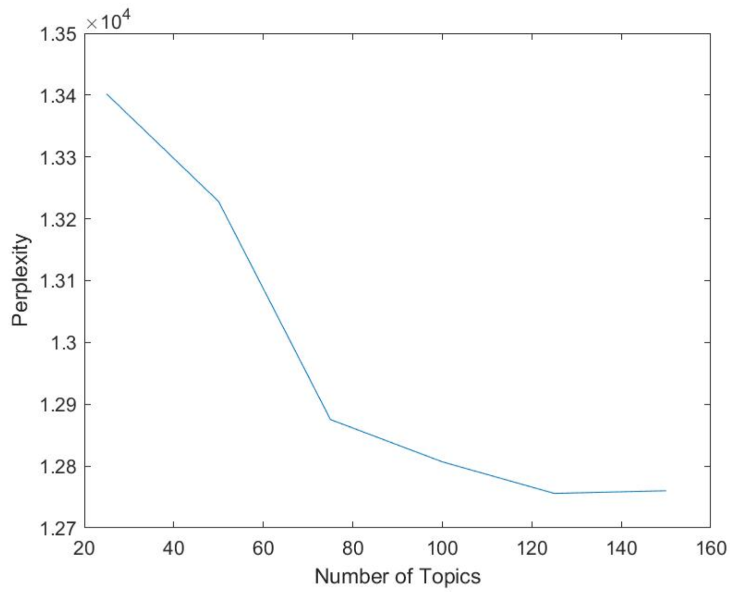
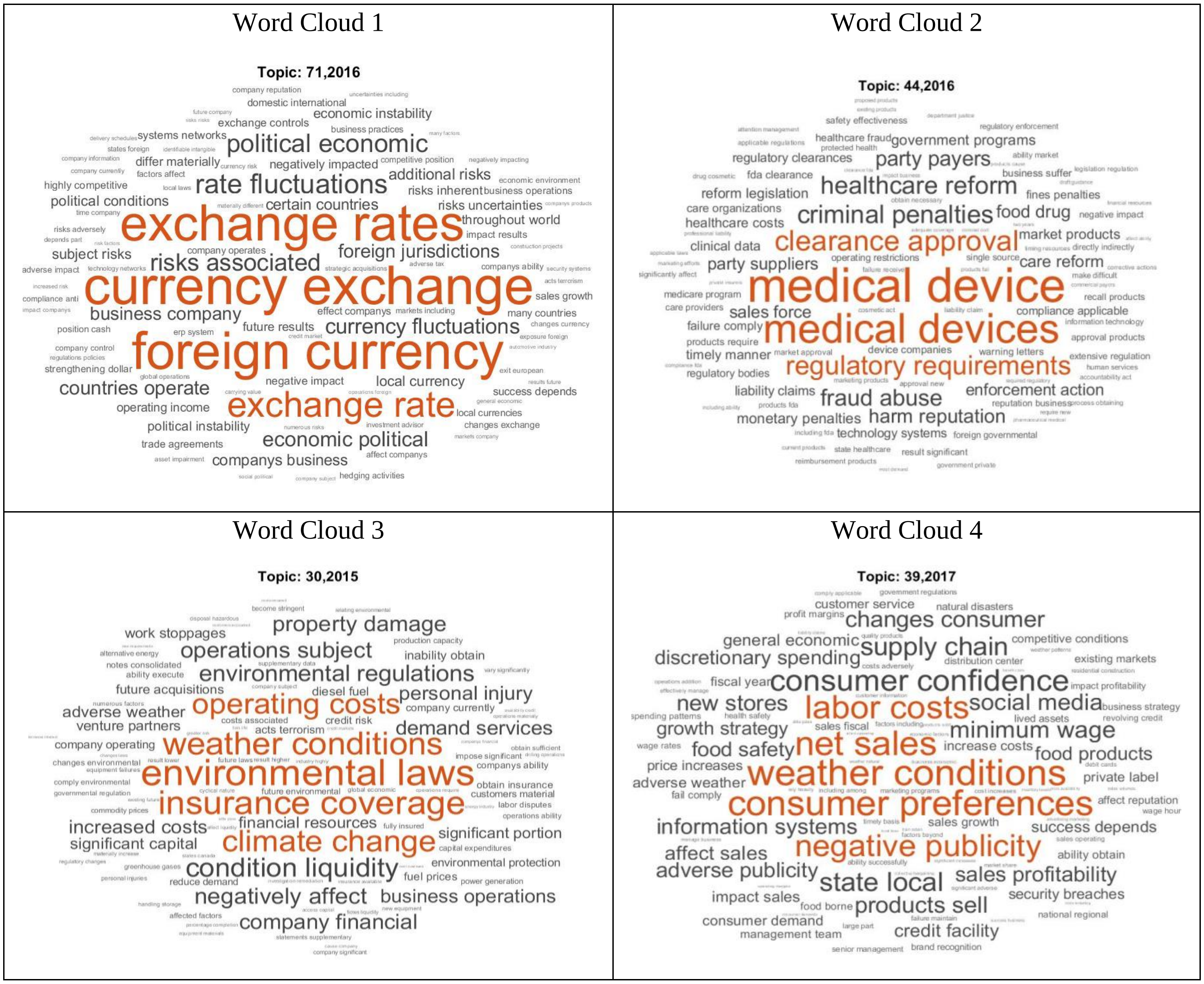


Figure 2: Word Clouds

The figure presents four word clouds generated from keywords associated with various topics that LDA has traced in firms' NRD in a year. The font size of each bigram represents its relative frequency of usage (and therefore relevance) to each topic.



Appendix: Variable Definition

Variable	Definition
NRDret	NRD-linked returns, defined as the NRD similarity weighted average of firms' monthly stock returns. Formally, NRDret of firm *i* in month *t* is calculated as the follows: $$NRDret_{i,t} \equiv \sum_{j \neq i} \omega_{ij} Ret_{j,k} \Big/ \sum_{j \neq i} \omega_{ijt},$$ Where the summation is over all firms in a given year.
Ret	Monthly raw stock returns.
HHI(NRD)	Concentration of risk topics distributions and defined by squaring the proportion of each firm's risk topics mentions in the distribution of a firm's Item 1A (i.e. NRD defined in Section 3.3) and then summing the resulting numbers.
BM	Book-to-market ratio defined as the book equity divided by total market value at the fiscal year end.
GP	Gross profitability ratio, defined as revenue less cost of goods sold, scaled by assets.
AG	Asset growth, defined as yearly growth rate of total assets.
R&D	Research and Development intensity, defined as research and development expenditure scaled by total sales.
MOM	Medium-term price momentum, defined as the focal firm's stock return over the last 6 months, excluding the most recent month.
Turnover	Share turnover, defined as the stock's turnover over the last 12 months.
FF48 Returns	Industry return, defined as the value weighted average industry return in each FF48 industry.

Table 1. Summary Statistics of key variables

This table provides summary statistics for the key variables in the regression. The sample includes all securities listed in NYSE/AMEX/NASDAQ with the share codes 10 or 11. All variables are winsorized at 1 and 99% and defined in the appendix.

	Mean	SD	P25	Median	P75	Obs
NRD_EXP	5.90	6.10	0.98	5.46	6.60	319216
NRDrett-1 (%)	1.080	5.370	-1.620	1.580	4.310	319216
HHI (NRD) (%)	2.730	0.820	2.220	2.550	2.980	319216
Log(Mkt Cap)	13.431	1.913	12.080	13.364	14.710	319216
BM	1.400	1.700	0.500	0.800	1.500	319216
GP	0.334	0.297	0.189	0.315	0.474	319216
AG	0.108	0.323	-0.034	0.054	0.174	319216
R&D	0.067	0.131	0.000	0.008	0.077	319216
Ret_{t-1} (%)	0.950	13.410	-6.300	0.490	7.330	319216
MOM	5.590	30.430	-11.010	5.050	20.960	319216
Turnover	0.205	0.181	0.088	0.156	0.259	319216
FF48 Returns (%)	0.014	0.051	-0.015	0.018	0.046	319216

Table 2. Characteristics of NRD Exposure Portfolios

This table reports the time series average of cross-sectional mean characteristics for portfolios of sorted by NRD Exposure (NRD_EXP). β^{M} denotes the CAPM beta, B/M is the book-to-market ratio, ME is the market equity quintile, Momentum is the 12-month cumulative return. Op. Profit., AG, LEV, and IK are operating profitability, asset growth, leverage, and investment, respectively. The sample period is from 2006 to 2018.

Quintile	NRD_EXP	β^{M}	B/M	ME	Momentum	Profitability	Asset Growth	LEV	IK
5	6.958	1.250	0.440	4.068	0.139	0.171	0.126	0.672	0.001
4	6.460	1.307	0.536	3.216	0.126	0.155	0.109	0.723	0.003
3	6.080	1.304	0.585	2.741	0.108	0.147	0.104	0.765	0.008
2	5.588	1.317	0.580	2.530	0.113	0.118	0.128	0.686	0.017
1	4.057	1.476	0.486	2.446	0.135	0.030	0.238	0.516	0.052

Table 3. NRD Exposure and Stock Returns – Portfolio Analysis

This table reports the relation between NRD Exposure (NRD_EXP) and stock return using portfolio analysis. Panel A (Panel B) reports monthly equal-weighted (size-weighted) average returns excess of risk-free rate (R), the Fama and French (1993) three-factor alpha (α^{3F}), the Carhart (1997) four-factor alpha (α^{4F}), the Fama and French (2015) five-factor alpha (α^{5F}), the six-factor alpha that is based on the Fama and French (2015) five-factor model and momentum factor of Carhart (1997). *t*–statistics are computed from standard errors that are adjusted for heteroscedasticity and serial correlation following Newey and West (1987). ***, **, * denote statistical significance at the 1%, 5%, or 10% levels, respectively. The sample period is from June 2008 to December 2018.

	Panel A. Equal-Weighted					Panel B. Size-Weighted				
	Ret	α^{3F}	α^{4F}	α^{5F}	α^{6F}	Ret	α^{3F}	α^{4F}	α^{5F}	α^{6F}
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Q5	1.071**	0.302***	0.314***	0.272***	0.278***	0.883**	0.127*	0.137**	0.111	0.116*
	(2.06)	(2.74)	(3.52)	(2.66)	(3.19)	(2.08)	(1.67)	(2.19)	(1.52)	(1.80)
Q4	0.890*	0.163	0.174**	0.149	0.154*	0.768*	0.048	0.053	0.037	0.039
	(1.67)	(1.56)	(2.06)	(1.43)	(1.76)	(1.85)	(0.59)	(0.69)	(0.46)	(0.53)
Q3	0.688	-0.029	-0.011	-0.021	-0.013	0.882**	0.150	0.158*	0.129	0.133
	(1.24)	(-0.20)	(-0.11)	(-0.15)	(-0.12)	(2.02)	(1.62)	(1.88)	(1.36)	(1.48)
Q2	0.611	-0.103	-0.089	-0.076	-0.069	0.675	-0.004	0.003	-0.060	-0.057
	(1.09)	(-0.66)	(-0.70)	(-0.52)	(-0.60)	(1.52)	(-0.03)	(0.03)	(-0.50)	(-0.52)
Q1	0.430	-0.438**	-0.424**	-0.197	-0.191	0.399	-0.326*	-0.318*	-0.336*	-0.332
	(0.67)	(-2.15)	(-2.12)	(-0.92)	(-0.92)	(0.83)	(-1.77)	(-1.72)	(-1.66)	(-1.65)
	Difference									
Q5-Q1	0.641***	0.739***	0.739***	0.469**	0.468**	0.484**	0.453**	0.455**	0.448**	0.448**
	(2.79)	(3.44)	(3.41)	(1.99)	(1.98)	(2.49)	(2.40)	(2.40)	(2.15)	(2.14)

Table 4. NRD Exposure and Stock Returns – Regression Analysis

This table reports the results from Fama and MacBeth (1973) regressions of monthly stock returns on NRD Exposure (NRD_EXP) and control variables including the natural logarithm of size, the natural logarithm of book-to-market ratio, past performance (Ret_{t-1} and $\mathrm{Ret}_{t-12:t-2}$), operating profitability (Op. Profit.), and asset growth (AG) as described in APPENDIX. Industry fixed effects are based on industries defined by the Fama and French 48 industry classifications. Industry fixed effects are based on the Fama and French (1997) 48 industries. The *t*-statistics, reported in parentheses, are computed from standard errors that are adjusted for heteroskedasticity and serial correlation following Newey and West (1987), with a lag of 12. ***, **, * denote statistical significance at the 1%, 5%, or 10% levels, respectively. The sample period is from June 2006 to December 2018.

	(1)	(2)	(3)	(4)	(5)	(6)
NRD_EXP	0.212***	0.203***	0.165**	0.179***	0.175***	0.142**
	(3.27)	(2.93)	(2.39)	(3.10)	(3.10)	(2.45)
Log(Size)	0.004	-0.001	-0.011	0.041	0.042	0.033
	(0.09)	(-0.03)	(-0.24)	(0.78)	(0.87)	(0.70)
Log(B/M)	0.023	0.011	0.064	0.127	0.124*	0.213**
	(0.20)	(0.11)	(0.58)	(1.62)	(1.70)	(2.31)
Ret_{t-1}		-1.456***	-1.513***		-1.694***	-1.722***
		(-3.02)	(-3.18)		(-3.65)	(-3.72)
$\mathrm{Ret}_{t-12:t-2}$		-0.187	-0.209		-0.334	-0.362
		(-0.28)	(-0.32)		(-0.51)	(-0.55)
Op. Profit.			0.357***			0.440***
			(3.04)			(4.03)
AG			-0.421**			-0.404***
			(-2.41)			(-2.79)
Intercept	-0.572	-0.841	-0.580	-1.386	-1.722	-1.469
	(-0.59)	(-0.91)	(-0.62)	(-0.90)	(-1.19)	(-1.02)
Industry FE	No	No	No	Yes	Yes	Yes
No of Month	150	150	150	150	150	150
Avg. N	1,688	1,688	1,688	1,688	1,688	1,688
Adj. R^2	0.017	0.032	0.035	0.056	0.067	0.070

Table 5. Characteristics of NRD Exposure Portfolios

This table considers NRD Exposure factor (equal-weighted in Panel A and size-weighted in Panel B) as an additional factor and uses five factors in regression to explain average returns on the sixth. MKTRF is the value-weighted return on the market portfolio of all sample stocks minus the one-month Treasury bill rate. SMB (small minus big) is the size factor. HML (high minus low B/M) is the value factor. RMW (robust minus weak OP) is the profitability factor, and CMA (conservative minus aggressive Inv) is the investment factor. *t*–statistics are computed from standard errors that are adjusted for heteroscedasticity and serial correlation following Newey and West (1987). ***, **, * denote statistical significance at the 1%, 5%, or 10% levels, respectively. The sample period is from June 2008 to December 2018.

	Panel A. NRD_EXP Premium (Equal-Weighted)							
	Intercept	MKTRF	SMB	HML	RMW	CMA	NRD	R2
MKTRF	0.922***		0.428***	0.418*	-0.729**	-0.641**	0.035	0.287
	(3.02)		(2.79)	(1.96)	(-2.36)	(-2.42)	(0.26)	
SMB	0.170	0.132***		0.180**	-0.125	0.080	-0.249***	0.311
	(1.09)	(3.08)		(2.39)	(-1.03)	(0.69)	(-3.79)	
HML	-0.347*	0.129*	0.180**		-0.280**	0.917***	0.136	0.415
	(-1.66)	(1.68)	(2.07)		(-2.31)	(8.23)	(1.60)	
RMW	0.158	-0.091**	-0.051	-0.113**		0.082	0.203***	0.363
	(1.13)	(-2.53)	(-1.02)	(-2.33)		(0.93)	(4.54)	
CMA	0.121	-0.070**	0.028	0.322***	0.072		0.019	0.330
	(1.13)	(-2.24)	(0.69)	(6.39)	(0.89)		(0.47)	
NRD_EXP	0.499**	0.017	-0.399***	0.218*	0.800***	0.088		0.328
	(2.12)	(0.26)	(-3.36)	(1.91)	(4.31)	(0.46)		
	Panel B. NRD_EXP Premium (Size-Weighted)							
MKTRF	0.929***		0.424***	0.427**	-0.701***	-0.639**	0.023	0.287
	(3.19)		(2.98)	(2.07)	(-2.75)	(-2.41)	(0.18)	
SMB	0.191	0.128***		0.112	-0.322***	0.066	-0.288***	0.321
	(1.14)	(3.15)		(1.38)	(-2.84)	(0.55)	(-4.02)	
HML	-0.272	0.136*	0.118		-0.177	0.957***	-0.031	0.398
	(-1.32)	(1.66)	(1.25)		(-1.53)	(8.00)	(-0.36)	
RMW	0.311**	-0.104***	-0.159***	-0.083		0.119	-0.004	0.240
	(2.21)	(-3.17)	(-2.73)	(-1.41)		(1.24)	(-0.08)	
CMA	0.127	-0.069**	0.024	0.327***	0.087		0.008	0.329
	(1.18)	(-2.23)	(0.56)	(6.12)	(1.22)		(0.16)	
NRD_EXP	0.507**	0.010	-0.392***	-0.040	-0.012	0.030		0.145
	(2.33)	(0.19)	(-3.75)	(-0.34)	(-0.08)	(0.16)		

Table 6. Portfolio Sorting: NRD Lead-lag Effects

The following tables report abnormal returns for the NRD momentum strategy. We first sort firms into two size groups: small and large firms in a month. Then the top *NRDret* portfolio (portfolio 5) combines 20% of the small and 20% of the large firms, both of which have the highest *NRDret* in the previous month. Similarly, the bottom *NRDret* portfolio (portfolio 1) contains 20% firms from each size group that have lowest *NRDret*. We exclude financial firms (one-digit SIC code=6) and stocks with price less than $1 at portfolio formation. All stocks are then balanced monthly to maintain equal weights in Panel A, and value weights in Panel B. Column (1) presents raw returns of the portfolio over the risk-free rate. Column (2) presents the intercept from the regression of monthly excess returns on factor returns. Column (3) to (6) use factor returns from Kenneth French Data Library for the following factors: Fama and French (1993) three factor model, including Carhart (1997) momentum to form four factor model, Fama and French (2015) five factor model, and six factor model (which incorporates momentum in the five factor model). 5-1 is the alpha of a zero-cost portfolio that goes long on the top 20% stocks ranked by *NRDret* and short in the bottom 20%. Returns are in monthly percent, those marked with ***, ** and * indicate statistical significance levels of 1%, 5%, and 10%, respectively. The sample period is July 2006 and December 2018.

Panel A. Equally-Weighted Portfolio

	(1)	(2)	(3)	(4)	(5)	(6)
	Excess Returns	CAPM alpha	3-Factor Model	4-Factor Model	5-Factor Model	6-Factor Model
1	0.234	0.019	0.005	0.033	0.213	0.226
	(0.47)	(0.04)	(0.01)	(0.07)	(0.41)	(0.44)
2	0.806	0.596	0.585	0.600	0.673	0.681
	(1.65)	(1.23)	(1.21)	(1.24)	(1.33)	(1.35)
3	0.813*	0.677	0.638	0.654	0.843*	0.851*
	(1.87)	(1.55)	(1.47)	(1.51)	(1.88)	(1.90)
4	0.935**	0.73	0.732	0.758*	0.879*	0.892*
	(2.03)	(1.60)	(1.60)	(1.67)	(1.84)	(1.88)
5	0.938**	0.821*	0.812*	0.829*	0.969**	0.978**
	(2.18)	(1.89)	(1.88)	(1.92)	(2.15)	(2.18)
5-1	**0.704*** **	**0.802*** **	**0.807*** **	**0.796*** **	**0.756*** **	**0.751*** **
	(2.96)	**(3.38)**	**(3.35)**	**(3.32)**	**(3.04)**	**(3.03)**

Panel B. Value-Weighted Portfolio

	(1) Excess Returns	(2) CAPM alpha	(3) 3-Factor Model	(4) 4-Factor Model	(5) 5-Factor Model	(6) 6-Factor Model
1	0.337	0.148	0.117	0.145	0.331	0.344
	(0.63)	(0.27)	(0.22)	(0.27)	(0.59)	(0.62)
2	0.811*	0.672	0.664	0.676	0.730	0.736
	(1.80)	(1.48)	(1.47)	(1.50)	(1.55)	(1.56)
3	0.844**	0.749**	0.703*	0.720**	0.892**	0.901**
	(2.37)	(2.09)	(1.96)	(2.02)	(2.41)	(2.45)
4	0.995**	0.839*	0.821*	0.845*	0.962*	0.974*
	(2.01)	(1.69)	(1.66)	(1.71)	(1.86)	(1.89)
5	0.914**	0.841**	0.835**	0.852**	0.968**	0.977**
	(2.53)	(2.30)	(2.28)	(2.33)	(2.54)	(2.57)
5-1	**0.577***	**0.693****	**0.718****	**0.708****	**0.637***	**0.633***
	(1.79)	**(2.15)**	**(2.22)**	**(2.19)**	**(1.90)**	**(1.88)**

Table 7. Cross-sectional Regression: NRD Lead-lag Effects

This table reports the results of the baseline Fama-MacBeth regression. The dependent variable is the focal firm's monthly return. The explanatory variables include NRD linked returns (*NRDret*). We control for HHI(NRD). HHI(NRD) is calculated by squaring the proportion of risk topic in the distribution of a firm's Item 1A (i.e., *NRD* defined in Section 3.3) and then summing the resulting numbers. It is used to capture the concentration of risk topics. We also include size [log(MktCap)], book-to-market ratio (BM), gross profitability (GP), asset growth (AG), R&D intensity (R&D), focal firm's own lagged return (Ret_{t-1}), medium-term price momentum (MOM), share turnover (Turnover), value-weighted FF48 industry returns (columns 1 and 3) and TNIC-2 peers momentum (column 2). Industry and year fixed effects are also included. All variables are defined in the Appendix. Fama-MacBeth t-statistics are reported in parentheses. The time-series standard errors are Newey-West adjusted (up to 12 lags) for heteroscedasticity and autocorrelation. Coefficients marked with ***, ** and * indicate statistical significance levels of 1%, 5%, and 10%, respectively. All variables are winsorized at 1 and 99% and defined in the appendix. The sample period is July 2006 and December 2018.

	(1)	(2)	(3)
	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$	0.748***	0.669***	
	(3.22)	(3.02)	
$NRDret(Decile)_{t-1}$			0.775***
			(4.54)
HHI(NRD)	0.090*	0.092**	0.081*
	(1.92)	(2.05)	(1.93)
Log(Mkt Cap)	-0.119	-0.109	-0.118
	(1.10)	(0.81)	(1.05)
BM	0.114	0.125*	0.116*
	(1.64)	(1.79)	(1.70)
GP	0.862***	0.838***	0.859***
	(4.63)	(4.85)	(4.71)
AG	-0.379**	-0.321**	-0.371**
	(-2.53)	(-2.05)	(-2.54)
R&D	0.239	0.352	0.233
	(0.43)	(0.61)	(0.40)
Ret_{t-1}	-0.021***	-0.020***	-0.021***
	(-4.62)	(-4.07)	(-4.54)
MOM	0.002	-0.000	0.000
	(0.10)	(-0.10)	(0.12)
Turnover	-1.470***	-1.452***	-1.470***

	(-4.03)	(-3.92)	(-4.01)
FF48 Returns	0.040***		0.038***
	(2.85)		(2.72)
TNIC mom		0.065***	
		(4.17)	
Industry and year FE	Y	Y	Y
Observations	319,216	307,472	319,216
Avg. R^2	0.099	0.100	0.099

Table 8. Controlling for Other Economic Linkages

This table reports results of the Fama-MacBeth regressions after incorporating several other economic linkages discussed in prior literature. The dependent variable is the focal firm's monthly return. The explanatory variables include NRD linked returns (*NRDret*). In column (1), in each year, we exclude all firms that sell to multiple industries to eliminate the conglomerate effect studied in Cohen and Lou (2012). In column (2), we incorporate lagged returns of firms that have similar segment sales distributions ($SaleRet_{t-1}$). In column (3), we drop the 1/3rd largest TNIC-2 peers to control for the intra-industry lead-lag effect. All controls from our baseline regression and industry and year fixed effects are included. The time-series standard errors are Newey-West adjusted (up to 12 lags) for heteroscedasticity and autocorrelation. Fama-MacBeth t-statistics are reported in parentheses. All variables are defined in the appendix. Coefficients marked with ***, ** and * indicate significance levels of 1%, 5%, and 10%, respectively. All variables are winsorized at 1 and 99% and defined in the appendix.

	standalone	segment sales	size diffusion
	(1)	(2)	(3)
	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$	0.924**	0.632**	0.665**
	(2.56)	(2.38)	(2.60)
$SaleRet_{t-1}$		0.021**	
		(2.86)	
Other Controls	Y	Y	Y
Industry and year FE	Y	Y	Y
Observations	155,792	278,554	209,045
Avg R^2	0.189	0.100	0.101

Table 9. Firm Heterogeneity and Cross-Sectional Tests

In this table, we report the results of our cross-sectional tests to examine the sensitivity of the NRD momentum to various firm characteristics. We run Fama-MacBeth forecasting regressions, with focal firm returns (Ret_t) being the dependent variables. Explanatory variables are *$NRDret_{t-1}$* (one-month lagged returns of NRD linked firms) interacted with dummy variables that equal one if the underlying variable is above the median of all firms in a given year, and zero otherwise:, *Market Cap*, *Google SVI* (Individual investor attention to the firm, measured by Google search trends), *Institutional Holdings, Number of Analysts* (logged number of analysts following the firm), and *illiq* (ratio of absolute change in stock price to excess demand for trading). The time-series standard errors are Newey-West adjusted (up to 12 lags) for heteroscedasticity and autocorrelation. Fama-MacBeth t-statistics are reported in parentheses. We include all controls and fixed effects from our baseline regression. All variables are defined in the appendix. Coefficients marked with ***, ** and * indicate significance levels of 1%, 5%, and 10%, respectively. All variables are winsorized at 1 and 99% and defined in the appendix.

	(1)	(2)	(3)	(4)	(5)
	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$	1.331***	1.072***	1.01***	1.146***	0.054
	(4.98)	(3.25)	(4.01)	(4.31)	(0.21)
$NRDret_{t-1}$×(Market Cap > median)	-1.227***				
	(-5.40)				
$NRDret_{t-1}$×(Google SVI > median)		-0.149*			
		(-1.75)			
$NRDret_{t-1}$×(Institutional holdings > median)			-0.612**		
			(-2.52)		
$NRDret_{t-1}$×(Number of analysts > median)				-0.890***	
				(-4.33)	
$NRDret_{t-1}$×(illiq > median)					1.084***
					(5.13)
Other controls	Y	Y	Y	Y	Y
Industry and year FE	Y	Y	Y	Y	Y
Observations	316,361	243,669	319,216	319,216	298,972
Avg R^2	0.10	0.11	0.10	0.10	0.11

Table 10. NRD Linkages and Firm Operations

This table examines the forecasting ability of NRD momentum for gross profit, return on assets, sales and sales growth in Columns 1 to 4 respectively. The independent variable of interests are the firms peers' lagged NRD-similarity weighted gross profit (column 1), return on assets (column 2), sales (column 3) and sales growth (column 4). We further control for the firm's own lagged value of the dependent variable. All control variables are lagged by a year: market capitalization [log(MktCap)], book-to-market (B/M) and R&D intensity (R&D). All regressions include industry and year fixed effects. All variables are defined in the appendix. The standard errors are clustered at the firm level. Coefficients marked with ***, ** and * indicate significance levels of 1%, 5%, and 10%, respectively. All variables are winsorized at 1 and 99% and defined in the appendix.

	(1)	(2)	(3)	(4)
	GP_y	ROA_y	$Sales_y$	$Sales\ Growth_y$
NRD GP_{y-1}	1.199***			
	(14.34)			
NRD ROA_{y-1}		0.567***		
		(4.07)		
NRD $Sales_{y-1}$			1.388***	
			(15.97)	
NRD $Sales\ Growth_{y-1}$				0.690**
				(2.42)
GP_{y-1}	0.730***			
	(56.22)			
ROA_{y-1}		0.539***		
		(23.86)		
$Sales_{y-1}$			0.722***	
			(64.10)	
$Sales\ Growth_{y-1}$				0.059***
				(4.13)
Log(Mkt $Cap)_{y-1}$	-0.003***	0.056***	-0.028***	0.013***
	(-3.56)	(12.03)	(-13.82)	(10.17)
B/M_{y-1}	-0.001***	0.011***	-0.006***	-0.007***
	(-2.84)	(9.20)	(-6.98)	(-8.68)
$R\&D_{y-1}$	0.003	-1.521***	0.282***	-0.001
	(0.17)	(-11.33)	(7.38)	(-0.04)
Industry and year FE	Y	Y	Y	Y

Observations	34,670	34,670	34,670	31,888
R^2	0.643	0.541	0.721	0.043

OA 1. Bigram and Unigram Results

These tables report Fama-MacBeth regressions NRDret generated from bigram and unigram LDA with different numbers of topics. In panel A, we re-run baseline regressions by specifying bigrams for topic detection in Item 1A and vary the number of topics to be inferred from 25 to 150. In Panel B, we switch from bigrams to unigrams for detecting topics in Item 1A, and vary the number of topics to be inferred from 25 to 150. The dependent variable is the focal firm's monthly return. The explanatory variables include NRD linked returns derived from various number of topics from bigrams and unigrams (*NRDret*). All regressions include controls and fixed effects from the baseline specification. The time-series standard errors are Newey-West adjusted (up to 12 lags) for heteroscedasticity and autocorrelation. Fama-MacBeth t-statistics are reported in parentheses. All variables are defined in the appendix. Coefficients marked with ***, ** and * indicate significance levels of 1%, 5%, and 10%, respectively. All variables are winsorized at 1 and 99% and defined in the appendix.

Panel A. Bigram LDA with different number of topics

	(1)	(2)	(3)	(4)	(5)	(6)
	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$ (25 topics)	0.496***					
	(3.05)					
$NRDret_{t-1}$ (50 topics)		0.595***				
		(2.74)				
$NRDret_{t-1}$ (75 topics)			0.705***			
			(3.03)			
$NRDret_{t-1}$ (100 topics)				0.748***		
				(3.22)		
$NRDret_{t-1}$ (125 topics)					0.836***	
					(3.47)	
$NRDret_{t-1}$ (150 topics)						0.829***
						(3.30)
Other controls	Y	Y	Y	Y	Y	Y
Industry and year FE	Y	Y	Y	Y	Y	Y
Observations	319,216	319,216	319,216	319,216	319,216	319,216
Avg R^2	0.100	0.099	0.099	0.099	0.099	0.099

Panel B. Unigram LDA with different number of topics

	(1)	(2)	(3)	(4)	(5)	(6)
	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$ (25 topics)	0.352***					
	(3.70)					
$NRDret_{t-1}$ (50 topics)		0.698***				
		(3.93)				
$NRDret_{t-1}$ (75 topics)			0.535***			
			(3.13)			
$NRDret_{t-1}$ (100 topics)				0.678***		
				(3.79)		
$NRDret_{t-1}$ (125 topics)					0.701***	
					(4.52)	
$NRDret_{t-1}$ (150 topics)						0.779***
						(3.77)
Other controls	Y	Y	Y	Y	Y	Y
Industry and year FE	Y	Y	Y	Y	Y	Y
Observations	319,216	319,216	319,216	319,216	319,216	319,216
Avg R^2	0.100	0.100	0.100	0.100	0.100	0.100

OA 2. Latent Semantic Analysis

The following tables display results using latent semantic analysis for detecting risk themes in Item 1A. In Panel A, we vary the number of themes to be detected from 25 to 150. The key explanatory variables is NRD similarity weighted peer returns where the topic distribution is determined by LSA. The variable is labelled as *NRDret*(LSA). The regression also includes other control variables and fixed effects from baseline specifications presented in Table 2. Fama-MacBeth t-statistics are reported in parentheses. The time-series standard errors are Newey-West adjusted (up to 12 lags) for heteroscedasticity and autocorrelation. Coefficients marked with ***, ** and * indicate statistical significance levels of 1%, 5%, and 10%, respectively.

Panels B and C are for portfolio results using *NRDret*(LSA) with 100 themes. We first sort firms into two size groups: small and large firms in a month. Then the top *NRDret* portfolio (portfolio 5) combines 20% of the small and 20% of the large firms, both of which have the highest *NRDret*(LSA) in the previous month. Similarly, the bottom *NRDret*(LSA)portfolio (portfolio 1) contains 20% firms from each size group that have lowest *NRDret*(LSA). We exclude financial firms (one-digit SIC code=6) and stocks with price less than $1 at portfolio formation. All stocks are then balanced monthly to maintain equal weights in Panel A, and value weights in Panel B. Column (1) presents raw returns of the portfolio over the risk-free rate. Column (2) presents the intercept from the regression of monthly excess returns on factor returns. Column (3) to (6) use factor returns from Kenneth French Data Library for the following factors: Fama and French (1993) three factor model, including Carhart (1997) momentum to form four factor model, Fama and French (2015) five factor model, and six factor model (which incorporates momentum in the five factor model). 5-1 is the alpha of a zero-cost portfolio that goes long on the top 20% stocks ranked by *NRDret*(LSA)and short in the bottom 20%. Returns are in monthly percent, those marked with ***, ** and * indicate statistical significance levels of 1%, 5%, and 10%, respectively. The sample period is July 2006 and December 2018.

Panel A. Fama-MacBeth Regressions with *NRDret*(LSA)

	(1)	(2)	(3)	(4)	(5)	(6)
	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t	Ret_t
$NRDret_{t-1}$(LSA 25)	0.500***					
	(3.14)					
$NRDret_{t-1}$(LSA 50)		0.381**				
		(2.41)				
$NRDret_{t-1}$(LSA 75)			0.402***			
			(2.85)			
$NRDret_{t-1}$(LSA 100)				0.380**		
				(2.47)		
$NRDret_{t-1}$(LSA 125)					0.444***	
					(2.87)	
$NRDret_{t-1}$(LSA 150)						0.477***
						(3.24)
Other controls	Y	Y	Y	Y	Y	Y

Industry and year FE	Y	Y	Y	Y	Y	Y
Observations	319,216	319,216	319,216	319,216	319,216	319,216
Avg R^2	0.099	0.099	0.099	0.099	0.099	0.099

Panel B. Equal-weighted Portfolio Results on *NRDret*(LSA 100)

	(1) Excess Returns	(2) CAPM alpha	(3) 3-Factor Alpha	(4) 4-Factor Alpha	(5) 5-Factor Alpha	(6) 6-Factor Alpha
1	0.295	0.079	0.075	0.098	0.257	0.269
	(0.59)	(0.16)	(0.15)	(0.20)	(0.50)	(0.52)
2	0.743	0.526	0.509	0.528	0.602	0.612
	(1.53)	(1.09)	(1.05)	(1.10)	(1.19)	(1.21)
3	0.900*	0.696	0.696	0.723	0.840*	0.854*
	(1.95)	(1.52)	(1.52)	(1.59)	(1.75)	(1.80)
4	0.806*	0.682	0.644	0.659	0.849*	0.857*
	(1.83)	(1.54)	(1.46)	(1.50)	(1.86)	(1.88)
5	0.968**	0.839*	0.823*	0.842*	0.983**	0.993**
	(2.25)	(1.93)	(1.91)	(1.96)	(2.19)	(2.22)
5-1	**0.674*****	**0.759*****	**0.748*****	**0.745*****	**0.726*****	**0.724*****
	(2.98)	**(3.35)**	**(3.26)**	**(3.24)**	**(3.05)**	**(3.03)**

Panel C. Value-weighted Portfolio Results on *NRDret*(LSA 100)

	(1) Excess Returns	(2) CAPM Alpha	(3) 3-Factor Alpha	(4) 4-Factor Alpha	(5) 5-Factor Alpha	(6) 6-Factor Alpha
1	0.395	0.215	0.188	0.207	0.372	0.381
	(0.75)	(0.40)	(0.35)	(0.39)	(0.67)	(0.69)
2	0.766*	0.624	0.616	0.635	0.682	0.692
	(1.70)	(1.38)	(1.37)	(1.41)	(1.45)	(1.47)
3	0.772**	0.664*	0.621*	0.635*	0.787**	0.795**
	(2.12)	(1.81)	(1.69)	(1.73)	(2.07)	(2.09)
4	0.956*	0.796	0.782	0.808	0.928*	0.942*

	(1.94)	(1.60)	(1.58)	(1.64)	(1.80)	(1.83)
5	0.973***	0.896**	0.875**	0.893**	0.103***	0.103***
	(2.75)	(2.50)	(2.44)	(2.50)	(2.76)	(2.80)
5-1	**0.578***	**0.681****	**0.688****	**0.686****	**0.653****	**0.653***
	(1.84)	**(2.15)**	**(2.17)**	**(2.16)**	**(1.98)**	**(1.97)**