NBER WORKING PAPER SERIES

LAZY PRICES

Lauren Cohen
Christopher Malloy
Quoc Nguyen

Working Paper 25084
http://www.nber.org/papers/w25084

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge MA 02139
September 2018

We would like to thank Christopher Anderson, Ulf Axelson, Nick Barberis, John Chalmers, Kent Daniel (discussant), Karl Diether, Irem Dimerci, Joey Engelberg, Umit Gurun, Gerald Hoberg (discussant), Xing Huang (discussant), Hunter Jones, Bryan Kelly, Jonathan Karpoff, Dana Kiku (discussant), Patricia Ledesma, Dong Lou, Asaf Manela, Ernst Maug, Craig Merrill, Mike Minnis (discussant), Toby Moskowitz, Peter Nyberg (discussant), Cesar Orosco (discussant), Chris Parsons (discussant), Frank Partnoy, Mitchell Petersen, Christopher Polk, Taylor Nadauld, Krishna Ramaswamy, Samantha Ross, Alexandra Niessen-Ruenzi, Stefan Ruenzi, Mark Seasholes, Dick Thaler, Pietro Veronesi, Sunil Wahal, Daniel Weagley (discussant), Hongjun Yan (discussant), Luigi Zingales, and seminar participants at Arizona State University, Brigham Young University, University of California at San Diego, University of Chicago, DePaul University, University of Edinburgh, Fuller and Thaler Asset Management, Hong Kong University, Hong Kong University of Science and Technology, University of Kansas, London Business School, London School of Economics, University of Mannheim, McGill University, Northwestern University, University of Oregon, University of San Diego, Shanghai Advanced Institute of Finance (SAIF), University of South Carolina, Temple University, Tsinghua PBC School of Finance, University of Washington, Washington State University, Yale University, Yeshiva University, American Finance Association Meetings, Barclays Quantitative Investment Conference, Ben Graham Centre for Value Investing at the Ivey Business School at Western University Symposium on Intelligent Investing, Chicago Booth Asset Pricing Conference, Chicago Quantitative Alliance (CQA), Macquarie Global Quantitative Conference, Conference on Professional Asset Management at Rotterdam University, Geneva Finance Research Institute (GFRI), Q Group Quantitative Investment Conference, Rodney White Conference on Financial Decisions and Asset Markets, the Public Company Accounting Oversight Board, and the United States Securities and Exchange Commission for incredibly helpful comments and discussions. We are grateful for funding from the National Science Foundation. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Lazy Prices
Lauren Cohen, Christopher Malloy, and Quoc Nguyen
NBER Working Paper No. 25084
September 2018
JEL No. G02,G12,G14

ABSTRACT

Using the complete history of regular quarterly and annual filings by U.S. corporations from 1995-2014, we show that when firms make an active change in their reporting practices, this conveys an important signal about future firm operations. Changes to the language and construction of financial reports also have strong implications for firms' future returns: a portfolio that shorts "changers" and buys "non-changers" earns up to 188 basis points in monthly alphas (over 22% per year) in the future. Changes in language referring to the executive (CEO and CFO) team, regarding litigation, or in the risk factor section of the documents are especially informative for future returns. We show that changes to the 10-Ks predict future earnings, profitability, future news announcements, and even future firm-level bankruptcies. Unlike typical underreaction patterns in asset prices, we find no announcement effect associated with these changes—with returns only accruing when the information is later revealed through news, events, or earnings—suggesting that investors are inattentive to these simple changes across the universe of public firms.

Lauren Cohen
Harvard Business School
Baker Library 273
Soldiers Field
Boston, MA 02163
and NBER
lcohen@hbs.edu

Quoc Nguyen
University of Illinois Chicago
University Hall
Department of Finance (MC 168)
601 S. Morgan St.
Chicago, IL 60607
quoc@uic.edu

Christopher Malloy
Harvard Business School
Baker Library 277
Soldiers Field
Boston, MA 02163
and NBER
cmalloy@hbs.edu

An internet appendix is available at: http://www.nber.org/data-appendix/w25084/

In a Grossman and Stiglitz (1976) world, agents are compensated for the marginal value of the information they collect, process, and impound into prices. While this model is static, the dynamics of these underlying processes have drastically changed for investors over time. Information production and dissemination have seen a drastic drop in cost over the past three decades. With this drop in cost, there has been a paired rise in the amount of information being produced, making the search and processing problem more complex. If investors have not kept up with the increasing magnitude and complexity of these changes, we might see simple instances of disclosed information going unattended to by even these Grossman-Stiglitz investors.

In this paper, we use the laboratory of firm annual statements to examine this tension. Prior literature has documented that while investors at one time responded contemporaneously to financial statement releases that contained large changes, over time this announcement effect has attenuated (Brown and Tucker, 2011 and Feldman et al., 2010). This prior literature thus concluded that changes to 10-K documents have become "less informative" over time.[1] While we replicate this fact that there is no significant announcement effect associated with changes to regular filings, we show that this misses a large and critically important component of these changes' impact on asset prices.

Namely, the lack of announcement returns is *not* due to financial statements becoming less useful over time. Instead, it is because investors are missing these subtle but important signals from annual reports at the time of the releases, perhaps due to their increased complexity and length.[2] When you isolate changes to corporate reports using our

[1] Note that while Feldman et al. (2010) find a modest contemporaneous and predictive effect of changes in sentiment in the MD&A section on stock returns, Brown and Tucker (2011) subsequently argue that announcement effects related to document changes have diminished over time, consistent with a decline in the informativeness of corporate filings.

[2] Also note that Loughran and McDonald (2017) point out that the average publicly traded firm has their

approach, one can see that document changes do impact stock prices in a large and significant way, but this happens with a lag: investors only *gradually* realize the implications of the news hinted at by document changes, but this news eventually does get impounded into future stock prices and future firm operations. Thus the message from our paper is quite different, in that our results point to a large amount of rich information that is being hidden in the 10-Ks, and that investors are missing (and continue to miss, even today), rather than the conclusion that corporate documents are becoming less informative and less useful to investors in today's capital markets. Indeed the findings in this paper indicate an extreme, broad-based form of investor inattention to an item that is foundational to the corporate reporting process—the quarterly and annual reports--and which leads to large return predictability.

As two motivating empirical facts for the increasing difficulty a Grossman-Stiglitz investor faces in the collection and processing of value-sensitive information, consider Figure 1. Figure 1 plots the simple average size of a firm's annual financial statement (10-K) as measured by the number of text words – so stripping out tables, ASCII embedded information, .jpeg files, etc.; and just focusing solely on actual text - over our sample period. From Panel A of Figure 1, one can see that over the last roughly 20 years, the length of the average 10-K has grown drastically – with the present-day 10-K being roughly 6 times as long as that in 1995.[3] Panel B shows that over this same time period, the number of textual changes[4] has also grown substantially (by over 12 times). Thus, not only are 10-

annual report downloaded from the SEC's website only 28.4 times by investors immediately after 10-K filings, suggesting that most investors may not be carefully examining the filings to begin with. Of course, investors may be accessing this information from other sources (Bloomberg, CapIQ, etc.), rather than solely through the SEC's website.

[3] Note that Li (2008) also documents this regularity.

[4] Here the number of textual changes is defined simply as the number of instances that a piece of text was removed, added, or modified as captured by Microsoft Word's "Compare Documents" function (Microsoft

Ks becoming longer over time (containing more text), but they also contain significantly more changes year-to-year. We use this laboratory to explore how investors have respond to these changes in information delivery – and how this information eventually translates into stock prices and firm operations.

To better understand our approach, consider the example of Baxter International, Inc. Baxter is a bioscience and medical products firm headquartered in Deerfield, IL. The firm was founded in 1931, trades on the NYSE (ticker: BAX), and is a member of the S&P 500 Index. The company had historically had Annual Reports (10-K's) that were very similar across time, but something changed in 2009. This can be seen in Figure 2, which shows the similarity between Baxter's 10-K from year to year.

What caused Baxter's 2009 10-K[5] to veer from the prior year in terms of the language used and information given? Figure 3 shows a number of news headlines that flooded the media in the months following the release of the 10-K (Baxter's 10-K was publicly released on February 23, 2010). For instance, a *New York Times* article published on April 24, 2010 reported that the FDA was clamping down on medical devices – in particular, on automated IV pumps used to deliver food and drugs. From the article: "*The biggest makers of infusion pumps include Baxter Healthcare of Deerfield, Ill.; Hospira of Lake Forest, Ill.; and CareFusion of San Diego.*" The article went on to quote an FDA official commenting that the new, tighter regulations would slow down the FDA approval process for automated pumps. Then, on May 4th (just 10 days later) the *New York Times* reported that the FDA had imposed a large recall on Baxter: "*Baxter International is* ***recalling*** *its* ***Colleague infusion pumps*** *from the American market under an agreement with federal regulators that*

Word Tools menu, point to Track Changes, and then click Compare Documents).

[5] Note that here we are referring to the 2009 calendar year 10-K that was released in February 2010.

sought to fix problems like battery failures and software errors."[6]

Moreover, the stock returns of Baxter International moved substantially surrounding the *New York Times* articles. In the two-week period around the articles, Baxter's price burned down more than -20%. This is shown in Figure 4, which also shows that the price remained depressed, not reverting over the subsequent 6-month period. In contrast, we see no significant reaction to Baxter's own disclosure of its 10-K on February 23, 2010, nearly two months before the news articles were published.

The question then is whether these two information releases were at all linked – i.e., could something about the changes to the 10-K from Figure 2 (and reported 2 months before) have hinted at the portending news regarding the automated pump issue. Figure 5 gives some suggestive evidence in this direction by showing the incidence of keywords in Baxter's 10-Ks over time related to the FDA's clamp-down, and the recall of Baxter's *Colleague Pump*. Figure 5 shows that Baxter's usage of these words spiked in their 2010 report relative to previous years. In particular Baxter's 2009 filing showed a 71% increase in mentions of the "*FDA*," a 50% increase in the usage of the term "*Recall*," and a 182% increase in the mentions of "*Colleague Pump*." Figure 6 then shows more detailed, suggestive evidence on this point. It shows a number of parallel passages: the 2009 version of the passage vs. the 2008 version, showing examples of Baxter's increased mentioning of these items.[7]

From Figure 6, for instance, one can see that Baxter changed the passage:

"It is possible that additional charges related to COLLEAGUE may be required in future

[6] The link to Baxter's 2009 full 10-K and accompanying exhibits, along with links to the full-length articles and excerpts, are included in Figure 2.

[7] See also Figure A-1 in the Appendix, for additional text passages related to this Baxter example. In addition, Appendix Figure A-2 presents selected text passages from successive 10-Ks from another example company (Herbalife).

periods." [2008]

to:

"It is possible that ***substantial*** *additional charges,* ***including significant asset impairments, related to COLLEAGUE may be required in future periods.*** *"* [2009]

Along with adding this passage to their 2009 10-K:

"The sales and marketing of our products and our relationships with healthcare providers are under increasing scrutiny by federal, state and foreign government agencies. The FDA, the OIG, the Department of Justice (DOJ) and the Federal Trade Commission have each increased their enforcement efforts..."

Circling back, would being attentive to the changes in the 10-K have made a difference to investors in Baxter? Going back to Figure 4, the answer appears to be yes. Not only did the price of Baxter not move at all around the public filing of the 10-K (February 23, 2010), but the price did not move for the next 2 months – until the news was reported in the *New York Times* on April 23. Reading and reacting to these negative changes by shorting Baxter at any point in the two months leading up to the NYT article would have allowed an investor to capture over 30% in returns in the month following the news' release.

We demonstrate that this pattern of behavior, investor response, subsequent events, and return evolution are systematic across the entire cross-section of U.S. publicly traded firms from 1995 to 2014.

We first show that firms that change their reports experience significantly lower future stock returns. In particular, a portfolio that goes long "non-changers" and short "changers" earns a statistically significant 34-58 basis points per month – up to 7% per year (t=3.59) - in value-weighted abnormal returns over the following year. These returns

continue to accrue out to 18 months, and do not reverse, implying that far from overreaction, these changes imply true, fundamental information for firms that only gets gradually incorporated into asset prices in the months after the reporting change. As all publicly traded firms are mandated to file 10-Ks (and 10-Qs), the sample over which we show these abnormal returns is truly the universe of firms (not a small, illiquid or otherwise selected subset).

We show that these findings cannot be explained by traditional risk factors, well-known predictors of future returns, unexpected earnings surprises, or news releases that coincide with the timing of these firm disclosures. Moreover, we find an economically and statistically zero announcement day return (much like Baxter) in the full sample. This is in contrast to a gradual information diffusion type explanation that is consistent with the empirical pattern of many other regularities (e.g., post-earnings announcement drift, momentum, etc.), in which there is an immediate large response followed by a much more modest – but persistent – drift in the same direction. Instead, the pattern we document is more consistent with investors simply failing to account for – or be attentive – to the systematic and rich information contained in simple changes to a firm's annual reports. Their stock prices exhibit little to no reaction at the time of public filing by the firm, even though there is a robust and systematic relationship (whereby changes predict future negative returns and negative real operational realizations) – with the information only being impounded into price in the future.

Next, we explore the mechanism at work behind these return results. We show that firms' reporting changes are concentrated in the management discussion (MD&A) section, which is the section of the reports where management has the most discretion and flexibility in terms of content. However, in terms of return-rich content, we find that while changes

in MD&A section wording do predict large and significant abnormal returns, changes in text in the Risk Factors section are even more informative for stock returns. For instance, the 5-factor alpha on (Non-Changers – Changers) particularly in this risk factors section is over 188 basis points per month (t=2.76), or over 22% per year. Further, we find that changes in language referring to the executive (CEO and CFO) team, and about litigation and lawsuits, are especially informative for future returns, as is the increased usage of so-called "negative sentiment" words. For instance, changes focused on litigation and lawsuits underperform the non-changers by over 71 basis points per month, or over 8.5% per year (t=3.29).

We then turn to measures of real activity and show that changes to the 10-Ks predict future earnings, profitability, future news announcements, and even future firm-level bankruptcies. Moreover, much like return realizations, these appear to be largely unanticipated, as the real operational changes are not taken into account by analysts covering the firm – resulting in the 10-K changes significantly predicting future negative earnings surprises and negative cumulative abnormal returns (CARs) around these events.

Lastly, theory does not predict that changes must lead to negative returns. It may be just as plausible, *ex ante*, that firms make positive changes in their 10-K text which are ignored by investors, and then lead to positive realizations in returns and firm outcomes. Thus, the loading on unsigned text "changes" would be ambiguous. We have two pieces of evidence speaking to the strong observed negative relationship we document in the data with respect to returns and future outcomes. First, when we use natural language processing (NLP) textual signing of the underlying text of the changes, we note that 86% of changes consist of "negative" sentiment changes. When we segregate out the 14% of changes that are "positive" changes, we do find that they predict significantly positive

returns in the future. Second, and perhaps causing part of the disproportionate (86-14) ratio of negative (bad news portending) changes, is that class-action lawsuits have been dominated by claims of the omission of negative news to existing shareholders (i.e., short-sellers have had more success suing firms for not properly disclosing material positive information in a timely fashion). This would asymmetrically increase the risk of failing to report negative information, leading to the asymmetric realizations of changes observed.

Lastly, we do a number of robustness checks across firm size, time, industry, firm-events, etc. The effect that we document is not driven by any of these factors. In particular, it is not something about special firm events (e.g., we exclude periods of M&A, SEOs, or other large firm events that might necessitate changing of the 10-K) or about certain industries, types, or characteristics of firms. In addition, this does not appear to be a function of transaction costs or limits to arbitrage. The return results we document have the following characteristics: they accrue over *months* following the release of the 10-K (so no high frequency trading is needed); the portfolios have very modest turnover (around the infrequent reporting dates); the effects show up in value-weighted returns across the universe of all publicly traded firms (and so are not concentrated in small firms); the average "changer" firm (to be shorted) is actually larger than the average long at $3.5 B market cap (vs. $2.5 B), and the average changer firm has relatively modest shorting fees – again actually less costly to short than the average stock in the long portfolio.

Our findings are also not driven solely by changes in the length of these documents. As mentioned above, while 10-Ks have seen a large increase in length over time, when we control for document length and document length changes, the impact of the changes we measure in the filings is a large and significant predictor of returns. In sum, controlling for these along with other characteristics and events (e.g., issuance, accruals, etc.), the act of

substantively altering a firm's 10-K remains an economically large and statistically robust predictor of future returns and real firm operating changes.

Stepping back, these results in some manner require a differential "laziness" of investors with respect to text compared with numerical financial statement entries. In particular, nearly every table in financial statements is shown with the current year's numbers along with a series of past years' comparable reported numbers. For instance, a sales revenue figure of 1.5 billion dollars would mean little without the context of comparing it prior years' sales revenues. In contrast, investors do not appear to be doing the same comparison of this year's text to last year's. That simple comparison, as we show throughout the paper, contains rich information for the future of a firm's operations.

In order to parameterize and examine the actions of investors in even more depth, we would ideally like to measure times at which investors allocate more attention to a firm's 10-Ks, and in particular changes to 10-Ks. While this has historically been difficult (to impossible) to measure, we attempt to do so using novel data from the Securities and Exchange Commission (SEC). Namely, we filed a Freedom of Information Act (FOIA) request with the SEC in order to obtain data documenting: i.) every downloaded filing from the SEC website's EDGAR downloadable service, ii.) the exact time-stamp of when the filing was downloaded, and iii.) the downloader's (partially masked for anonymity) IP address.[8] From this, we construct a panel dataset of 10-K downloading activity (which filings and when) by each investor over time. We use this data to identify 10-K releases (e.g., Apple in 2011) in which a large percentage of investors download not only the current year's 10-K, but where these *same* investors also download the prior year's 10-K in tandem.

[8] Note that this data is now publicly available on an ongoing basis.

These investors plausibly have a higher likelihood of wanting to compare the two – given their joint downloading - than situations in which the majority of investors are simply downloading this year's 10-K filing alone. We find that when more investors are potentially comparing 10-Ks, and possibly "paying attention" to changes – downloading both this year's and last year's 10-K – this attenuates the key return predictability effects we document in this paper, consistent with inattention being a mechanism behind our documented findings.[9] Finally, we investigate the nature of this inattention and show that investors have an easier time digesting qualitative changes when they are explicitly drawn to these changes through comparative statements included in the text (e.g., with statements such as "relative to prior year EBITDA" or "compared to last year"). In this sense, our paper new, granular evidence on the origins and characteristics of investor inattention by pinpointing which specific phrases and language patterns can help investors improve their ability to process textual information.

The remainder of the paper is organized as follows. Section I provides a brief background and literature review. Section II describes the data we use and explores the particular construction of firms' annual and quarterly reports. Section III examines the impact of these choices, and Section IV explores the mechanism driving our results in more detail. Section V concludes.

I. Background and Related Literature

Our paper contributes to several growing literatures, including (but not limited to):

[9] Note, however, that in these situations where investors are presumably being more attentive (by simultaneously downloading and comparing year-on-year documents), we do find that the *short-run* announcement effects are *more* pronounced, likely because investors immediately detect the changes to the reports and quickly impound these changes into stock prices.

a) the broad topic of underreaction in stock prices and the impact of investor inattention; b) the use of textual analysis in finance and accounting; and c) the information content of firms' disclosure choices.

The magnitude and nature of our return predictability results add new evidence and much-needed granularity to the existing stock price underreaction and inattention literature. As described in Tetlock (2014)'s review article, several papers document that underreaction is strongest when investors fail to pay attention to informative content. See, for example, Tetlock (2011), who constructs measures of "stale" news stories and demonstrates that investors overreact to stale information (and correspondingly, underreact to novel information). In addition, Da, Engelberg, and Gao (2011, JF) use Google search activity to pinpoint retail investor attention, while Ben-Raphael, Da, and Israelson (2017) measure institutional attention using Bloomberg search activity; the latter shows that stock price drift is most pronounced for stocks with the least amount of institutional attention. Another novel measure of attention is employed in Engelberg, Sasseville, and Williams (2012), who show that spikes in TV ratings (presumably driven by retail investors) during the Jim Cramer "Mad Money" show are linked to overreaction in stock prices for the companies recommended during the show. By contrast, what we document in this paper is an acute form of investor inattention that impacts a large cross-section of firms, is centered on the most important corporate disclosure that firms make, and which leads to large return predictability. Further, we use novel data from the SEC log files to demonstrate that variation in attention to this exact same item (the annual report) produces variation in these return predictability patterns. And finally, we dig into the nature of this inattention and show that investors have an easier time digesting qualitative changes (i.e., changes in text, as opposed to numbers) when they are explicitly drawn to

these changes through comparative statements included in the text; but when such comparisons are not included, investors simply cannot decipher meaningful changes to these documents. So it is not merely the difference between quantitative and qualitative information that matters for investors (as in Engelberg (2008)), but also the way in which that qualitative information is constructed and presented.[10] In these ways, our paper helps to micro-found some of the more general evidence on inattention and underreaction in stock prices by clarifying exactly what it is that investors fail to recognize.

In attempting to pinpoint textual changes at the document level, our paper also contributes to the large and fast-growing field of textual analysis. As a result of increased computing power and advances in the field of natural language processing, many recent papers have tried to employ automated forms of textual analysis to answer important questions in finance and accounting; Loughran and McDonald (2016) provide a helpful survey of some of these papers. Most relevant to our study are the articles that analyze the link between textual information in firm disclosures (such as the 10-Ks and 10-Qs that feature in our analysis) and firm behavior and performance.[11] For example, Li (2008) employs a form of textual analysis and finds that the annual reports of firms with lower earnings (as well as those with positive but less persistent earnings) are harder to interpret. Li (2010a) also finds that firms' tone in forward-looking statements in the MD&A section can be used to predict future earnings surprises. Meanwhile, Nelson and Pritchard (2007)

[10] Note that we also explicitly show (in Tables V, VI and Appendix Table A-9) that our document similarity measure is distinct from previously used textual metrics that focus on sentiment and/or negative words (such as those in Tetlock, Saar-Tsechansky, and Macskassy (2008), or Loughran and McDonald (2011)).

[11] Note that before the advent of advanced computing techniques, several papers focused on hand-coded analysis of disclosure content, for example in the management discussion (MD&A) section of annual reports (see Bryan, 1997, and Rogers and Grant, 1997). Others used survey rankings in order to quantify the level of disclosure (see Clarkson, Kao, and Richardson, 1999, and Barron, Kile, and O'Keefe, 1999) in the MD&A sections. See Cole and Jones (2005) and Feldman et al. (2010) for a survey of the earlier evidence.

explore the use of cautionary language designed to invoke the safe harbor provision under the Private Securities Litigation Reform Act of 1995, and find that firms that are subject to greater litigation risk change their cautionary language to a larger degree relative to the previous year; but after a decrease in litigation risk, they fail to remove the previous cautionary language. In addition, Feldman et al. (2010) find that a positive tone in the MD&A section is associated with modestly higher contemporaneous and future returns, and that an increasingly negative tone is associated with lower contemporaneous returns.[12] In our paper, we show that the document similarity measure we employ predicts future returns even controlling for any impact of disclosure sentiment.

Finally, a handful of additional papers explore other aspects of firm-level annual reports, in studying the impact of different types of corporate disclosure. For instance, Lee (2012) finds that less of the earnings-related information is incorporated into a firm's stock price during the three days following the 10-Q filings for firms with longer or less readable 10-Q. Meanwhile Dyer, Lang, and Stice-Lawrence (2017) show that 10-Ks have become longer and more complex, and examine some of the reasons behind these trends. And closest to our paper is perhaps Brown and Tucker (2011), who focus on year-on-year changes in the text of the MD&A section, and find that changes in the MD&A section are related to future operating changes in the business (e.g., accounting-based measures of performance, as well as liquidity measures); they also find that contemporaneous returns around 10-K filing dates are increasing in changes to MD&A. Importantly, Brown and Tucker (2011) report that "While MD&A disclosures have become longer over time, they have become more like what investors saw in the previous year... Moreover, we find that

[12] See also Muslu et al. (2015), Li (2011), Loughran and McDonald (2016), and Das (2014) for a survey of various textual analysis approaches.

the price responses to MD&A modifications have weakened over time... suggesting a decline in MD&A usefulness." What we show in this paper, however, is that changes in 10-Ks are actually remarkably useful for investors, as they predict large negative returns in the future. So while we confirm their finding from recent years that announcement effects associated with document changes are close to zero,[13] this is *not* because they have become less useful. Rather, it is because investors are missing these subtle but important signals from annual reports at the time of the releases, perhaps due to their increased complexity and length. Isolating changes using our approach, these changes have powerful predictability for future asset prices. Far from them becoming less informative as past literature has posed, our results point to a large and significant role of 10-Ks and 10-Qs through the present day. And yet the rich information in these documents – and the changes to the information conveyed – appears to be largely missed by investors. Instead price revelation occurs only gradually over time, with both asset prices and real activity reacting slowly over the next 6-12 months to the document changes.

II. Data and Summary Statistics

We draw from a variety of data sources to construct the sample we use in this paper. We download all complete 10-K, 10-K405, 10-KsB and 10-Q filings from the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) website[14] from 1995 to 2014. All complete 10-K and 10-Q filings are in HTML text format and contain an aggregation of all information that are submitted with each firm's file, such as exhibits, graphics, XBRL

[13] Note that our results pertain to the entire 10-K, but we can confirm that the announcement effects associated with changes to the various sub-sections (including the MD&A) are also statistically indistinguishable from zero.
[14] (https://www.sec.gov/edgar/)

files, PDF files, and Excel files. Similar to Loughran and McDonald (2011), we concentrate our analysis on the textual content of the document. We only extract the main 10-K and 10-Q texts in each document and remove all tables (if their numeric character content is greater than 15%), HTML tags, XBRL tables, exhibits, ASCII-encoded PDFs, graphics, XLS, and other binary files.[15]

We obtain monthly stock returns from the Center for Research in Security Prices (CRSP) and firms' book value of equity and earnings per share from Compustat. We obtain analyst data from the Institutional Brokers Estimate System (IBES). We obtain sentiment category identifiers from Loughran and McDonald (2011)'s Master Dictionary.[16]

We measure the quarter-on-quarter similarities between 10-Q and 10-K filings using four different similarity measures taken from the literature in linguistics, textual similarity, and natural-language processing (NLP): i.) cosine similarity, ii.) Jaccard similarity, iii.) minimum edit distance, and iv.) simple similarity. We describe each measure, and its respective calculation, below.

The first measure is called the cosine similarity, which has also been used in the finance literature by Hanley and Hoberg (2010). It is computed between two documents - D_1 and D_2 - as follows. Let D_{S1} and D_{S2} be the set of terms occurring in D_1 and D_2, respectively. Define T as the union of D_{S1} and D_{S2}, and let t_i be the i^{th} element of T. Define the term frequency vectors of D_1 and D_2 as:

$$D_1^{TF} = [nD_1(t_1), nD_1(t_2), \dots, nD_1 t_N];\ D_2^{TF} = [nD_2(t_1), nD_2(t_2), \dots, nD_2 t_N]$$

where $nD_k(t_i)$ is the number of occurrences of term t_i in D_k. The cosine similarity

[15] Bill McDonald provides a very detailed description on how to strip 10-K/Q down to text files: http://sraf.nd.edu/data/stage-one-10-x-parse-data/
[16] http://sraf.nd.edu/textual-analysis/resources/

between two documents is then defined as:

$$Sim\_Cosine = \frac{D_1^{TF} \cdot D_2^{TF}}{\left||D_1^{TF}|\right| \times \left||D_2^{TF}|\right|}$$

where the dot product, $\cdot$, is the scalar product and norm, $\left|| \quad ||\right|$, is the Euclidean norm.

For a textual and numerical example, consider these three short texts:

D_A: We expect demand to increase.

D_B: We expect worldwide demand to increase.

D_C: We expect weakness in sales.

It is easy to see that D_A is very similar to D_B and that D_A is more similar to D_B than it is to D_C. The cosine similarity of $D_A$and D_B is computed as follow. First, the union $T(D_A, D_B)$ is:

$$T(D_A, D_B) = [\text{we, expect, worldwide, demand, to, increase}]$$

The term frequency vectors of D_A and D_B are:

$$D_A^{TF} = [1,1,0,1,1,1];\ D_B^{TF} = [1,1,1,1,1,1]$$

The cosine similarity score of D_A and D_B is therefore

$$Sim\_Cosine(D_A, D_B) = \frac{(1\times1+1\times1+0\times1+1\times1+1\times1+1\times1)}{\left(\sqrt{1^2+1^2+1^2+1^2+1^2}\right)\times\left(\sqrt{1^2+1^2+1^2+\ 1^2+1^2+1^2}\right)} = \ 0.91$$

Similarly, the cosine similarity of D_A and D_C is computed as follow. The union $T(D_A, D_C)$ of D_A and D_C is:

$$T(D_A, D_C) = [\text{we, expect, demand, to, increase, weakness, in, sales}]$$

The term frequency vectors of D_A and D_C:

$$D_A^{TF} = [1,1,1,1,1,0,0,0];\ D_C^{TF} = [1,1,0,0,0,1,1,1]$$

The cosine similarity score of DA and D_C is therefore:

$$Sim\_Cosine(D_A, D_C) = \frac{(1\times1+1\times1+1\times0+1\times0+1\times0+0\times1+0\times1+0\times1)}{\left(\sqrt{1^2+1^2+1^2+1^2+1^2}\right)\times\left(\sqrt{1^2+1^2+1^2+1^2+1^2}\right)} = \ 0.40$$

Clearly, D_A is more similar to D_B than to D_C and the cosine similarity measures captures this difference in similarity. The Jaccard similarity measure uses the same term frequency vectors/sets as in the cosine similarity measure, and is defined as:

$$Sim\_Jaccard = \frac{\left|D_1^{TF} \cap D_2^{TF}\right|}{\left|D_1^{TF} \cup D_2^{TF}\right|}$$

In other words, the Jaccard similarity is the size of the intersection divided by the size of the union of the two term frequency sets; also note that the Jaccard measure is binary (each word is counted only once in a given set) while the cosine similarity measure includes frequency (and hence includes counts of each word).

In the same textual examples $D_A, D_B,$ and D_C as above, the Jaccard similarities are:

$$Sim\_Jaccard(D_A, D_B) = \frac{|\{\text{we, expect, demand, to, increase}\}|}{|\{\text{we, expect, worldwide, demand, to, increase}\}|} = \frac{5}{6} = 0.83$$

$$Sim\_Jaccard(D_A, D_B) = \frac{|\{\text{we, expect}\}|}{|\{\text{we, expect, demand, to, increase, weakness, in, sales}\}|} = \frac{2}{8} = 0.25$$

The third similarity measure we employ is called $Sim\_MinEdit$ and is computed by counting the smallest number of operations required to transform one document into the other. In the same textual examples $D_A, D_B,$ and D_C as above, transforming D_A to D_B only requires adding the word "*worldwide,*" while transforming D_A to D_C requires deleting 3 words "*demand,*" "*to,*" and "*increase*" and adding 3 words *"weakness," "in," and "sales"*.

Finally, the fourth similarity measure we use is called Sim_Simple, and uses a simple side-by-side comparison method. We utilize the function "Track Changes" in Microsoft Word or the function "diff" in Unix/Linux terminal to compare the old document D_1 with the new document D_2. We first identify the changes, additions, and deletions while comparing the old document with the new document. We first count the number of words in those changes, additions, and deletions and normalize the total count by the average size

of the old document D_1 and the new document D_2.

$$c = [additions + deletions + changes] \; / \; [(Size\, D_1 + Size\, D_2)/2]$$

In order to obtain a similarity measure that has values between $[0, 1]$, where 1 means the two documents are identical, as in the prior three similarity measures, we then normalize by feature scaling c to compute Sim_Simple as follows:

$$Sim\_Simple = [c_{max} - c]/c_{max}$$

Note that every annual 10-K report contains 15 schedules and every quarterly 10-Q report contains 10 schedules. The common schedules for both 10-K and 10-Q reports are: Management's Discussion and Analysis of Financial Condition and Results of Operations, Risk Factors, Legal Proceedings, Quantitative and Qualitative Disclosure about Market Risks, Control and Procedures, and Other information. These schedules (or "Items") of 10-K and 10-Q reports are listed in Figure 8. We identify the textual content of each schedule by capturing regular expressions that contain the word "item" and the schedule name. Since the labels for schedules are very inconsistent across filings, we process all 10-K and 10-Q filings many times to capture those exceptions. First, we use regular expressions to capture the most common structure for schedule titles: lines that start with "Item" + "a number" + "title name." Then we start capturing exceptions to the common rule, for example, lines that only has "Item" + "a number," lines with only "number" + "title names," etc., while also making sure that each schedule is captured exactly once. We repeat that process many times and incorporate new exceptions each time.

Table I presents summary statistics from our final dataset, which consists of all 10-Ks and 10-Qs downloaded from the SEC Edgar websites from 1995 to 2014. *Document Size* refers to the number of characters in each report, and the *Size of Change* refers to the number of characters that change relative to a prior report (in the case of a 10-K, the

change is measured relative to last year's 10-K, and in the case of a 10-Q, the change is measured relative to the same quarter's 10-Q in the prior year). Panel A of Table I shows that the average 10-K contains 308,633 characters, while the average 10-Q contains roughly one-third as many characters (114,848).

For some of our tests of the mechanism, we also draw sentiment category identifiers and word lists (e.g., measures of negative words, positive words, uncertainty, litigiousness, etc.) from Loughran and McDonald (2011)'s Master Dictionary.[17] In Panel A, the *Sentiment of Change* refers to the number of positive words minus the number of negative words normalized by the size of the change. The *Uncertainty of Change* and the *Litigiousness of Change* are the number of words categorized by "uncertainty" and "litigiousness," respectively, normalized by the size of the change. Finally, we also parse 10-K/Q documents for mentions of CEO or CFO turnover and define two indicator variables *Change CEO* and *Change CFO* as indicator variables set equal to one if the 10-K or 10-Q mentions a change in CEO or change in CFO, respectively. More specifically, we search for instances where a word from the set {appoint, elect, hire, new, search} and a word from the set {CEO, CFO, Chief Executive Officer, Chief Financial Officer} appear within 10 words of each other. Table I shows that CEO and CFO changes are mentioned in roughly 2-5% of the reports, on average.

Panel B of Table I presents summary statistics of the four similarity measures. Each of the measures ranges from 0 to 1, but the ranges differ across the measures. For example, the distribution of the *Sim_Cosine* measure is fairly narrow, with a mean of 0.86 and a

[17] These words are available at: https://sraf.nd.edu/textual-analysis/resources/.

standard deviation of 0.21, while the distribution of the *Sim_Simple* measure is centered at a much lower level, with a mean of 0.12 and a standard deviation of 0.12. Recall that higher values indicate a higher degree of document similarity across years between the 10-Ks (or 10-Qs), while lower values indicate more changes across documents. Also note that we winsorize any outliers of these measures (at the 1st and 99th percentiles) before including them in our subsequent analyses.

Panel C of Table I reports the correlations between the measures. All four measures are strongly positively correlated with each other, although the *Sim_Simple* measure is correlated only 0.25 with the *Sim_Cosine* measure; all of the other pairwise correlations between the four measures exceed 0.5.

III. The Implications of Changes in Reporting Behavior

In this section we examine the implications of firms' decisions to change the language and construction of their SEC filings. In particular, we explore the nature of these changes and their implications for firms' future actions and outcomes.

We begin by analyzing the future stock returns associated with firms who change their reports substantially, versus those who do not. First, we compute standard calendar-time portfolios, and then we control for additional determinants of returns by employing Fama-MacBeth monthly cross-sectional regressions.

A. Calendar-Time Portfolio Returns

For each of the four similarity measures described in the previous section, we

compute quintiles each month based on the prior month's distribution of similarity scores across all stocks. For firms with a fiscal year-end in December, we use the following reports: for calendar quarter Q1, we use the release of a firm's 10-Q, which generally occurs in April or May; for calendar quarter Q2, we use another release of a firm's 10-Q, which generally occurs in July or August; for calendar quarter Q3, we use another release of a firm's 10-Q, which generally occurs in October or November; and finally for the year-end results we use the release of the full-year 10-K, which typically occurs in February of March.[18] Similarity scores are computed relative to the prior year report that lines up in calendar time with the report in question (such that 2005 Q1 10-Qs are compared with 2004 Q1 10-Qs, for example).[19] Stocks enter the portfolio in the month after the public release of one of their reports, which induces a lag in our portfolio construction. Note that in all of our tests, firms are held in the portfolio for 3 months. Portfolios are rebalanced monthly, and the average monthly returns are reported in Table II.

Panel A of Table II presents equal-weighted calendar-time portfolio returns. Quintile 1 (Q1) refers to firms that have the least similarity between their document this year and the one last year; hence this portfolio consists of the "big changers." Quintile 5 (Q5) refers to firms that have the most similarity in their documents across years, and hence this portfolio represents the "little to no changers." Q5-Q1 represents the long-short (L/S)

[18] See Appendix Figure A-4 for a depiction of the average clustering of release dates, by month, for the 10-Ks and 10-Qs in our sample. Also note that for firms with "off-cycle" fiscal year-ends we simply use their reports in an analogous way to that presented here, but incorporating the different timing. E.g., firms with a fiscal-year end in June typically release their annual 10-Ks in August and September; and for the other 3 calendar quarters we would analyze their 10-Qs instead.

[19] Note that due to seasonality in sales and operations (and the associated discussion of those seasonal patterns in the company filings), the most comparable report for a given 10-K is the prior year 10-K (as opposed to the prior quarter 10-Q), and the most comparable report for a given 10-Q is the prior year 10-Q in that same quarter. As shown in Appendix Table A-13, when we restrict our sample to only look at year-on-year changes in the text of 10-Ks (and hence remove all the 10-Q changes), or look only at year-on-year changes in the text of 10-Qs (and hence remove all the 10-K changes), our main portfolio result from Table II is similar.

portfolio that goes long Q5 and short Q1 each month.

Panel A shows that this L/S portfolio earns a large and significant abnormal return, ranging in magnitude between 18-45 basis points per month. This result is unaffected by controlling for the 3 Fama-French factors (market, size, and value), or for two additional momentum and liquidity factors. This suggests that the return spreads we see between these portfolios are not driven by systematic loadings on commonly known risk factors. Notably, all 4 measures of similarity deliver this pattern, suggesting moreover that our results are not driven by the particular way we compute year-over-year changes in the documents. This finding indicates that firms that make significant changes to their disclosures in a given year experience lower future returns. Later in the paper we explore the possible mechanisms behind this return result.

Panel B of Table II then presents value-weight portfolio returns, computed as in Panel A except that each stock in the portfolio is weighted by its (lagged) market capitalization. Panel B shows that the value-weight portfolio returns are similar but somewhat larger in magnitude to the equal-weight results, with the value-weight L/S portfolio earning up to 58 basis points per month (t=3.59), depending on the similarity measure employed. In addition, Appendix Figure A-3 plots the annual time-series of excess returns documented in Table II; this figure shows that the number of positive excess return years is distributed quite evenly throughout the sample period, reaffirming the conclusion that the abnormal returns documented in this paper are not concentrated in just a few quarters or years.[20]

We explore the evolution of both the long and short legs of this portfolio using event-

[20] To further confirm that our results are not driven by a few special years of quarters, we also exclude the years 2000, 2001, 2008, and 2009 in our portfolio tests, and find that the abnormal returns documented in

time returns in Figure 7. As seen from the event-time returns in Figure 7, any positive alpha on the Q5 long side (the "little to no changers") quickly reverts to zero, while the negative alpha persists and increases up to 6 months out – never reversing. In particular, Figure 7-A explores the longer-term returns by computing the average cumulative abnormal return for each quintile portfolio sorted based on firms' similarity scores (here the *Sim_Jaccard* measure is used), for 1 month out to 6 months after portfolio formation. Figure 7-A shows that L/S returns accrue gradually over the course of the subsequent 6 months, and do not reverse. Additionally, the long-term poor performance of Q1 (the "changers") is particularly strong and persistent in this figure.

Figure 7-B then takes an even more granular look at the L/S return effect, by exploring the event-time announcement returns around the public release of these filings (from t-10 days out to t+10 days). Figure 7-B shows that – like the example of Baxter International, Inc. - there is no statistically or economically detectable effect around the announcement of these filings, but rather that the return effect we document in this paper accrues gradually over the course of the following 6 months and does not revert. Taken as a whole, Figure 7 suggests that the information contained in a firm's decision to significantly change its reporting practices has a long-lasting impact on firm value that does not accrue upon release of reports, but instead only gradually through price revelation over time.

B. Characteristics of Quintile Portfolios

The finding that a significant portion of the return spread documented in Table II

Table II are still large and significant.

and Figure 7 comes from the short side begs the question of the composition and characteristics of both sides of this L/S portfolio. For example, it could be the case that the short side simply contains a set of smaller firms that are difficult (and expensive) to short. Or perhaps, there is not significant turnover of small or illiquid stocks to trade. Both of these might make the returns we document fall within simple limits of arbitrage. Table III presents the average size, turnover, shorting costs (in basis points), and sentiment (as defined in Table I) for all five quintile portfolios. As Table III shows, there is little evidence that the short side contains an unusual set of firms on average; if anything, the firms in Q1 appear to be slightly larger, and have lower average shorting costs. The only notable difference appears to be in the Sentiment of the text of the firms' filings, a finding we explore in greater depth below. Moreover, given that turnover is so modest, that VW returns are actually a bit larger than EW, that our sample is the entire universe of publicly traded firms (i.e., we are not restricted to a small set of firms or industries as every publicly traded firm is mandated to file 10-Ks and 10-Qs) resulting in large diversified portfolios for each quintile, and that returns only accrue slowly over the following 6 months, we do not believe limits to arbitrage are a significant contributor to the return regularities we see.

C. Fama-MacBeth Regressions

We next run monthly Fama-MacBeth cross-sectional regressions of future individual firm-level stock returns on a host of known return predictors, plus our 4 similarity measures. As Table IV shows, each similarity measure is a positive and significant predictor of future stock returns, implying that firms who make large changes to their reports experience lower future returns. This result holds when we include a variety of additional return predictors

as well, including the following: last month's (or last quarter's) standardized unexpected earnings surprise (*SUE*); *Size,* the log market value of equity; *log(BM)*, the log book value of equity over market value of equity; *Ret(-1,0),* the previous month's return; and *Ret(-12, -2),* the cumulative stock return from month *t*-12 to month *t*-2. SUE is defined as the Compustat-based standardized unexpected earnings, and is computed as in Livnat and Mendenhall (2006), where Compustat-based Earnings Surprise is based on the assumption that EPS follows a seasonal random walk, where the best expectation of the EPS in quarter *t* is the firm's reported EPS in the same quarter of the previous fiscal year. In terms of magnitude, the coefficient on *Sim_Simple* in column 12 (=0.0292, *t*=2.11), for example, implies that for a one-standard deviation decline in a stock's document similarity across years, returns are 36 basis points lower per month in the future.

IV. Mechanism

In this section we explore the mechanism at work behind our key return results.

A. Explaining Changes in Reporting Behavior

We begin by regressing our similarity measures on a host of characteristics of the documents in question. The goal of this exercise is to better understand what helps explain changes in similarity across years for a given firm's document.

We construct a variety of measures based on specific words, as well as sentiment-type measures based on available word dictionaries. As noted above in our discussion of the summary statistics in Table I, we use sentiment category identifiers and word lists (e.g., measures of negative words, positive words, uncertainty, litigiousness, etc.) from Loughran

and McDonald (2011)'s Master Dictionary. Specifically, the variable *Sentiment of Change* refers to the number of positive words minus the number of negative words normalized by the size of the change; *Uncertainty of Change* and the *Litigiousness of Change* refer to the number of words categorized by "uncertainty" and "litigiousness," respectively, normalized by the size of the change; and *Change CEO* and *Change CFO* are indicator variables set equal to one if the 10-K or 10-Q mentions a change in CEO or change in CFO, respectively.

Table V shows the results of panel regressions of document similarity (here measured as *Sim_Simple*)[21] on these characteristics of the document, with firm and time fixed effects included, and clustering done at the firm level. Table V shows that lower similarity (i.e., more changes) across documents is associated with lower sentiment, higher uncertainty, more litigiousness, and more frequent mentions of CEO and CFO changes.[22] Each of these findings is highly statistically significant and suggests that the changes in reporting practices that we identify are associated with significant changes in the operations or prospects of the firm in question.

In Table VI we also explore the extent to which our return results are driven by other aspects of the filings, as opposed to our specific measure of changes in the similarity of year-over-year documents. For example, low sentiment itself, or just the length of the document, or just changes in the length of the document might be more important predictors of returns than our measure of (dis)similarity, or might drive out the forecasting power of our measure. We also construct a measure entitled *Sentiment of Change is Positive* to focus on the sentiment of the changes we document, and to separate out the

[21] The results for the other three measures of similarity yield the same conclusions.

[22] Note that in Appendix Table A-12 we show that this sentiment correlation is even stronger if we only include negative words in the construction of this variable. This is consistent with Tetlock (2007) and Tetlock, Saar-Tsechansky, and Macskassy (2008), who show that investors pay special attention to negative words used in media reports.

positive and negative components. As Table VI shows, however, that even after controlling for these document-level characteristics (in a Fama-MacBeth monthly return predictability set-up as in Table V) that similarity remains a large and significant predictor of future returns (t=3.82). Decreases in sentiment alone do predict negative returns, as do increases in the length of the filings, but neither of these measures drives out the predictability of year-over-year changes in document similarity.[23]

B. Isolating Key Sections of Reports

Next, we try to isolate the particular sections of the quarterly and annual reports that are associated with the largest declines in similarity across years for a given firm.

Figure 8 lists the standard sections that are present in firms' annual (10-K) and quarterly (10-Q) reports, respectively. Figure 9-A then plots the average similarity score for different items in firms' 10-Ks and shows that Item 7 (Management's Discussion and Analysis of Financial Condition and Results of Operations—commonly known as the MD&A section) displays a significantly lower average similarity across years than the other categories. Notably, this is the section of the 10-K where management presumably has the most discretion over the content. Similarly, Figure 9-B reports the average similarity score for different items of firms' 10-Qs, and again shows that the MD&A section (here Item 2) displays the lowest average similarity to the other items in the report (although a number

[23] Note that in Appendix Table A-2 we also explore interactions of similarity with document-level characteristics such as Sentiment, and find even stronger return predictability results. See also Appendix Table A-9, which shows that tone/sentiment changes measured across the entire 10-K predict stock returns: negative tone changes predicts negative returns, and positive tone changes predicts positive returns; but that the return predictability of our document similarity measure is unaffected by the inclusion of these variables.

of 10-Q Items are closer in changes year-to-year).

C. Return Predictability of Key Sections of Reports

We then take the item/section categories listed in Figure 8 and examine the return predictability associated with changes to each section. To do so we construct similarity measures for each item of the 10-K using only the textual portion contained within that specific item. As before, for each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores across all stocks. We report the key sections where the return predictability is most pronounced, and report these calendar-time portfolio returns in Table VII. Table VII indicates that changes in the MD&A section are consistently associated with significant future return predictability, although interestingly the magnitude of this effect (ranging between 11-22 basis per month) is often smaller than the effects associated with the "Legal Proceedings" category (Item 3 in the 10-K), the "Quantitative and Qualitative Disclosures About Market Risk" category (Item 7a), and particularly the "Risk Factors" section (Item 1A). Changes concentrated in the Risk Factors section, for example, yield L/S portfolio return alphas (Non-Changers minus Changers) of up to 188 basis points per month (t=2.76) in Panel A of Table VII, or over 22% in risk-adjusted abnormal returns per year. These results suggest that changes to some sections may be quite subtle, and difficult for the market to detect, even though they may have large implications for future returns.

Given the potential structural break in reporting about risk-related items in the wake of Sarbanes-Oxley (see Li, 2010b), we also re-run our analysis for the Risk Factors section in the post-Sarbanes-Oxley period (2003-2014). Appendix Table A-1 shows that

we continue to find large and significant return predictability associated with changes in the Risk Factors section in this most recent time period. Finally, in Figure 10 we plot the value-weighted portfolio alphas by document section in a bar chart, which again highlights the large predictability of the risk factor section.

D. Interacting with Investor Attention

Next, we explore our mechanism in even greater depth by trying to isolate cases where we believe investors *are* paying more attention to these filings, meaning that our return effects should be muted in such instances if we believe our return predictability results are primarily a result of investor inattention. To identify variation in investor attention, we exploit a new database that captures investor behavior at a very granular level: the SEC Edgar traffic log download file. This database contains records of all downloads of corporate filings, matched to the IP addresses of the downloading agent/entity (see Loughran and McDonald, 2017 and Chen, Cohen, Gurun, Lou, and Malloy, 2017 for details). As in Loughran and McDonald (2017), we first remove the impact of "robot requests" which consist of mass downloads by large institutional investors (often quantitative investment firms), and try to test the hypothesis that firms with more "attentive" investor bases see a more muted return predictability effect.

To test this hypothesis, we run Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our similarity measures plus interactions of these similarity measures with a measure of investor attention computed from the SEC log file. Specifically, we construct a variable called *IPAccessMultipleYear*, which is a proxy for having an attentive investor base: it is measured as the number of unique IP addresses that

access *both* the current 10-K/10-Q and the previous year's 10-K/10-Q of the same firm (normalized by the total unique IP addresses that access the current 10-K/10-Q). The idea behind this variable is that if many investors are downloading simultaneously both this year's and last year's filings, we conjecture that it is more likely that they would pick up on the document changes driving our return results; as a result, we would expect them to impound this information into prices more quickly upon the release of the current year's filing, resulting in lower future return predictability.

Table VIII shows that this pattern exists in the data. The interaction term on *IPAccessMultipleYear* x *Similarity* is consistently negative, and significant for 3 of the 4 similarity measures. For example, in Column 8 of Table VIII, the coefficient on this interaction term is negative and significant (=-0.0953, t=2.05), implying a reduction of -0.0136 (or 22% less) in the predictability of *Similarity* for a 1 standard deviation increase in the number of unique IP addresses that check the changes in 10-Ks/10-Qs. These results confirm the idea that when investor attention to year-over-year corporate filings is higher, the return predictability results we document in this paper are weaker.[24]

Next we dig even deeper into the nature of the investor inattention by trying to pinpoint the precise manner in which markets fail to incorporate changes in "qualitative" information as opposed to quantitative information. To do so, we attempt to isolate those firms who make explicitly comparative statements in the text of their annual and quarterly

[24] We also look at filing dates when investors are potentially distracted (as in Hirshleifer, Lim, and Teoh (2006)), by examining filing dates with over 100 earnings announcements on that same date (which we view as high distraction, hence low attention days), relative to filing dates with fewer than 100 earnings announcements. We show in Appendix Table A-10 that the return predictability associated with these high-distraction filing dates (the L/S portfolio spread is equal to 47 basis points per month, t=2.75); and is indeed higher than the return predictability associated with low distraction filing dates (where the L/S portfolio spread average 30 basis points, t=1.43); again these results are consistent with investor inattention serving as an important driver of our findings.

filings, and compare them to firms who do not. For instance, we isolate all the cases where firms include text and phrases like "compared to last year (quarter)" or "relative to last year (quarter)" as well as references to the prior year (e.g., for a 2017 annual report, we isolate mentions such as "compared to 2016" or "relative to prior year") This procedure indicates that roughly 1/3 of the sample contains reports that make explicit comparative textual statements in their filings, while 2/3 of firm-filings do not. We then further divide this comparative sample into the firms that make explicit textual comparisons to specific accounting variables (e.g., "relative to prior year EBITDA", etc.), with those who do not.

We find that our primary return predictability results are driven by the firms who do *not* make explicit textual comparisons in their filings to prior time periods. Our results are consistent with the behavioral interpretation that firms that explicitly draw attention to prior years in their text and actively facilitate superior information processing on the part of investors are less likely to have changes in their reports go un-noticed by the markets; indeed, in Appendix Table A-7 we show that our basic return predictability portfolio result from Table II is concentrated among the firms who do not make these kinds of explicit comparisons.

In addition, we also find that the short-run announcement effect of document changes is significantly more pronounced for those firms that have investors who *do* make the multi-year downloads on the SEC server (see Appendix Table A-8). Recall that we previously showed that the longer-term predictability results were weaker for these firms, but a natural implication of these findings is that the short-run announcement effects should plausibly then be stronger. More precisely, while we find that there is no announcement effect overall in our sample associated with document changes, and no announcement effect for the firms where investors are not downloading multi-year filings;

we find that there *is* a significant short-run announcement effect associated with document changes for the firms where investors are executing multi-year downloads. Since these firms plausibly have an investor base that is more attentive to the year-on-year document changes (as proxied for by our multi-year SEC download measure), it is sensible that the immediate announcement effects associated with document changes would be more pronounced for these firms, as investors (and prices) quickly respond to these changes.

E. Real Effects

To explore the drivers of the return predictability results at the heart of our paper, we also examine the extent to which changes in document similarity predict declines in future operating performance at the firms in question. In Table IX, we explore the predictability of a firm's similarity score for future operating income, net income, and sales. All the future accounting variables are measured 2-quarters ahead. Specifically, we define the following real measures of performance: (Oibdpq/L1atq) as operating income before depreciation (Oidbpq) divided by lagged total assets (L1atq); (Niq/L1atq) as net income (Niq) divided by lagged total assets (L1atq); and (Saleq/L1atq) as sales (Saleq) divided by lagged total assets (L1atq). All of these accounting variables are winsorized at the 1% level throughout the table, and these regressions include month, industry, and firm fixed effects. We also adjust the standard errors for clustering at the monthly level.

Consistent with the idea that the return effects we document in this paper are driven by *future* real declines in operating performance at the firms in question, Table IX shows that all four similarity measures significantly predict these three measures of operating performance (profitability, operating profitability, and sales). For example, focusing on the

Sim_Jaccard measure in the first row of Table IX, we see that decreased similarity (i.e., more changes in the filings) is a significant predictor of lower future operating income, lower net income, and lower sales. These findings highlight the fact that the subtle changes to the filings that we identify in this paper are associated with fundamental changes at these firms.

F. Future News and Events: 8-Ks, Short Interest, Earnings Surprises, and Bankruptcy Events

In this section we examine if document changes predict a wide variety of other types of changes for the firms in question, ranging from future news releases, changes in investor behavior, to other notable events at these companies. In particular, Table X reports the predictability of a firm's similarity score on a firm's future 8-K releases, short interest, earnings surprises (SUEs), and future bankruptcy events. The dependent variables are defined more precisely as follows: in Panel A, the number of 8-Ks ("future releases") that a company files with the SEC to announce major events that shareholders should know about in the next 6 months; in Panel B, the change in short interest in the next month following the document release; in Panel C, the standardized unexpected earnings (SUE) of the subsequent earnings announcement; and in Panel D, a dummy variable equal to one if there is a bankruptcy event in the next year. Bankruptcy events are defined and taken from Capital IQ. All of these regressions include month and firm fixed effects, and again the variables are winsorized at 1% and standard errors are adjusted for clustering at the monthly level.

Panel A of Table X shows that decreases in year-over-year similarity predict a

significant increase in the incidence of 8-K filings by firms over the following 6 months after the 10-K/Q release in question. This is consistent with the idea that changes in document similarity precede important public disclosures by firms. Then in Panel B we report suggestive evidence that document similarity is negatively related to future short interest (in the month following the document release), consistent with the notion that the negative news that gets impounded in prices following a document change is reflected in the shorting market. Next in Panel C we explore future earnings surprises, and the ability of our document similarity measure to forecast future earnings news (measured by SUEs). Panel C indicates that document changes do indeed forecast future negative earnings surprises (at least for 3 of the 4 similarity measures), even controlling for firm and time fixed effects. Finally, in Panel D of Table X we again find suggestive evidence that similarity is negatively related to future bankruptcy events, meaning that document changes are positive predictors of future bankruptcy. Collectively, the results in Table X all point to the idea that document changes have some ability to forecast future (bad) news at the firms in question.

Moreover, in Appendix Table A-4 we drop all instances of special events from the data (e.g., years of M&A, joint ventures, divestitures, or strategic alliances). These are cases in which mechanically there would be changes in 10-Ks and 10-Qs, and we do not want to be capturing a continuation of any return impacts of these special events. From Appendix Table A-4, our results remain strong, robust, and statistically significant upon

dropping these special events from the sample.

G. Other Sorts and Tests of the Mechanism

We also run a slew of additional tests that we now tabulate in the Appendix of this paper. For example, we run additional double-sorts of our portfolio tests, such as for samples of high and low levels of Sentiment, Uncertainty, and Litigiousness, where "low" and "high" are defined as less than the median and higher than median, respectively. For each pair of Low and High samples, we compute quintile portfolios similar to Table II. Appendix Table A-2 shows that the return results documented earlier are concentrated in the Low Sentiment, High Uncertainty, and High Litigiousness subsamples.[25] For instance, the L/S 5-factor alpha for the Jaccard similarity measure is 71 basis points per month (t=3.29) in the High Litigiousness subsample, and 72 basis points per month (t=3.51) in the High Uncertainty subsample, or over 8% in abnormal returns per year.

We also examine the idea that textual similarity may be related to the life cycle of the firm. To measure the life cycle of a firm, we follow Spence (1979), Kotler (1980), and Anthony and Ramesh (1992) and use these four variables as proxies for a firm life cycle stage: (1) annual dividend as a percentage of income, (2) percent sales growth, (3) capital expenditure normalize by total asset, and (4) age of the firm. We then run a regression of our Jaccard similarity on the lagged five-year average of depreciation rate, sales growth, capital expenditure, and age. The regression is run over the entire sample from 1994 to 2014, and the results are shown in Appendix Table A-11. We find that depreciation rate

[25] Appendix Table A-3 also examines the impact of specific law firms that corporations employ to file their 10-Ks, and provides suggestive evidence that in-house lawyers are associated with more year-on-year changes in filings.

and age of a firm are negatively related to Jaccard similarity and sales growth and capital expenditure are positively related to Jaccard similarity. This suggests that firms increasingly modify their financial disclosures as they mature.[26]

H. Robustness Checks

Lastly, we perform a series of robustness checks to ensure that our key findings are not simply repackaging a set of previously known return predictors. To do so, we re-run the Fama-MacBeth regressions from Table IV, but include a series of additional firm-level characteristics, such as accruals (to ensure that the accruals anomaly (see Sloan (1996)) is not driving our findings), investment, gross profitability, and free cash flow. Table XI indicates that none of these variables drive out the return predictability associated with changes to a firm's reporting practices (as captured by our similarity scores).[27]

We then also directly examine the impact of industry concentration in the results we document. In particular, we test whether the results we document are concentrated in any specific industry (or industries) which are driving the results for the whole sample. For instance, if all "changers" were coming from a certain industry, and "non-changers" from another, we'd simply be longing and shorting different industries. To control for this, we run an industry-adjusted version of the calendar-time portfolios of Table II. Namely, within each industry, we sort *that industry* into Q1-Q5 based on changes in documents. We then

[26] We also decompose the Jaccard similarity measure into the expected and unexpected components based on the above predictors for a firm's life cycle. We find that the unexpected component of Jaccard similarity is slightly stronger, both in terms of magnitude and statistical significance, in predicting future stock returns.
[27] In addition, to examine omitted variable biases – and their potential impacts on our estimation – in more depth we follow Oster (2016) and Altonji, Elder, and Taber (2005) to evaluate the robustness to omitted variable bias by observing coefficient and R-squared movements after inclusion of controls. We show this in Appendix Table A-6, which shows that the predictability between changes in documents and future returns (*Similarity*) are unlikely to be significantly impacted by omitted variable concerns.

aggregate each industry's Q1 –Q5 portfolios together into market wide Q1-Q5, now equivalently representing each industry by construction. Appendix Table A-5 shows that the results are strong and significant after making the industry adjustment, suggesting the changes in documents results we find are not linked to specific industries.

Lastly, we also check that our results are not affected by including so-called "stop words" (see Loughran and McDonald (2011)), or by our particular filtering of the SEC filings; for example, when we remove stop words and use the cleaned 10-K/10-Q publicly available database provided by Loughran and McDonald (2011),[28] we show in Appendix Table A-14 that our main portfolio results are even larger and more significant than those reported in Table II.

Collectively our findings indicate that these subtle changes in firms' reporting behavior have substantial predictability for future returns in a manner that has not previously been documented in the literature.

V. Conclusion

In this paper we show that the most comprehensive annual information windows that firms provide to the markets – in the form of their mandated, annual reports – have changed dramatically over time: these reports have become significantly longer and more complex. Moreover, while past literature has found a diminishing announcement effect to these statements, concluding that they have become less informative over time, our evidence points to a different conclusion. Namely, we find that observing simple changes in reports yields a powerful, and robust indicator of future firm performance, from future

[28] http://sraf.nd.edu/textual-analysis/resources/

short sales, to profitability, to probability of bankruptcy. When firms break from their routine phrasing and content - breaking from former language, sections, etc. in their annual and quarterly reports – this action contains rich, important information for future firm outcomes.

However, investors are inattentive to the valuable information in these simple changes. A portfolio that shorts "changers" and buys "non-changers" in annual and quarterly financial reports earns 30-50 basis points per month over the following year. The returns continue to accrue out to 18 months, and do not reverse; this finding suggests that these return movements are not overreactions, but instead reflect true, fundamental changes to firms that only get gradually incorporated into asset prices over the 12-18 months after the reporting change. Importantly, these return patterns are found across the entire universe of publicly traded firms (since public companies are mandated to file annual reports), exist in large firms, inexpensive to short firms, and take place over months, and so are unlikely to be driven by a limit to arbitrage. Moreover, unlike other traditional drift regularities (e.g., return momentum, industry momentum, PEAD), these document changes are not accompanied by any significant announcement returns, and so are inconsistent with a standard underreaction story (as there is no initial reaction). Instead, they are more consistent with a setting where investors are inattentive to this rich information, which is then only impounded into prices with a significant delay. Indeed, when we measure investors' propensity to "compare" this year's filings to prior years—and hence explicitly overcome the laziness/inattention mechanism that we propose in this paper—we find that the returns are significantly attenuated.

Technological advancements reducing the cost of information production and dissemination have made the job of a Grossman Stiglitz investor more complex. And while

technology could also aid in the collection and processing of this same information, we show that far from needing complicated state-of-the-art solutions, simple changes in documents from year-to-year contain powerful information that is seemingly being ignored by the capital markets. This insight likely applies more broadly to other forms of transmitted firm information. Documents such as bond covenants, lease arrangements, securities offering documents, M&A prospectuses – i.e., documents for which there is a regular cadence and repeated use – may be rich places for researchers to explore further. More broadly, the implications of breaks from repeated behaviors in the corporate setting provide a critical, yet understudied area, in both corporate finance and asset pricing.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, 2005, Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools, *Journal of Political Economy* 113, 151–184.

Barron, Orie E., Charles O. Kile, and Terrence B. O'Keefe, 1999, MD&A Quality as Measured by the SEC and Analysts' Earnings Forecasts, *Contemporary Accounting Research* 16, 75–109.

Ben-Raphael, Azi, Zhi Da, and Ryan Israelson, 2017, It depends on where you search: Institutional investor attention and underreaction to news, *Review of Financial Studies* 30, 3009-3047.

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

Bryan, Stephen H., 1997, Incremental information content of required disclosures contained in management discussion and analysis, *Accounting Review* 72, 285–301.

Chen, Huaizhi, Lauren Cohen, Dong Lou, and Christopher J Malloy, 2017, IQ from IP: Simplifying Search in Portfolio Choice, Working Paper, Harvard Business School.

Clarkson, Peter M, Jennifer L. Kao, and Gordon D. Richardson, 1999, Evidence That Management Discussion and Analysis (MD&A) is a Part of a Firm's Overall Disclosure Package, *Contemporary Accounting Research* 16, 111–134.

Cole, C. J., and C. L. Jones, 2005, Management Discussion and Analysis: A review and Implications for Future Research, *Journal of Accounting Literature* 24, 135–174.

Da, Zhi, Joey Engelberg, and Pengjie Gao, 2011, In search of attention, *Journal of Finance* 66, 1461-1499.

Das, S. R. (2014). Text and Context: Language Analytics in Finance. Foundations and Trends in Finance, 8(3), 145-261.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence, 2017, The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation, *Journal of Accounting and Economics* 64, 221-245.

Engelberg, Joey, 2008, Costly information processing: Evidence from earnings announcements, Working Paper, University of California at San Diego.

Engelberg, Joey, Caroline Sasseville, and Jared Williams, 2012, Market madness? The case of *Mad Money*, *Management* Science 58, 351-364.

Feldman, Ronen, Suresh Govindaraj, Joshua Livnat, and Benjamin Segal, 2010, Management's Tone Change, Post Earnings Announcement Drift and Accruals, *Review of Accounting Studies* 15, 915–953.

Grossman, Sanford J., and Joseph E. Stiglitz, 1976, Information and Competitive Price Systems, *American Economic Review* 66, 246–253.

Hanley, Kathleen W., and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review of Financial Studies* 23, 2821-2864.

Hirshleifer, David, Sonya Seongyeon Lim, and Siew Hong Teoh, 2009, Driven to distraction: Extraneous events and underreaction to earnings news, *Journal of Finance* 64, 2289-2325.

Lee, Yen-Jung, 2012, The effect of quarterly report readability on information efficiency of stock prices, *Contemporary Accounting Research* 29, 1137-1170.

Li, Feng, 2008, Annual Report Readability, Current Earnings, and Earnings Persistence, *Journal of Accounting and Economics* 45, 221–247.

Li, Feng, 2010a, The Information Content of Forward-Looking Statements in Corporate Filings - A Naïve Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

Li, Feng, 2010b, Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports? Working Paper, Shanghai Jiaotong University.

Li, Feng, 2011, Textual Analysis of Corporate Disclosures: A Survey of the Literature, *Journal of Accounting Literature* 29, 143-165.

Livnat, Joshua and Richard Mendenhall, 2006, Comparing the post-earnings announcement drift for surprises calculated from analyst and time series forecasts, *Journal of Accounting Research* 44, 177-205.

Loughran, Tim, and Bill McDonald, 2011, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, *Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill McDonald, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187-1230.

Loughran, Tim, and Bill McDonald, 2017, The Use of EDGAR Filings by Investors, *Journal of Behavioral Finance* 18, 231–248.

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

Nelson, Karen K., and Adam C. Pritchard, 2007, Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language, Working Paper, Texas Christian University.

Oster, Emily, 2017, Unobservable Selection and Coefficient Stability: Theory and Evidence, *Journal of Business and Economic Statistics 0,* 1-18.

Rogers, Rodney K., and Julia Grant, 1997, Content Analysis of Information Cited in Reports of Sell-Side Financial Analysts, *Journal of Financial Statement Analysis* 3, 17-30.

Sloan, Richard G, 1996, Do Stock Prices Fully Reflect Information in Accruals and Cash Flows About Future Earnings? *The Accounting Review* 71, 289–315.

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

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *Journal of Finance* 63, 1437-1467.

Tetlock, Paul C., 2011, All the news that's fit to reprint: Do investors react to stale information?, *Review of Financial Studies* 24, 1481-1512.

Tetlock, Paul C., 2014, Information transmission in finance, *Annual Review of Financial Economics* 6, 365-384.

Table I: Summary Statistics on Firms 10-Ks and 10-Qs

Panel A reports the summary statistics of 10-Ks and 10-Qs from 1995 to 2014. *Document Size* is the number of characters (not words) in each document. *Sentiment of Change* is the number of positive words minus the number of negative words normalized by the size of the *Change*. *Uncertainty of Change* and *Litigiousness of Change* are the number of words categorized as uncertainty and litigiousness, respectively, normalized by the size of the *Change*. *Change CEO* and *Change CFO* are indicator variables that equal to one if the 10-K or 10-Q mentions a change in CEO or CFO, respectively. Sentiment category identifiers (e.g., negative, positive, uncertainty, litigious) are taken from Loughran and McDonald (2011)'s Master Dictionary. *Panel B* reports the summary statistics of four different measures of document similarity. *Panel C* reports the correlation between the four similarity measures used in this paper. *Sim_Cosine* is the cosine similarity measure, *Sim_Jaccard* is the Jaccard similarity measure, *Sim_MinEdit* is the minimum edit distance similarity measure, and *Sim_Simple* is the simple side-by-side comparison. Details on how we compute the four similarity measures can be found in the Data section.

Panel A: Summary Statistics on Firms 10-Ks and 10-Qs

	Count	Mean	SD	Min	Max
Document Size - 10K	90198	308633	282473	34660	5.24e+07
Document Size - 10Q	263537	114848.4	286663.9	18824	3.14e+07
Sentiment of Change	353735	-0.0003371	.0011069	-0.00409	.0048492
Uncertainty of Change	353735	.0007317	.0009165	0	.004885
Litigiousness of Change	353735	.0003252	.0009358	0	.0037628
Change CEO	353735	.0539817	.2259819	0	1
Change CFO	353735	.0238223	.1524956	0	1

Panel B: Summary Statistics on Similarity Measures

	Count	Mean	SD	Min	Max
Sim_Cosine	349513	0.8582	0.2118	0.0004	.9999
Sim_Jaccard	349513	0.4234	0.1957	0.0001	.9950
Sim_MinEdit	349513	0.3846	0.1881	0.0000	.9993
Sim_Simple	332821	0.1247	0.1157	0.0000	.9966

Panel C: Correlation

	Sim_Cosine	*Sim_Jaccard*	*Sim_MinEdit*	*Sim_Simple*
Sim_Cosine	1.0000			
Sim_Jaccard	0.6485	1.0000		
Sim_MinEdit	0.5494	0.8159	1.0000	
Sim_Simple	0.2473	0.5811	0.6317	1.0000

Table II: Main Results – Calendar Time Portfolio Returns

This Table reports the calendar-time portfolio returns. *Sim_Cosine* is the cosine similarity measure, *Sim_Jaccard* is the Jaccard similarity measure, *Sim_MinEdit* is the minimum edit distance similarity measure, and *Sim_Simple* is the simple side-by-side comparison. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity measures across all stocks. Stocks then enter the quintile portfolios in the month after the public release of one of their 10-K or 10-Q reports. Stocks are held in the portfolio for 3 months. We report Excess Returns (return minus risk free rate), Fama-French 3-factor Alphas (market, size, and value), and 5-factor Alphas (market, size, value, momentum, and liquidity). *Panel A* reports equal-weight portfolio returns, and *Panel B* reports value-weight portfolio returns. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Panel A: Equally Weighted

Sim_Cosine

	Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0063*	0.0072*	0.0072**	0.0085**	0.0092***	0.0031***
Return	(1.6844)	(1.9562)	(2.1098)	(2.5915)	(2.7958)	(3.1295)
3-Factor	-0.0015**	-0.0008	-0.0005	0.0009	0.0018***	0.0034***
Alpha	(-2.1917)	(-1.0984)	(-0.7182)	(1.2102)	(2.6597)	(4.4527)
5-Factor	-0.0012*	-0.0005	-0.0004	0.0010	0.0021***	0.0032***
Alpha	(-1.7529)	(-0.7409)	(-0.5334)	(1.2891)	(3.2830)	(4.2050)

Sim_Jaccard

	Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0059	0.0067*	0.0069*	0.0082**	0.0098***	0.0038***
Return	(1.4795)	(1.7358)	(1.8874)	(2.3469)	(3.0051)	(2.6525)
3-Factor	-0.0016**	-0.0010	-0.0006	0.0008	0.0028***	0.0044***
Alpha	(-1.9903)	(-1.2193)	(-0.8056)	(1.0453)	(3.4748)	(4.5582)
5-Factor	-0.0014*	-0.0007	-0.0006	0.0009	0.0028***	0.0042***
Alpha	(-1.8428)	(-0.9348)	(-0.8598)	(1.1935)	(3.5744)	(4.3051)

Sim_MinEdit

	Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0061	0.0066*	0.0070*	0.0086**	0.0099***	0.0036***
Return	(1.5972)	(1.7821)	(1.9375)	(2.5803)	(3.3628)	(2.6851)
3-Factor	-0.0019**	-0.0014*	-0.0010	0.0010	0.0030***	0.0048***
Alpha	(-2.5589)	(-1.9084)	(-1.5170)	(1.3714)	(3.9961)	(5.9594)
5-Factor	-0.0015**	-0.0011	-0.0008	0.0012*	0.0030***	0.0045***
Alpha	(-2.1401)	(-1.5907)	(-1.3126)	(1.7002)	(4.1087)	(5.4649)

Sim_Simple

	Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0072*	0.0079**	0.0082**	0.0090***	0.0090***	0.0018
Return	(1.8671)	(2.1185)	(2.3413)	(2.7340)	(3.0359)	(1.2038)
3-Factor	-0.0008	-0.0002	0.0003	0.0014**	0.0020**	0.0028***
Alpha	(-1.0934)	(-0.2075)	(0.3834)	(2.0139)	(2.5730)	(3.2194)
5-Factor	-0.0006	0.0003	0.0004	0.0016**	0.0021***	0.0027***
Alpha	(-0.8898)	(0.3700)	(0.6345)	(2.3037)	(2.6774)	(3.0117)

Panel B: Value Weighted

	Sim_Cosine							*Sim_Jaccard*					
	Q1	Q2	Q3	Q4	Q5	Q5 - Q1		Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0043	0.0047	0.0055*	0.0073**	0.0078**	0.0034**	Excess	0.0023	0.0032	0.0048	0.0061*	0.0079**	0.0056***
Return	(1.3175)	(1.4452)	(1.7378)	(2.3510)	(2.4006)	(2.5277)	Return	(0.6428)	(0.8759)	(1.3267)	(1.8440)	(2.4684)	(3.7529)
3-Factor	-0.0015*	-0.0015*	-0.0004	0.0010	0.0020*	0.0035***	3-Factor	-0.0032***	-0.0021	-0.0009	0.0007	0.0023**	0.0054***
Alpha	(-1.8420)	(-1.7869)	(-0.4884)	(1.1653)	(1.9677)	(2.6264)	Alpha	(-2.9705)	(-1.2966)	(-0.7260)	(0.6030)	(2.0125)	(4.0784)
5-Factor	-0.0012	-0.0019**	-0.0006	0.0012	0.0023**	0.0034**	5-Factor	-0.0023**	-0.0017	-0.0007	0.0013	0.0023**	0.0046***
Alpha	(-1.3826)	(-2.1346)	(-0.6463)	(1.3562)	(2.2318)	(2.5306)	Alpha	(-2.1957)	(-1.0379)	(-0.5917)	(1.1810)	(2.1099)	(3.4439)
	Sim_MinEdit							*Sim_Simple*					
	Q1	Q2	Q3	Q4	Q5	Q5 - Q1		Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0042	0.0045	0.0062*	0.0076**	0.0083***	0.0039**	Excess	0.0024	0.0061*	0.0077**	0.0078**	0.0074**	0.0050***
Return	(1.2513)	(1.3755)	(1.8790)	(2.4179)	(2.9217)	(2.3077)	Return	(0.6879)	(1.8821)	(2.4476)	(2.5284)	(2.4775)	(2.6924)
3-Factor	-0.0018**	-0.0016*	-0.0001	0.0017*	0.0028**	0.0046***	3-Factor	-0.0039***	0.0002	0.0018*	0.0019*	0.0019	0.0058***
Alpha	(-2.2916)	(-1.9110)	(-0.1420)	(1.7441)	(2.4895)	(3.0576)	Alpha	(-3.8893)	(0.1802)	(1.8704)	(1.8797)	(1.4452)	(3.5865)
5-Factor	-0.0017**	-0.0014*	0.0000	0.0017*	0.0021*	0.0037**	5-Factor	-0.0036***	0.0005	0.0018*	0.0018*	0.0015	0.0051***
Alpha	(-2.0184)	(-1.6724)	(0.0397)	(1.7814)	(1.8437)	(2.4488)	Alpha	(-3.4960)	(0.6607)	(1.7835)	(1.7139)	(1.1461)	(3.1419)

Table III: Characteristics of Quintile Portfolios

This table reports *Size,* the log of market value of equity, *Monthly turnover*, *Shorting fees*, and *Sentiment,* which is is the sentiment of changes, of the five quintile portfolios.

	Q1	Q2	Q3	Q4	Q5
Size	3507587	3219430	2829955	2504717	2464603
Monthly turnover	0.0663	0.0850	0.0804	0.0867	0.0706
Shorting fees (bps)	71.69582	80.63605	92.05002	87.06895	73.54532
Sentiment	-0.0033	-0.0011	-0.0007	-0.0005	-0.0004

Table IV: Main Results – Fama MacBeth Regressions

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our four similarity measures and a host of known return predictors. *Sim_Cosine* is the cosine similarity measure, *Sim_Jaccard* is the Jaccard similarity measure, *Sim_MinEdit* is the minimum edit distance similarity measure, and Sim_Simple is the simple side-by-side comparison. *Size* is log of market value of equity, *log(BM)* is log book value of equity over market value of equity, *Ret(-1,0)* is previous month's return, and *Ret(-12, -1)* is the cumulative return from month -12 to month -1. *SUE* is the standardized unexpected earnings and computed as actual earnings per share minus average analyst forecast earnings per share, divided by the standard deviation of forecasts. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
							Ret					
Sim_Cosine	0.0045***	0.0031**	0.0037**									
	(2.6469)	(2.5103)	(2.1751)									
Sim_Jaccard				0.0082***	0.0066***	0.0059***						
				(3.2607)	(3.8197)	(3.4063)						
Sim_MinEdit							0.0054**	0.0041***	0.0029**			
							(2.5398)	(2.7795)	(1.9970)			
Sim_Simple										0.0404**	0.0302**	0.0292**
										(2.1031)	(2.2484)	(2.1099)
Size		0.0000	0.0000		0.0001	0.0001		0.0001	0.0001		0.0001	0.0000
		(0.1111)	(0.0507)		(0.2496)	(0.1133)		(0.2558)	(0.0980)		(0.2385)	(0.0485)
log(BM)		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*		0.0017*	0.0016*
		(1.8936)	(1.7142)		(1.8797)	(1.7047)		(1.8955)	(1.7163)		(1.8740)	(1.6957)
Ret(-1,0)		-0.0260***	-0.0243***		-0.0263***	-0.0244***		-0.0263***	-0.0244***		-0.0263***	-0.0245***
		(-3.9281)	(-3.6827)		(-3.9704)	(-3.7026)		(-3.9731)	(-3.6930)		(-3.9852)	(-3.7105)
Ret(-12,-1)		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0036		0.0064**	0.0037
		(2.3394)	(1.2457)		(2.3407)	(1.2502)		(2.3357)	(1.2438)		(2.3469)	(1.2934)
SUE			0.0007***			0.0007***			0.0007***			0.0007***
			(6.5591)			(6.5442)			(6.5584)			(6.4993)
Cons	0.0058	0.0058	0.0067	0.0064	0.0046	0.0069	0.0076**	0.0057	0.0084	-0.0238	-0.0176	-0.0142
	(1.4516)	(0.6721)	(0.5684)	(1.6348)	(0.5171)	(0.5814)	(1.9765)	(0.6369)	(0.7057)	(-1.3069)	(-1.0217)	(-0.7060)
R-Squared	0.0006	0.0427	0.0485	0.0017	0.0432	0.0489	0.0017	0.0432	0.0488	0.0019	0.0435	0.0492
N	713451	713451	496084	713451	713451	496084	713451	713451	496084	713680	713680	495931

Table V: Potential Mechanism

This Table explores the potential mechanism behind our results. We regress our similarity measure on a host of characteristics of the document in question. *Sentiment* is the number of positive words in the change minus the number of negative words in the change normalized by the size of the change. *Uncertainty* and are the number of words categorized as uncertainty and litigiousness, respectively, normalized by the size of the Change. *Change CEO* and *Change CFO* are indicator variables that equal to one if the 10-K or 10-Q mentions a change in CEO or CFO, respectively. Sentiment category identifiers (e.g., negative, positive, uncertainty, litigious) are taken from Loughran and McDonald (2011)'s Master Dictionary. All regressions include firm fixed effects and month fixed effects. Standard errors are clustered at the firm level. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)
			Sim_Simple		
Sentiment	3.5229***				
	(48.5432)				
Uncertainty		-3.5698***			
		(-34.1502)			
Litigiousness			-0.1226**		
			(-2.1148)		
Change CEO				-0.0064***	
				(-7.0998)	
Change CFO					-0.0076***
					(-5.7458)
Cons	0.1898***	0.1854***	0.1841***	0.1849***	0.1844***
	(17.9689)	(17.3996)	(17.2517)	(17.3104)	(17.2873)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes
R-Squared	0.0871	0.0679	0.0652	0.065	0.0649
N	338138	338138	338138	338138	338138

Table VI: Fama MacBeth Regressions, Controlling for Sentiment and Document Size

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our Sim_Jaccard similarity measures and a host of known return predictors. S*im_Jaccard* is the Jaccard similarity measure. *Sentiment of Change is Positive* is the number of positive words in the change, normalized by the size of the change. *Size* is log of market value of equity, *log(BM)* is log book value of equity over market value of equity, *Ret(-1,0)* is previous month's return, and *Ret(-12, -1)* is the cumulative return from month -12 to month -1. *Log(Document Size)* is the logarithm of the number of words in a document. *Δ Log(Document Size)* is the quarter-on-quarter change in *Log(Document Size)*. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)
		Ret	
Sim_Jaccard	0.0057***	0.0058***	0.0058***
	(3.4507)	(3.7844)	(3.8217)
Sentiment of Change is Positive	0.0019***	0.0021***	0.0021***
	(3.8515)	(4.2135)	(4.3310)
Log(Document Size)		0.0001	0.0003
		(0.6479)	(1.3957)
Δ Log(Document Size)			-0.0041**
			(-2.3017)
Size	0.0000	0.0000	-0.0000
	(0.1037)	(0.0668)	(-0.0102)
log(BM)	0.0017	0.0016	0.0016
	(1.6360)	(1.5858)	(1.5471)
Ret(-1,0)	-0.0267***	-0.0269***	-0.0269***
	(-4.1535)	(-4.1932)	(-4.1974)
Ret(-12,-1)	0.0074***	0.0074***	0.0074***
	(2.7110)	(2.6950)	(2.6856)
Cons	0.0055	0.0041	0.0025
	(0.5990)	(0.4805)	(0.2961)
R-Squared	0.0437	0.0445	0.0448
N	713451	713451	713451

Table VII: Portfolio Sorts - By Document Section

This Table reports the calendar-time portfolio returns for the common sections of firms' 10-K and 10-Q financial reports: Management's Discussion and Analysis, Legal Proceedings, Quantitative and Qualitative Disclosures About Market Risk, Risk Factors, and Other Information. Similarity measures for each section are computed using only the textual portion in that section. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity measures across all stocks. Stocks then enter the quintile portfolio in the month after the public release of one of their 10-K or 10-Q reports. Firms are held in the portfolio for 3 months. We report Excess Returns (return minus risk free rate), Fama-French 3-factor Alphas (market, size, and value), and 5-factor Alphas (market, size, value, momentum, and liquidity) of the top minus bottom quintile portfolio (Q5 – Q1). *Panel A* reports equal-weight portfolio returns, and Panel B reports value-weight portfolio returns. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Panel A: Equally Weighted

	Sim_Cosine			*Sim_Jaccard*		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0013	0.0011*	0.0012*	0.0021**	0.0022***	0.0020***
	(1.5648)	(1.6579)	(1.6751)	(2.5054)	(3.1451)	(2.8061)
Legal Proceedings	0.0036**	0.0037***	0.0033***	0.0028	0.0030**	0.0025*
	(2.2428)	(3.0939)	(2.6989)	(1.5729)	(2.3602)	(1.9341)
Quant. and Qual. Disclosures About Market Risk	0.0069***	0.0068***	0.0068***	0.0020**	0.0021***	0.0019***
	(2.7465)	(2.6923)	(2.6481)	(2.3738)	(2.9594)	(2.6049)
Risk Factors	0.0114	0.0118	0.0118	0.0143**	0.0144**	0.0188***
	(1.6111)	(1.6308)	(1.6365)	(2.1325)	(2.4497)	(2.7601)
Other Information	0.0020	0.0027	0.0036*	0.0031*	0.0037**	0.0040**
	(1.0839)	(1.4684)	(1.9179)	(1.7849)	(2.1854)	(2.2959)
	Sim_MinEdit			*Sim_Simple*		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0018*	0.0022***	0.0019***	0.0019***	0.0019**	0.0017**
	(1.9519)	(3.1616)	(2.6652)	(2.6673)	(2.5405)	(2.3253)
Legal Proceedings	0.0022	0.0025**	0.0022*	0.0013	0.0016	0.0012
	(1.2706)	(2.3030)	(1.9347)	(0.8157)	(1.4119)	(1.1042)
Quant. and Qual. Disclosures About Market Risk	0.0016	0.0023*	0.0022*	0.0013	0.0011	0.0007
	(1.1822)	(1.7374)	(1.6712)	(0.1581)	(0.1319)	(0.0801)
Risk Factors	0.0102	0.0185***	0.0138**	0.0125*	0.0154**	0.0177**
	(1.1928)	(2.7728)	(2.1663)	(1.9310)	(2.1914)	(2.1156)
Other Information	0.0009	0.0014	0.0016	0.0022	0.0026**	0.0022*
	(0.5773)	(0.9649)	(1.0514)	(1.2731)	(2.3091)	(1.9525)

Panel B: Value Weighted

	Sim_Cosine			*Sim_Jaccard*		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0027*	0.0028*	0.0022	0.0047***	0.0043***	0.0033**
	(1.8009)	(1.8471)	(1.4237)	(2.8834)	(2.6347)	(2.0151)
Legal Proceedings	0.0035*	0.0032	0.0032	0.0018	0.0010	0.0005
	(1.6643)	(1.5347)	(1.4722)	(0.8050)	(0.4609)	(0.2127)
Quant. and Qual. Disclosures About Market Risk	0.0039	0.0044	0.0045	0.0047***	0.0042***	0.0038**
	(1.3980)	(1.5716)	(1.6159)	(2.8918)	(2.6005)	(2.3723)
Risk Factors	0.0144*	0.0150**	0.0156**	0.0118*	0.0165***	0.0156**
	(1.9625)	(2.0069)	(2.0470)	(1.8999)	(2.7450)	(2.5669)
Other Information	0.0073**	0.0075**	0.0080**	0.0054	0.0049	0.0043
	(2.1343)	(2.2083)	(2.3014)	(1.5574)	(1.4249)	(1.2049)
	Sim_MinEdit			*Sim_Simple*		
	Excess Return	3-Factor Alpha	5-Factor Alpha	Excess Return	3-Factor Alpha	5-Factor Alpha
Management's Discussion and Analysis	0.0047***	0.0044***	0.0033*	0.0038**	0.0037**	0.0025
	(2.6718)	(2.6389)	(1.9706)	(2.0562)	(2.1179)	(1.4231)
Legal Proceedings	0.0014	0.0005	0.0007	0.0030	0.0024	0.0027
	(0.6083)	(0.2467)	(0.2985)	(1.2640)	(1.0351)	(1.1573)
Quant. and Qual. Disclosures About Market Risk	0.0000	0.0014	0.0012	0.0013	0.0011	0.0007
	(0.0149)	(0.6396)	(0.6135)	(0.1581)	(0.1319)	(0.0801)
Risk Factors	0.0095	0.0151**	0.0105*	0.0125	0.0133	0.0085
	(1.1777)	(2.2874)	(1.6658)	(1.5388)	(1.6108)	(1.0385)
Other Information	0.0022	0.0011	0.0009	0.0013	0.0002	0.0000
	(0.6272)	(0.3286)	(0.2515)	(0.3783)	(0.0678)	(0.0146)

Table VIII: Interacting with Investor Attention

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our similarity measures and interactions of the similarity measures with *IPAccessMultipleYear*. *IPAccessMultipleYear* is a proxy for the investor bases of firms that do check the changes in 10-Ks/10-Qs and is measured as the number of unique IP addresses that access both the current 10K/10-Q and previous year's 10-K/10K of the same firm normalized by the total unique IP addresses that access the current 10-K/10-Q. We download EDGAR traffic log file from the SEC and remove robot requests as in Loughran and McDonald (2015). *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Return							
	Sim_Cosine		*Sim_Jaccard*		*Sim_MinEdit*		*Sim_Simple*	
Similarity	0.0044**	0.0042**	0.0078***	0.0084***	0.0065***	0.0073***	0.0546**	0.0614**
	(2.5611)	(2.3725)	(2.8998)	(3.0773)	(2.7005)	(2.9432)	(2.1312)	(2.2997)
IPAccessMultipleYear x Similarity		-0.0027		-0.0084**		-0.0079*		-0.0953**
		(-0.6473)		(-2.0764)		(-1.7247)		(-2.0525)
IPAccessMultipleYear		0.0011		0.0015		0.0011		0.0786**
		(0.3070)		(0.8637)		(0.4996)		(2.0532)
Cons	0.0052	0.0054	0.0059	0.0057	0.0065	0.0063	-0.0362	-0.0418*
	(1.1637)	(1.1962)	(1.3588)	(1.3140)	(1.5014)	(1.4360)	(-1.5279)	(-1.6955)
R-Squared	0.0006	0.0014	0.0016	0.0024	0.0017	0.0025	0.0019	0.0027
N	547918	547918	547918	547918	547918	547918	548912	548912

Table IX: Real Effects

This Table reports regressions of operating income, net income, and sales on a firm's lagged similarity measures. *Oibdpq/L1atq* is operating income before depreciation (Oidbpq) divided by lagged total assets (L1atq). *Niq/L1atq* is net income (Niq) divided by lagged total assets (L1atq). *Saleq/L1atq* is sales (Saleq) divided by lagged total assets (L1atq). All variables in the table are winsorized at the 1% level. All regressions include month, industry, and firm fixed effects. Standard errors are adjusted for clustering at the monthly level. *t*-statistics calculated using the robust clustered standard errors are reported in parentheses. Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Oibdpq/L1atq				*Niq/L1atq*				*Saleq/L1atq*			
Sim_Jaccard	0.0068***				0.0089***				0.0133***			
	(10.6827)				(10.4802)				(7.8262)			
Sim_Cosine		0.0050*				0.0048				0.0131*		
		(1.9548)				(1.4433)				(1.9446)		
Sim_MinEdit			0.0065***				0.0075***				0.0200***	
			(12.4811)				(10.8781)				(14.4801)	
Sim_Simple				0.0051***				0.0071***				0.0118***
				(7.7947)				(8.4065)				(6.8514)
Cons	-0.0040***	-0.0135***	-0.0120***	-0.0161***	-0.0442***	-0.0418***	-0.0409***	-0.0423***	0.2149***	0.2126***	0.2151***	0.1944***
	(-3.0493)	(-4.7095)	(-8.5891)	(-6.3316)	(-24.0674)	(-11.1680)	(-23.5648)	(-12.7560)	(51.474)	(27.3337)	(53.7334)	(27.6678)
Month FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-Squared	0.0116	0.0585	0.2858	0.0558	0.0549	0.0581	0.2859	0.0588	0.0563	0.0596	0.287	0.2864
N	284151	284151	284151	325717	295031	295031	295031	338477	295031	295031	295031	338476

Table X: Future 8Ks, Short Interest, SUE, and Bankruptcy Events

This Table reports regressions of firm's future 8-K releases, insider selling activities, short interest, SUE, and future bankruptcy events on a firm's lagged similarity measures. All variables in the table are winsorized at the 1% level throughout the table. The dependent variable in *Panel A* is the number of 8-Ks (current reports) that a company must file with the SEC to announce major events that shareholders should know about in the next 6 months. The dependent variable in *Panel B* is change in short interest in the next month. The dependent variable in *Panel C* is the next standardized unexpected earnings. The dependent variable in *Panel D* is a dummy that equals to one if there is a bankruptcy event in the next year. Bankruptcy events are from CapitalIQ. All regressions include month and firm fixed effects. Standard errors are adjusted for clustering at the monthly level. *t*-statistics calculated using the robust clustered standard errors are reported in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

Panel A

	(1)	(2)	(3)	(4)
	Number of 8Ks in the next 6 months			
Sim_Jaccard	-0.0671**			
	(-2.1913)			
Sim_Cosine		0.0302		
		(0.2940)		
Sim_MinEdit			-0.0634***	
			(-2.5538)	
Sim_Simple				-0.0930***
				(-2.7641)
Cons	0.3111***	0.2224*	0.2910***	0.0468
	-5.0651	-1.6877	-5.0321	-0.3553
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R-Squared	0.0016	0.0015	0.0016	0.001
N	295560	295560	295560	337478

Panel B

	(1)	(2)	(3)	(4)
	Average short interest in the next month			
Sim_Jaccard	-0.2855**			
	(-2.0340)			
Sim_Cosine		-0.0821**		
		(-2.4103)		
Sim_MinEdit			0.0151	
			(0.5475)	
Sim_Simple				-0.0262
				(-0.8282)
Cons	0.5652***	0.3483***	0.2738***	0.1955**
	-3.8009	-5.915	-5.0587	-2.0651
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R-Squared	0.0077	0.0076	0.0077	0.0079
N	158259	158259	158259	185279

Panel C

	(1)	(2)	(3)	(4)
	Next SUE			
Sim_Jaccard	0.6049***			
	(8.0356)			
Sim_Cosine		1.5094***		
		(4.0017)		
Sim_MinEdit			0.1934***	
			(3.3112)	
Sim_Simple				-0.0291
				(-0.4018)
Cons	-0.7914***	-1.7978***	-0.4023***	0.143
	(-6.1113)	(-4.5942)	(-3.4014)	-0.6235
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R-Squared	0.0103	0.0099	0.0097	0.0098
N	180265	180265	180265	208196

Panel D

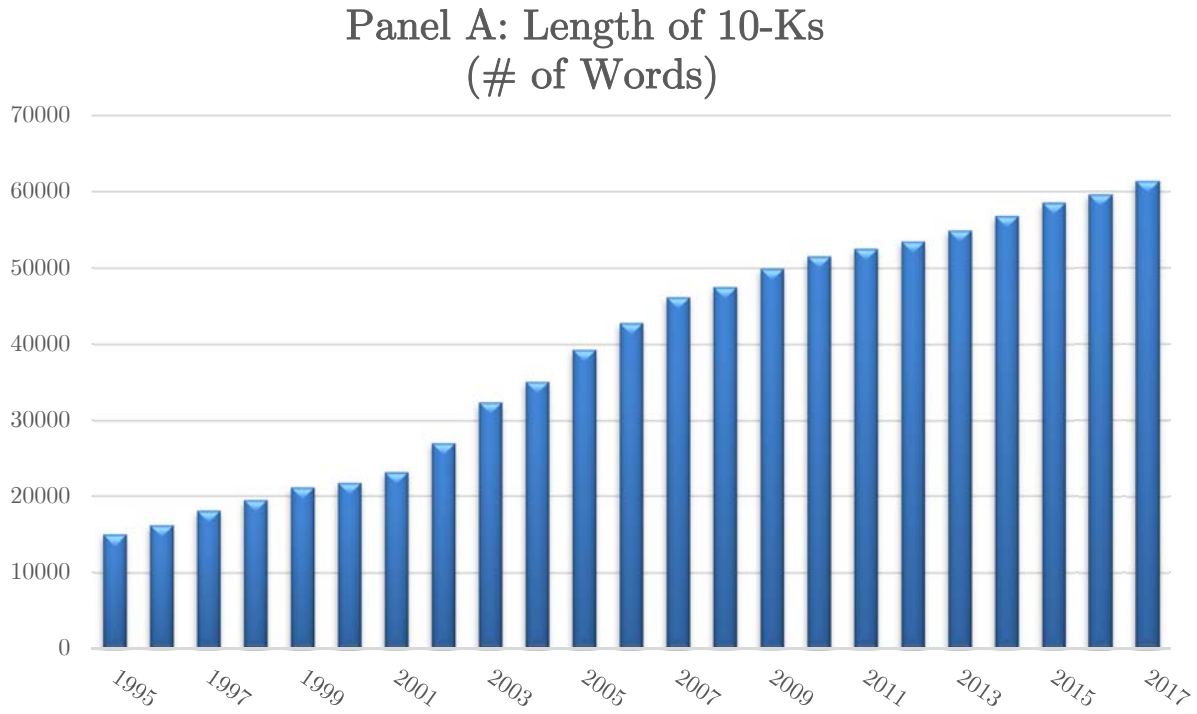
	(1)	(2)	(3)	(4)
	Future bankruptcy events			
Sim_Jaccard	-0.0019**			
	(-1.9993)			
Sim_Cosine		-0.001		
		(-0.2423)		
Sim_MinEdit			-0.0019**	
			(-2.1867)	
Sim_Simple				-0.0008
				(-0.7458)
Cons	0.0133***	0.0127***	0.0128***	0.0100***
	-8.2415	-2.9313	-8.5568	-14.1149
Time FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R-Squared	0.0007	0.0007	0.0008	0.0006
N	296074	296074	296074	338133

Table XI: Robustness – Fama MacBeth with more controls

This Table reports Fama-MacBeth cross-sectional regressions of individual firm-level monthly stock returns on our four similarity measures and a host of known return predictors. *Size* is log of market value of equity, *log(BM)* is log book value of equity over market value of equity, *Ret(-1,0)* is previous month's return. *Ret(-3, -1), Ret(-6, -1), Ret(-9, -1), and Ret(-12, -1)* are the cumulative return from month -3 to month -1, month -6 to month -1, and month -12 to month -1, respectively. *Invest* is capx/ppent. *GrossProfit* is (revt-cogs)/at. *FreeCashFlow* is (ni + dp - wcapch - capx)/at. *Accrual* is (Δact - chech - Δlct + Δdct + Δtxp - dp) scaled by average assets (at/2 + lag(at)/2). SUE is Compustat-based standardized unexpected earnings, and is computed as in Livnat and Mendenhall (2006), where Compustat-based Earnings Surprise is based on the assumption that EPS follows a seasonal random walk, where the best expectation of the EPS in quarter t is the firm's reported EPS in the same quarter of the previous fiscal year. *t*-statistics are shown below the estimates, and statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

	(1)	(2)	(3)	(4)
		Ret		
Sim_Cosine	0.0038***			
	(3.1920)			
Sim_Jaccard		0.0055***		
		(4.1972)		
Sim_MinEdit			0.0035***	
			(3.0559)	
Sim_Simple				0.0318**
				(2.3869)
Ret(-1,0)	-0.0298***	-0.0300***	-0.0301***	-0.0295***
	(-5.7678)	(-5.8082)	(-5.8257)	(-5.5795)
Ret(-3,-1)	0.0000	-0.0001	0.0000	-0.0005
	(-0.0108)	(-0.0130)	(-0.0053)	(-0.1088)
Ret(-6,-1)	(0.0006)	(0.0005)	-0.0005	0.0001
	(0.1739)	(0.1615)	(0.1539)	(0.0310)
Ret(-12,-1)	0.0057**	0.0057**	0.0056**	0.0059**
	(2.4081)	(2.4048)	(2.4005)	-2.4812
Size	0.0000	0.0001	0.0001	-0.0001
	(0.0292)	(0.1438)	(0.1578)	(-0.1942)
log(BM)	0.0012**	0.0013**	0.0013**	0.0012*
	(2.0156)	(2.0586)	(2.0671)	(1.8952)
Invest	-0.0026	-0.0024	-0.0025	-0.0023
	(-0.8126)	(-0.7535)	(-0.7703)	(-0.6920)
GrossProfit	0.0033*	0.0032*	0.0032*	0.003
	(1.8797)	(1.8398)	(1.8247)	(1.6332)
Accrual	-0.0098***	-0.0098***	-0.0098***	-0.0107***
	(-4.1605)	(-4.1842)	(-4.1717)	(-4.6210)
FreeCashFlow	0.0084**	0.0080**	0.0081**	0.0086**
	(2.3188)	(2.2151)	(2.2518)	(2.3296)
SUE	0.0011***	0.0011***	0.0011***	0.0011***
	(5.5288)	(5.5463)	(5.5718)	(4.8752)
Cons	0.0055	0.0053	0.0060	-0.0206
	(0.7191)	(0.6811)	(0.7817)	(-1.2576)
R-Squared	0.0649	0.0651	0.0651	0.0674
N	630081	630081	630081	569180

Figure 1: Length of 10-Ks and Changes to 10-Ks over Time



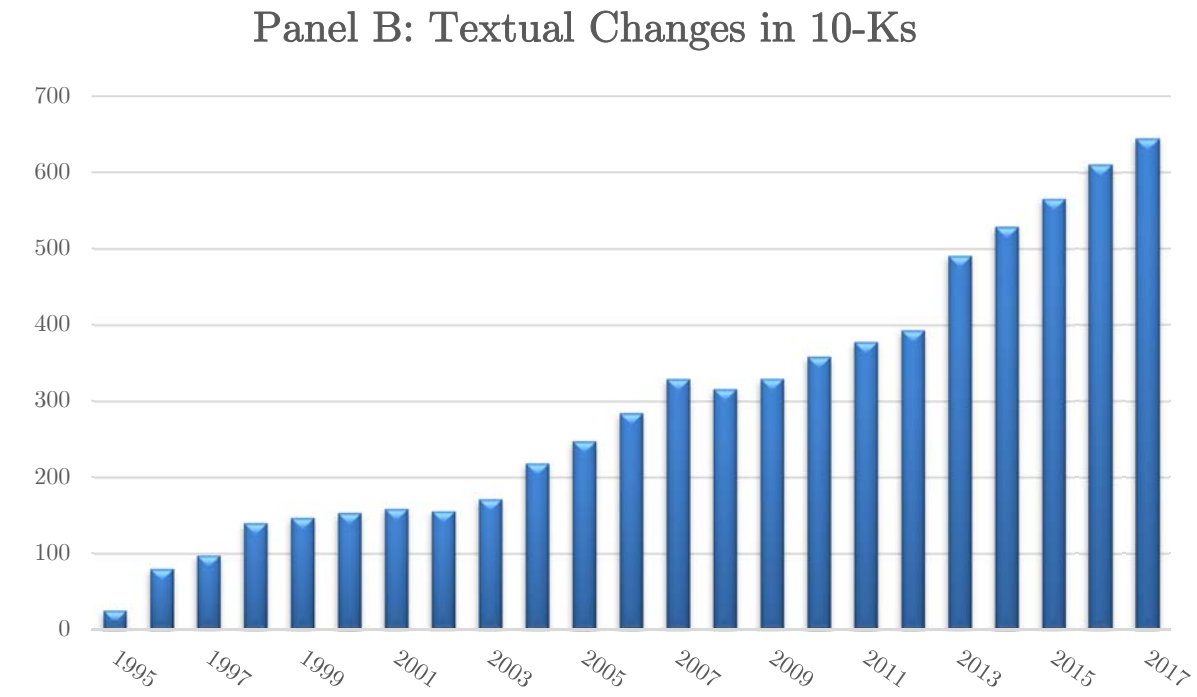


Figure 2: Similarity of Baxter International Inc.

This figure plots the Jaccard Similarity of Baxter International Inc. (NYSE: BAX) 10-K reports from 1997 to 2014, by year of filing/release date.

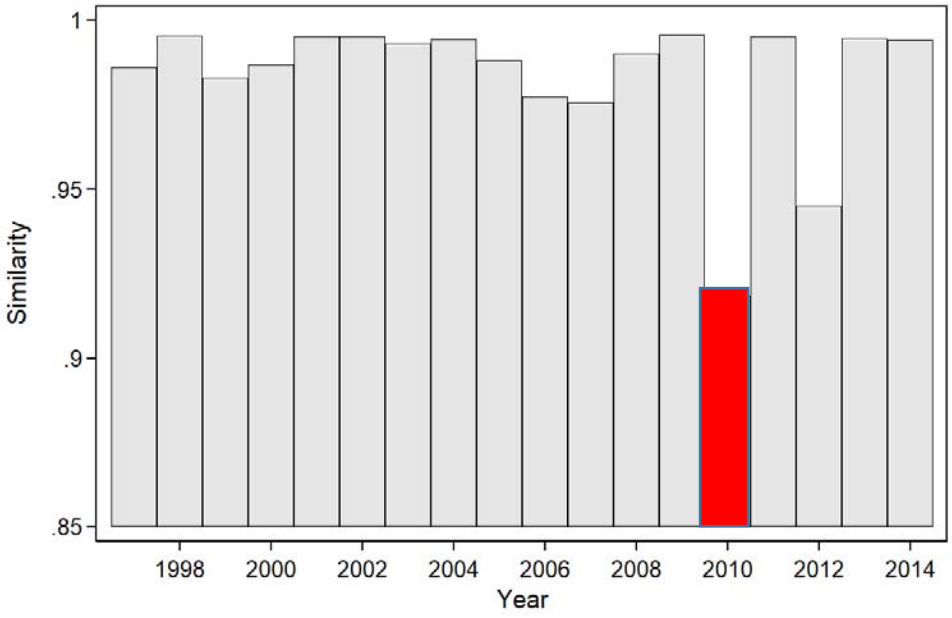
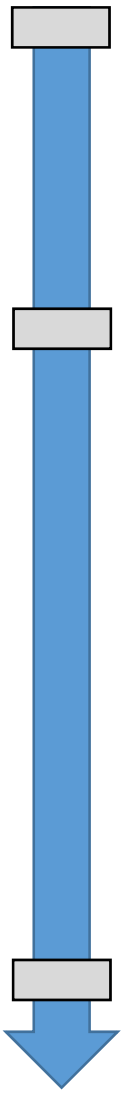


Figure 3: Main events and news articles regarding Baxter's recall of Colleague pumps in 2010



02/23/2010: Baxter filed its 2009 10-K financial report with the SEC

https://www.sec.gov/Archives/edgar/data/10456/000095012310015380/0000950123-10-015380-index.htm

04/23/2010: *The New York Times "F.D.A. Steps Up Oversight of Infusion Pumps"*

http://www.nytimes.com/2010/04/24/business/24pump.html

"Federal regulators say they are moving to tighten their oversight of medical devices, including one of the most ubiquitous and problematic pieces of medical equipment — automated pumps that intravenously deliver drugs, food and other solutions to patients.**"**

"The biggest makers of infusion pumps include Baxter Healthcare of Deerfield, Ill.; Hospira of Lake Forest, Ill.; and CareFusion of San Diego.**"**

"Dr. Shuren said he expected that the new requirements would initially slow down the rate of the agency's approval for new pumps that manufacturers are seeking to market.**"**

05/04/2010: *The New York Times "F.D.A. Deal Leads to Recall of Infusion Pumps"*

http://www.nytimes.com/2010/05/04/business/04baxter.html

"Baxter International is recalling its Colleague infusion pumps from the American market under an agreement with federal regulators that sought to fix problems like battery failures and software errors.**"**

"Baxter expects to record a pretax charge of $400 million to $600 million in the first quarter related to the recall, the company said Monday in a statement. The company isn't otherwise revising its 2010 forecast.**"**

Figure 4: Baxter Stock return

This figure reports the daily returns and the cumulative returns of Baxter International Inc. (NYSE: BAX) in the months folling the release of Baxter's 2009 10-K report.

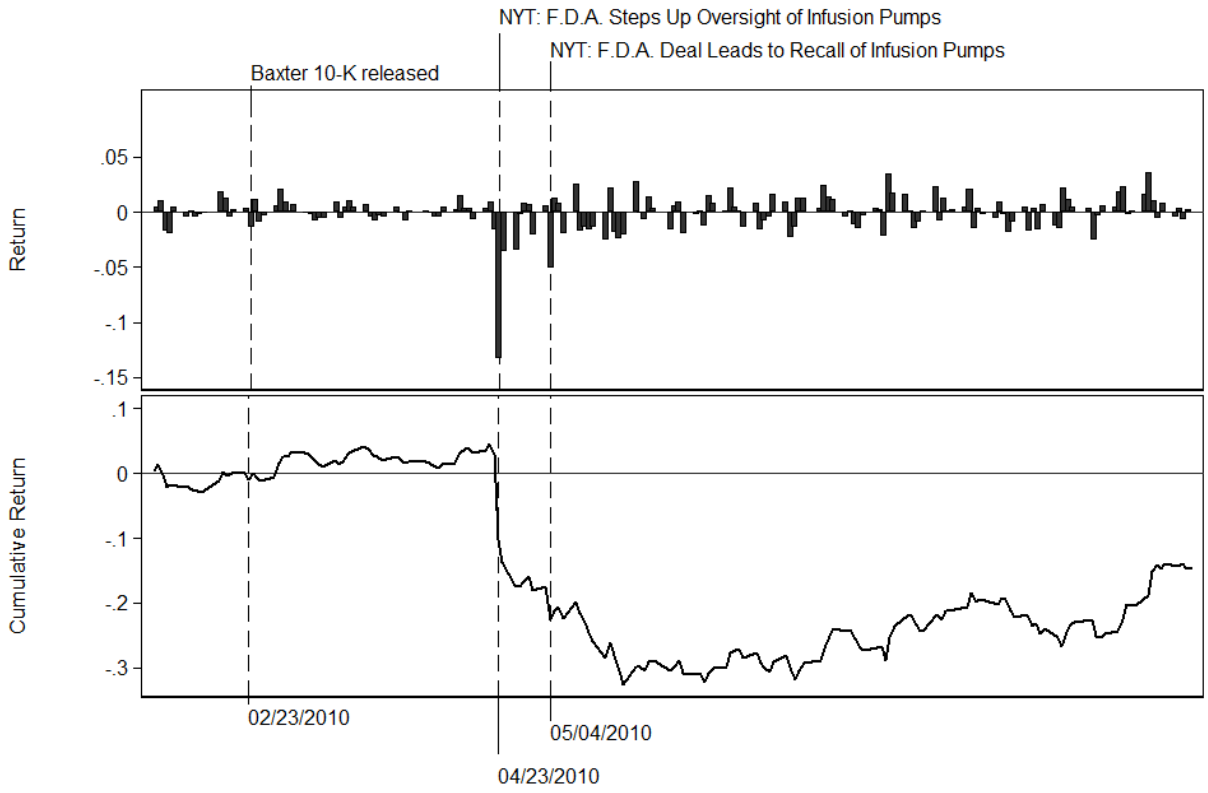


Figure 5: Important Keywords

This table reports the count of keywords that are related to events related to the recall of Baxtrer's Colleague pumps in 2010.

Word counts	**2007 10-K**	**2008 10-K**	**2009 10-K**
FDA	33	28	48
Recall	16	20	30
Colleague Pump	29	28	79

Figure 6: Example Passages and the changes made to them from Baxter's 10-Ks in 2008 and 2009

2008:	2009:
With respect to COLLEAGUE, the company remains in active dialogue with the FDA about various matters, including the company's remediation plan and reviews of the Company's facilities, processes and quality controls by the company's outside expert pursuant to the requirements of the company's Consent Decree. The outcome of these discussions with the FDA is uncertain and may impact the nature and timing of the company's actions and decisions with respect to the COLLEAGUE pump. The company's estimates of the costs related to these matters are based on the current remediation plan and information currently available. It is possible that additional charges related to COLLEAGUE may be required in future periods, based on new information, changes in estimates, and modifications to the current remediation plan as a result of ongoing dialogue with the FDA.	The company remains in active dialogue with the FDA regarding various matters with respect to the company's COLLEAGUE infusion pumps, including the company's remediation plan and reviews of the company's facilities, processes and quality controls by the company's outside expert pursuant to the requirements of the company's Consent Decree. The outcome of these discussions with the FDA is uncertain and may impact the nature and timing of the company's actions and decisions with respect to the COLLEAGUE pump. The company's estimates of the costs related to these matters are based on the current remediation plan and information currently available. It is possible that substantial additional charges, including significant asset impairments, related to COLLEAGUE may be required in future periods, based on new information, changes in estimates, and modifications to the current remediation plan.

2008:	2009:
In the third quarter of 2008, as a result of the company's decision to upgrade the global pump base to a standard software platform and other changes in the estimated costs to execute the remediation plan, the company recorded a charge of $72 million. This charge consisted of $46 million for cash costs and $26 million principally relating to asset impairments and inventory used in the remediation plan. The reserve for cash costs primarily consisted of costs associated with the deployment of the new software and additional repair and warranty costs. The following summarizes cash activity in the company's COLLEAGUE and SYNDEO infusion pump reserves through December 31, 2008.	In the third quarter of 2008, as a result of the company's decision to upgrade the global pump base to a standard software platform and other changes in the estimated costs to execute the remediation plan, the company recorded a charge of $72 million. This charge consisted of $46 million for cash costs and $26 million principally relating to asset impairments and inventory used in the remediation plan. The reserve for cash costs primarily consisted of costs associated with the deployment of the new software and additional repair and warranty costs. In 2009, the company recorded a charge of $27 million related to planned retirement costs associated with SYNDEO and additional costs related to the COLLEAGUE infusion pump. This charge consisted of $14 million for cash costs and $13 million related to asset impairments. The reserve for cash costs primarily related to customer accommodations and additional warranty costs. The charges were recorded in cost of sales in the company's consolidated statements of income, and were included in the Medication Delivery segment's pre-tax income. The following summarizes cash activity in the company's COLLEAGUE and SYNDEO infusion pump reserves through December 31, 2009.

Figure 7: Event Time Returns

This figure plots the average cumulative abnormal return for the top (highest similarity) and bottom (lowest similarity) quintile portfolios. We compute quintiles based on the prior year's distribution of similarity measures across all stocks. Abnormal return is return adjusted for market return. Events are dates of public release of a 10-K or a 10-Q. *Figure 7-A* shows the average monthly cumulative abnormal returns for month one to six. *Figure 7-B* shows the average daily cumulative abnormal returns for 10-day-before to 10-day-after the public release of a 10-K or a 10-Q event.

Figure 7-A

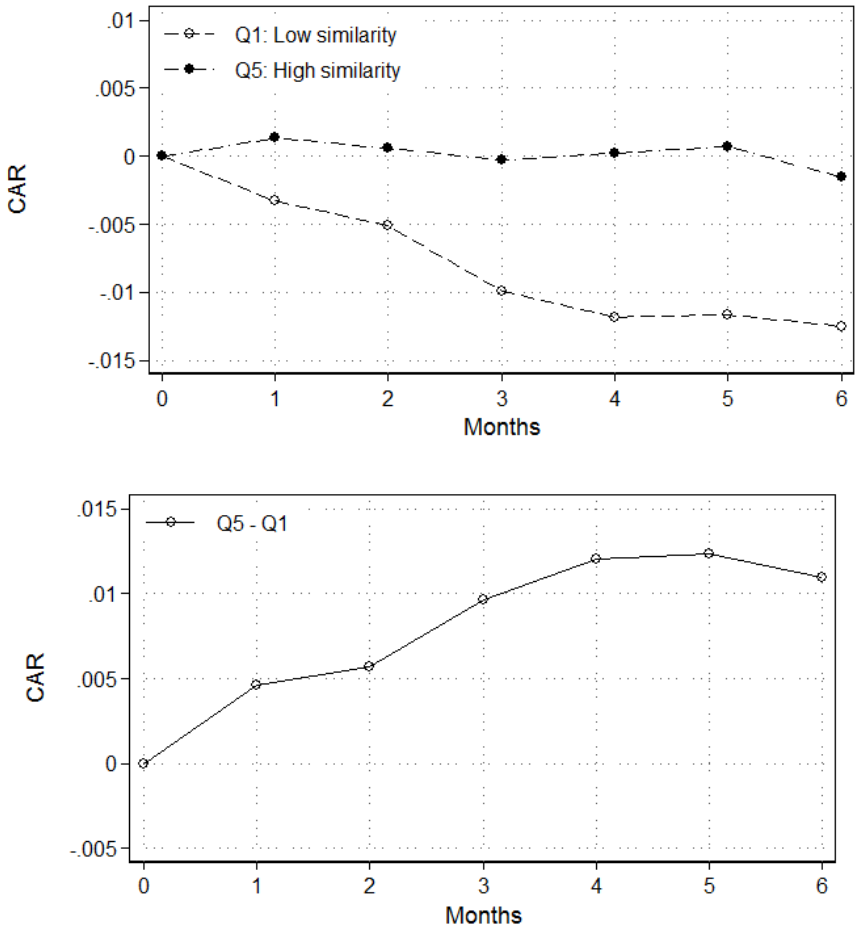


Figure 7-B

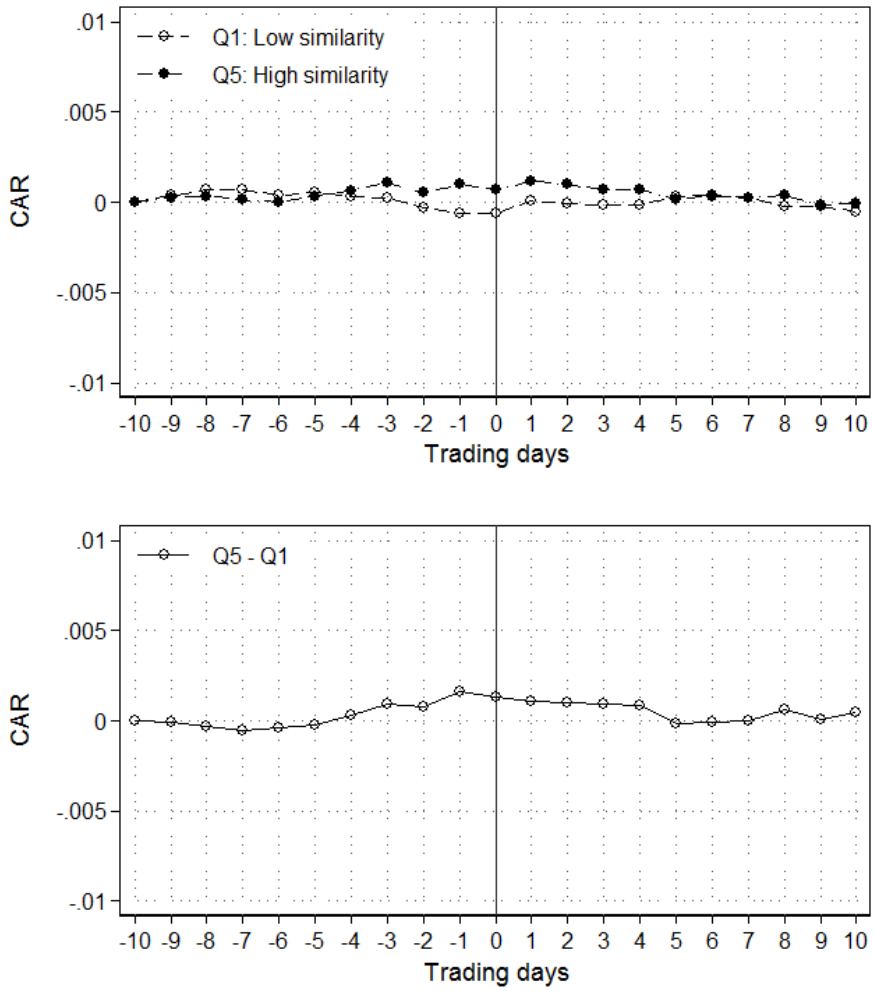

Q1: Low similarity
Q5: High similarity
CAR
.01
.005
0
-.005
-.01
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Trading days
Q5 - Q1
CAR
.01
.005
0
-.005
-.01
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10
Trading days

Figure 8: Section Definitions in 10Ks and 10-Qs

	Form 10-K
Item 1	Business
Item 1A	Risk Factors
Item 2	Properties
Item 3	Legal Proceedings
Item 4	Mine Safety Disclosures
Item 5	Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities
Item 6	Selected Financial Data
Item 7	Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 7A	Quantitative and Qualitative Disclosures About Market Risk
Item 8	Financial Statements and Supplementary Data
Item 9	Changes in and Disagreements With Accountants on Accounting and Financial Disclosure
Item 9A	Controls and Procedures
Item 9B	Other Information
Item 10	Directors, Executive Officers and Corporate Governance
Item 11	Executive Compensation
Item 12	Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters
Item 13	Certain Relationships and Related Transactions, and Director Independence
Item 14	Principal Accounting Fees and Services

	Form 10-Q
Item 1	Financial Statements
Item 2	Management's Discussion and Analysis of Financial Condition and Results of Operations
Item 3	Quantitative and Qualitative Disclosures About Market Risk
Item 4	Controls and Procedures
Item 21	Legal Proceedings
Item 21A	Risk Factors
Item 22	Unregistered Sales of Equity Securities and Use of Proceeds
Item 23	Defaults Upon Senior Securities
Item 24	Mine Safety Disclosures
Item 25	Other Information

Figure 9: Change by Section

Figure 9A reports the average Jaccard similarity for different sections of a firm's 10-K. Section definitions can be found in Figure 8. Figure 9B reports the average Jaccard similarity for different sections of a firm's 1Q-K.

Figure 9-A: Change by Section – 10K

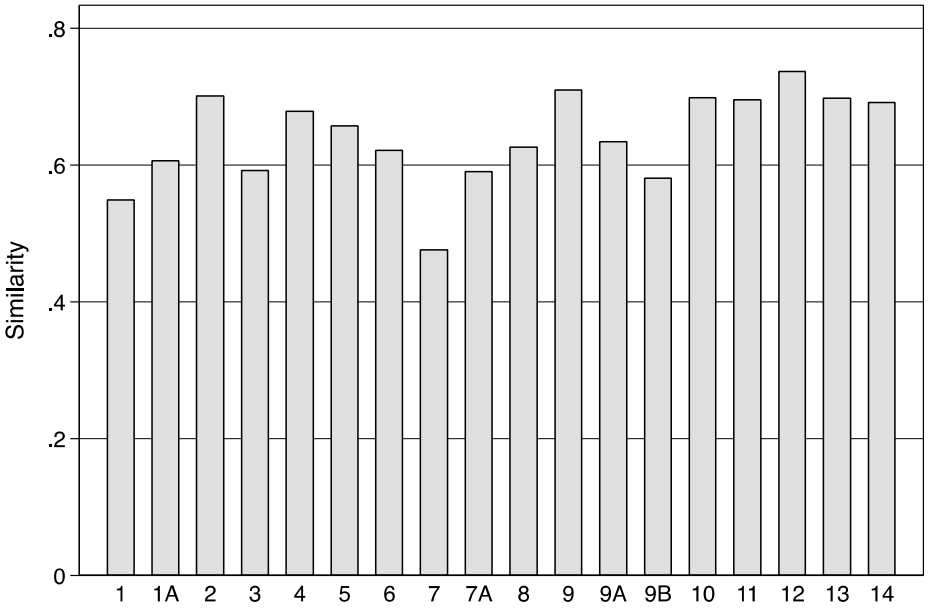


Figure 9-B: Change by Section – 10Q

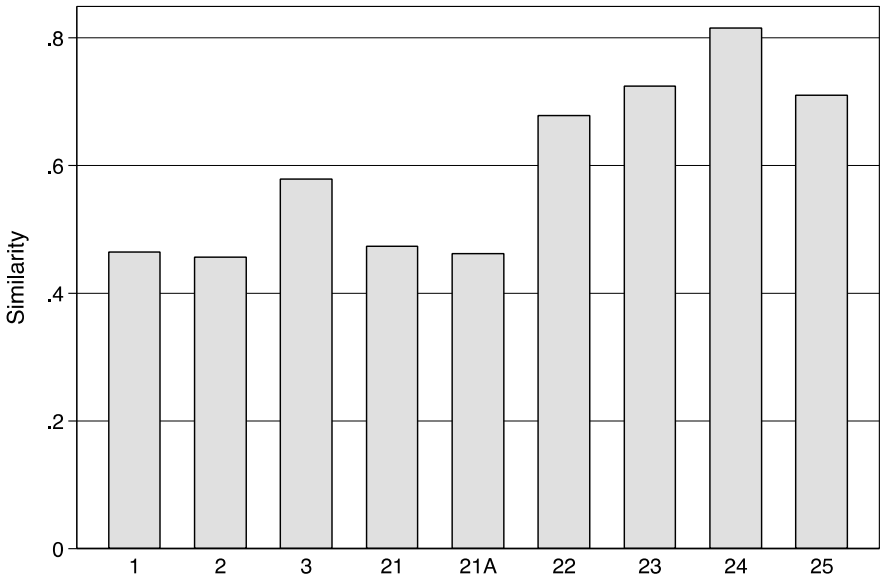


Figure 10: 5-factor Alphas for Portfolio Sort - by Important Common Sections for 10-Ks and 10-Qs

This Figure reports the 5-factor alphas (market, size, value, momentum, and liquidity) of the top (highest similarity) minus bottom (lowest similarity) quintile portfolio (Q5 – Q1) for the common sections of a firm's 10-K and 10-Q financial report: Management's Discussion and Analysis, Legal Proceedings, Quantitative and Qualitative Disclosures About Market Risk, Risk Factors, and Other Information:

10K/Item7+10Q/Item2: Management's Discussion and Analysis of Financial Condition and Results of Operations
10K/Item3+10Q/Item2.1: Legal Proceedings
10K/Item7A+ 10Q/Item3: Quantitative and Qualitative Disclosure About Market Risk
10K/Item1A+10Q/Item2.1A: Risk Factors
10K/Item9B+10Q/Item2.5: Other Information

Similarity measures for each section are computed using only the textual portion in that section. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity measures across all stocks. Stocks then enter the quintile portfolio in the month after the public release of one of their 10-K or 10-Q reports. Firms are held in the portfolio for 3 months.