Lazy Prices*

Lauren Cohen
Harvard Business School and NBER

Christopher Malloy
Harvard Business School and NBER

Quoc Nguyen
University of Illinois at Chicago

This Draft: February 10, 2016
First Draft: November 11, 2015

* We would like to thank Christopher Anderson, Karl Diether, Irem Dimerci, Umit Gurun, Bryan Kelly, Asaf Manela, Ernst Maug, Craig Merrill, Mike Minnis, Taylor Nadauld, Alexandra Niessen-Ruenzi, Stefan Ruenzi, and seminar participants at Brigham Young University, Fuller and Thaler Asset Management, University of Edinburgh, University of Kansas, University of Mannheim, the Chicago Booth Asset Pricing Conference, and the United States Securities and Exchange Commission for helpful comments and discussions. We are grateful for funding from the National Science Foundation.

Lazy Prices

ABSTRACT

We explore the implications of a subtle "default" choice that firms make in their regular reporting practices, namely that firms typically repeat what they most recently reported. 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 the firm. Changes to the language and construction of financial reports have strong implications for firms' future returns: a portfolio that shorts "changers" and buys "non-changers" earns up to 188 basis points per month (over 22% per year) in abnormal returns in the future. These reporting changes are concentrated in the management discussion (MD&A) section. Changes in language referring to the executive (CEO and CFO) team, or regarding litigation, are especially informative for future returns.

JEL Classification: G12, G14, G02

Key words: Information, default behavior, inertia, firms, annual reports, lazy investors

All agents face repetitive tasks. This includes financial agents. CEOs, CFOs, accountants, and treasurers alike face the decision of how to deal with these repetitive tasks. One response is to codify rules that lower the cost of performing these tasks. For instance, form e-mails, saved user information, and automatic-payment plans are examples of responses to repetitive tasks. And while there are efficiency arguments for codifying repetitive tasks, what this means observably is that it can take a large perturbation to call agents into action (and out of inaction) in order to "break" the rules designed to perform these repetitive tasks. While most studies have focused on the effects this will have on inertia, and try to measure that inertia, here we do the opposite. Specifically, we examine situations where agents actively intervene, break inertia, and fail to take the path of least resistance; and then analyze the information content of these actions.

We focus on the behavior of corporations, and show that when firms break from former language or well-codified text in their annual and quarterly reports, that there is a substantial amount of information embedded in this action for important future firm outcomes.

Out approach is grounded in a long line of research in psychology and behavioral economics that stresses the importance of "default" choices. A default typically refers to a baseline choice, setting, or policy that is pre-selected. Defaults generally take effect when an agent fails to make an active choice, fails to update his selection, or fails to "opt out" of a given baseline selection. Defaults can be critical because of a combination of behavioral and situational factors, most notably inertia. Inertia in decision-making has been demonstrated in many settings, and can be especially prevalent when an agent is faced with complex tasks.

In response to a growing body of evidence on the widespread nature and potentially problematic implications of inertial behavior, many interventions focusing on "smart defaults" have been implemented. For example, in areas ranging from lab experiments, to organ donation settings on drivers' licenses, to 401k retirement plan menu choices, the use of carefully constructed defaults designed to offer agents and social planners a more "desirable" set of outcomes in the event that all agents are inert—i.e., fail to make active choices--has become increasingly popular.

In this paper we explore the implications of default choices in an entirely non-experimental setting, by analyzing the behavior of corporations. We focus specifically on a setting where defaults appear to be commonly used by firms, namely in their reporting decisions. We show that the particular construction of firms' annual and quarterly reports suggests that firms are using simple default choices, the most obvious of which is simply repeating the information that they previously reported to the markets.

Consistent with the experimental evidence on the importance of active choices, we show that when firms do make an active decision to significantly change the wording and language choices embedded in their quarterly and annual reports, that these active changes have large but subtle implications for future firm behavior, and future firm outcomes.

To better understand our approach, consider the example of NetApp, Inc. NetApp is an American computer storage, big data, and visualization company that competes broadly in the data management space. The company was founded in 1994 as Network Appliance, Inc. (later changed to NetApp, Inc.) and is headquartered in Sunnyvale, California. The company had historically had Annual Reports (10-k's) that were very

similar across time, but something changed in 2011. This can be seen in Figure 1, which shows the similarity between NetApp's 10k from year to year.

What caused 2011's 10-k to veer from the prior year in terms of the language used and information given? Figure 2 shows a number of news headlines that flooded media nearly 6 months following the release of the 10-k (NetApp's 10-k was publicly released on June 23, 2011). They allege that the Syrian government had been using NetApp equipment to conduct numerous intelligence-gathering and stealth-surveillance activities against its citizens. The articles claim that they acquired NetApp equipment through an Italian re-seller (Area SpA) of their equipment in violation of the sanctions that were in place against doing business with the Syrian government. Further, the articles document actions - and email evidence regarding them – sent in early 2011, before the filing of NetApp's 10-k. This culminated in Senator Robert P. Casey Jr (PA), Senator Mark Kirk (IL), and Senator Christopher A. Coons (DE) authoring an open letter requesting a formal federal inquiry into NetApp requesting that "pending conclusion of an investigation, officials consider suspending all U.S. government work with NetApp, which received more than $111 million in U.S. contracts since 2001."[1]

The question then is whether these two were at all linked – i.e., could something about the changes to the 10-k have hinted at the portending government inquiry. Figure 3 provides suggestive evidence on this point. It shows a number of parallel passages: the 2010 version of the passage vs. the 2011 version. From Panel A, for instance, you see that NetApp changed the passage:

"The failure to comply with U.S. government regulatory requirements could subject us to fines and other penalties, which could have a material adverse effect on our revenues,

[1] The link to all of the full-length articles, as well as the full Senatorial letter, are included in Figure 2.

operating results and financial position." [2010]

to:

"Failure to comply with U.S. government regulatory requirements ***by us or our reseller partners*** *could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position."* [2011]

in 2011, in clear reference to the liability that their reseller Area SpA exposed them to through resale of NetApp's equipment. Panels B-D report similar revealing statements through uses of phrases such as the change of:

"*We are a party to lawsuits in the normal course of our business..."* [2010]

to

"We may be a party to lawsuits and other claims in the normal course of our business from time to time, including... ***governmental and other regulatory investigations and proceedings.****"* [2011]

In addition, they inserted this phrase newly in 2011: "***We are currently discussing contract compliance matters regarding sales made through a channel partner with the DOJ and GSA...,"*** again referencing the Syria resale of their equipment. Again, all of these were published as new additions to the otherwise nearly identical paragraphs of their 10-k six months before any news story broke.

Lastly, would being aware of the changes in the 10-k have made a difference to investors? Figure 4 shows that the answer is yes. NetApp dropped 20% over the 6 months between the *public* release of the 10-k and the first news broke of the Syria connection and inquiry.

We demonstrate that this pattern of behavior and subsequent events is systematic across the entire cross-section of U.S. publicly traded firms from 1995 to 2014. First we

show that firms rarely make substantive changes to the language and construction of their reports. Employing a variety of textual analysis approaches, we demonstrate that only a small percentage of firms make large changes to their reports on a quarterly or annual basis.

Next we explore the implications of these active changes. We show that firms that change their reports in a significant fashion are associated with lower future returns. In particular, a portfolio that goes long "non-changers" and short "changers" earns a statistically significant 30-60 basis points per month – up to 7.6% per year (t=4.44) - in 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 over the following 12-18 months after the reporting change.

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.

We also 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 instance, the 5-factor alpha on (Non-Changers – Changers) particularly in this 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 also show that firms that hire outside lawyers produce filings that look decidedly more similar each year, relative to firms that use in-house counsel (indicating that outside counsel is more inert, less inclined to make changes, or perhaps less informed).

What we find most intriguing about these results is that they 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. That simple comparison, as we show throughout the paper, contains rich information for the future of a firm's operations.

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 adds to a growing literature examining the information content of firms' disclosure choices. Initially 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), Barron, Kile, and O'Keefe (1999)) in the MD&A sections.[2]

More recently, as a result of increased computing power and advances in the field of natural language processing, the focus has shifted to more automated forms of textual analysis. 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 (2010) 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) 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. Meanwhile Feldman et al. (2010) find that a positive tone in the MD&A section is associated with higher contemporaneous and future returns, and that an increasingly negative tone is associated with lower contemporaneous returns.[3] Closest to our paper is perhaps Brown and Tucker (2011), who focus on year-on-year changes (as opposed to

[2] See Cole and Jones (2005) and Feldman et al. (2010) for a survey of the evidence.
[3] See also Muslu et al. (2009); and Li (2011) for a survey of various textual analysis approaches.

levels) 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. Our paper is unique in that we explore a simple set of measures that capture firm-level changes in reporting behavior across the entire 10-K and 10-Q, not just the MD&A sections; and more importantly we explore the impact of these changes on *future* stock returns and *future* litigation events, rather than past or contemporaneous events.

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[4] from 1994 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 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.[5]

We obtain monthly stock returns from the Center for Research in Security Prices

[4] (https://www.sec.gov/edgar/)
[5] Bill McDonald provides a very detailed description on how to strip 10-K/Q down to text files: http://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/privides

(CRSP) and firms' book value of equity and earning 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.[6]

We measure the quarter-on-quarter similarities between 10-Q and 10-K filings using four different similarity measures: cosine similarity, Jaccard similarity, minimum edit distance, and simple similarity. The first measure is called the cosine similarity between two documents D_1 and D_2 and is computed as follow. 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:

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

where $nD_1(t_1)$ is the number of occurrences of term t_i in D_1 and $nD_1(t_1)$ is the number of occurrences of term t_i in D_2. The cosine similarity between two documents is defined as:

$$Sim\_Cosine = \boldsymbol{D_1^{TF} * D_2^{TF}} \ /\ ||\boldsymbol{D_1^{TF}}||\text{x}||\boldsymbol{D_2^{TF}}||$$

where the dot product, *, is the scalar product and norm, || ||, 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:

[6] http://www3.nd.edu/~mcdonald/Word_Lists.html

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

The term frequency vectors of D_1 and D_2 are:

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

The cosine similarity score of D_A and D_B is therefore:

$$\begin{aligned} Sim\_Cosine(D_A, D_B) &= \boldsymbol{D_A}^{\boldsymbol{TF}} * \boldsymbol{D_B}^{\boldsymbol{TF}} \,/\, ||\boldsymbol{D_A}^{\boldsymbol{TF}}||\text{x}||\boldsymbol{D_B}^{\boldsymbol{TF}}|| \\ &= (1\text{x}1+1\text{x}1+0\text{x}1+1\text{x}1+1\text{x}1+1\text{x}1) \\ &\quad [\sqrt{1^2+1^2+1^2+1^2+1^2})\text{x}(\sqrt{1^2+1^2+1^2+1^2+1^2+1^2})] \\ &= 0.91 \end{aligned}$$

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:

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

The cosine similarity score of D_A and D_C is therefore:

$$\begin{aligned} Sim\_Cosine(D_A, D_C) &= \boldsymbol{D_A}^{\boldsymbol{TF}} * \boldsymbol{D_C}^{\boldsymbol{TF}} \,/\, //\boldsymbol{D_A}^{\boldsymbol{TF}}||\text{ x}||\boldsymbol{D_C}^{\boldsymbol{TF}}|| \\ &= (1\text{x}1+1\text{x}1+1\text{x}0+1\text{x}0+1\text{x}0+0\text{x}1+0\text{x}1+0\text{x}1) \\ &\quad [\sqrt{1^2+1^2+1^2+1^2+1^2})\text{x}(\sqrt{1^2+1^2+1^2+1^2+1^2})] \\ &= 0.40 \end{aligned}$$

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:

$$Sim\_Jaccard = |D_1^{TF} \cap D_2^{TF}| \ / \ |D_1^{TF} \cup D_2^{TF}|$$

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

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

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

The third similarity measure we employ is called *Sim_MinEdit* (also known as *Sim_String)* 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", "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 Words 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 next count the number of words in those changes, additions, and deletions and normalized the total count by the size of old document D_1.

In our mechanism tests, we measure the sentiment of document changes by counting the number of positive words minus the number of negative words in the changes between the old document and the new document, normalized by the size of the changes. We further compute the uncertainty and litigious nature of the change by counting the number of words categorized as uncertainty and litigious, respectively, normalized by the size of the changes. Sentiment category identifiers (e.g., negative, positive, uncertainty, litigious) are taken from Loughran and McDonald (2011)'s Master Dictionary.

We parse 10-K/Q documents for mentioning of CEO or CFO turnover and define two indicator variables Change CEO and Change CFO that take the value of 1 if the 10-K/Q documents mention a change in CEO or CFO.

Lastly, we obtain firms' auditor information from AuditAnalytics. However, AuditAnalytics only covers approximately one third of CRSP/Compusat universe, we further collect firms' auditor information directly from 10-K and 10-Q documents by parsing and capturing auditor names in sections that contain the phrase or variation of the phrase "Independent Registered Public Accounting Firm."

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 words in each report, and the *Size of Change* refers to the number of words 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). Table I shows that the average 10-K contains 308,633 words, while the average 10-Q contains roughly one-third as many words (111,789).

As noted above, 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.[7] In Table I, 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, *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 I shows that CEO and CFO changes are mentioned in roughly 2-5% of the reports, on average.

Table II 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 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.

[7] These words are available at: (http://www3.nd.edu/~mcdonald/Word_Lists.html)

Panel B 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. Our hypothesis is that large changes in reporting, when they do occur, will have significant implications for firms' future actions and outcomes, given the tendency of firms to simply report what they previously reported (i.e., to not change their reports).

We begin by analyzing the future stock returns associated with firms who change their reports, 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 January or February.[8] Similarity scores are computed relative to the prior year report that lines up in calendar time with the report in question (such that 2004 Q1 10-Qs are compared with 2005 Q1 10-Qs, for example). Stocks enter the portfolio in the month after the public release of one of their reports, which induces a lag in our portfolio construction. Firms are held in the portfolio for 3 months. Portfolios are rebalanced monthly, and the returns are reported in Table III.

Panel A of Table III 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) 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 18-46 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. Notably, all 4 measures of similarity deliver this same finding, suggesting that our results are not driven by the particular way we compute year-over-year changes in the documents. The interpretation of this finding is 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.

[8] 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 July and August; and for the other 3 calendar quarters we would analyze their 10-Qs instead.

Panel B of Table III 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 63 basis points per month (t=4.45), depending on the similarity measure employed.

Panel B of Table III also shows that the majority of the L/S spread comes from the short side of the portfolio. For example, using the Jaccard similarity measure, the Q1 short portfolio earns -44 basis points per month (t=4.56), while the Q5 long portfolio earns only +19 basis points (t=1.87).

B. Fama-MacBeth Regressions

Next we run monthly Fama-MacBeth cross-sectional regressions of 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 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, such as last month's (or last quarter's) standardized unexpected earnings surprise (SUE). SUE is computed as actual earnings per share minus average analyst forecast earnings per share, divided by the standard deviation of the forecasts.

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.

C. Long-Term Event Returns

We also examine longer-term returns by computing cumulative event-time returns extending out one year after the release of each document. Figure 5 shows the average cumulative abnormal return for each quintile portfolio sorted based on firms similarity scores (here the Sim_Simple measure is used), for 1 month to 12 months after portfolio formation. Figure 5 shows that returns accrue gradually over the course of the subsequent year, and do not reverse. Additionally, the long-term poor performance of Q1 (the "changers") is particularly striking in this figure. Taken as a while, Figure 5 suggests that the information contained in a firm's decision to significantly change its reporting practices has a long-lasting impact on firm value.

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 decreases in similarity across years for a given 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)[9] on these characteristics of the document, with firm and time fixed effects included. Table V shows that lower similarity across documents is associated with lower sentiment, higher uncertainty, more litigiousness, and more frequent mentions of CEO and CFO changes. Each of these findings is highly statistically significant, and suggests that substantive changes in reporting practices are associated with significant changes in the operations or prospects of the firm in question.

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 6 lists the standard sections that are present in firms' annual (10-K) and

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

quarterly (10-Q) reports, respectively. Figure 7 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 8 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 relative to the other items in the report. Collectively, these figures indicate that changes in reporting practices, to the extent that they exist, are concentrated in the key sections of the reports over which management exercises the most discretion, as opposed to in purely mechanical/legal sections or in the sections that simply report standard disclosures or procedures.

C. Return Predictability of Key Sections of Reports

We then take the item/section categories listed in Figure 6 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 VI. Table VI 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 12-20 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 returns (Non-Changers minus Changers) of up to 188 basis points per month (t=2.76).[10] 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). Table VII 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.

D. Double-Sorts on Characteristics of the Documents

Table VIII then takes these three variables and performs double-sorts of the calendar time portfolio tests shown earlier in Table III. Specifically, Table VIII reports the calendar-time value-weight portfolio returns 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 III. Table VIII shows that the return results documented earlier are concentrated in the Low Sentiment, High

[10] Note that this return result is still large in magnitude (over 100 basis points per month) and highly statistically significant even in the post-Sarbanes Oxley (2003-onward) sample period.

Uncertainty, and High Litigiousness subsamples. For instance, the L/S spread 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.

E. The Influence of Specific Law Firms

In this section we explore the impact of law firm characteristics on our firm-level results. Since lawyers and the law firms they work for are the agents who ultimately execute the changes in these documents, we examine the cross-sectional heterogeneity in document similarity according to law firm type. To do so, we extract and hand-code law firm names from 10-Ks and 10-Qs and create a standardized list that corrects for slight differences and errors in law firm listings across filings. In examining the impact of law firms, we find two interesting results. First, as shown in Panel A of Table IX, we find that outside lawyers, as opposed to in-house lawyers, are associated with significantly higher document similarity at the firm level. This suggests that outside lawyers are either more inert (i.e., lazier), less inclined to make substantive changes, or perhaps less informed; the net effect of this tendency is that firms represented by outside lawyers on average report filings that change very little from year-to-year relative to firms represented by in-house counsel.

Second, as shown in Panel B of Table IX, if we re-run our baseline similarity regressions from Table V (which seek to explain the factors that predict document similarity), but include law-firm fixed effects in these regressions—in addition to the time- and firm-fixed effects already present, we find a substantial increase in the R^2 in these regressions. Further, an F-test on the joint significance of law firm fixed effects in

these regressions is highly significant. Collectively these results indicate that the behavior of specific law firms, as well as specific law firm types (in-house versus outside counsel), is important in understanding the cross-sectional differences in reporting behavior by corporations.

F. 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 profit, and free cash flow. Table X 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). 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 explore a subtle, but economically important "default" choice that firms make in their regular reporting practices, namely that firms overwhelmingly repeat what they most recently reported. Further, we find that when firms break with routine - breaking from former language, sections, etc. in their annual and quarterly reports – that this action contains rich, important information for future firm outcomes.

A portfolio that shorts "changers" and buys "non-changers" in annual and quarterly financial reports earns 30-60 basis points per month over the following year. The returns continue to accrue out to 18 months, and do not reverse, implying that these return movements are 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. Changes in language referring to the executive (CEO and CFO) team, and about litigation, are especially informative for future returns. Litigation language changes, for instance, imply underperformance relative to non-changers of over 8.5% per year (t=3.29). Reporting changes are concentrated in the management discussion (MD&A) section, which is the portion of the documents where management has the most discretion. However, more subtle changes outside the MD&A section, for example in the Risk Factors section of the 10-k, have even larger predictability for future returns (with the long-short portfolio earning up to 188 basis points per month, or 22% annually).

The systematic patterns we document throughout the paper are consistent with a differential level of "laziness" of investors with respect to text relative to numerical financial statement entries. Investors appear quite adept at comparing numerical items across years – with financial reports themselves well-constructed for exactly these comparisons. In contrast, our evidence suggests that investors do not appear to be doing the same "comparison" of this year's text to last, leading to the rich information contained in these differences being largely missed by investors and the market.

Further, our results speak to a broader literature on the power of defaults, and the implications of inertia in decision-making. By taking a twist on this literature – examining *breaks* from default behavior – we show that deviations from default behavior can have rich information for future outcomes. In an entirely non-experimental setting, across thousands of firms and almost 20 years of data, breaks from default behavior have large implications for corporations, and asset prices more generally. Given the pervasiveness of inertia in agents' behavior across settings, and the amount of individual behaviors that are subject to these defaults within firms, the implications of *breaks* from these default behaviors in the corporate setting provide a critical, yet understudied area, in both corporate finance and asset pricing.

References

Agrawal, Anup, Sahiba Chadha and Mark A. Chen, 2006, Who is afraid of Reg FD? The behavior and performance of sell-side analysts following the SEC's Fair Disclosure Rules, *Journal of Business* 79, 2811-2834.

Barron, O. E.; C. O. Kile; and T. B. O'Keefe. "MD&A Quality as Measured by the SEC and Analysts' Earnings Forecasts." *Contemporary Accounting Research* 16 (1999): 75–109.

Beaver, William, 1968, The information content of annual earnings announcements, *Journal of Accounting Research* 6, 67-92.

Bergman, Nittai and Sugata Roychowdury, 2008, Investor Sentiment, Expectations, and Corporate Disclosure, *Journal of Accounting Research* 46, 1057-1083.

Chen, Shuping, and Dawn Matsumoto. 2006. Favorable versus unfavorable recommendations: The impact on analyst access to management-provided information, *Journal of Accounting Research* 40 (4): 657-689.

Chen, Shuping, Stephan Hollander, and Kelvin Law, 2014, The price of silence: When no one asks questions during conference calls, Working paper, University of Texas.

Clarkson, P. M.; J. L. Kao; and G. D. Richardson. "Evidence that Management Discussion and Analysis (MD&A) Is a Part of a Firm's Overall Disclosure Package." *Contemporary Accounting Research* 16 (1999): 111–34.

Cohen, Lauren, Andrea Frazzini, and Christopher Malloy, 2010. Sell side school ties, *Journal of Finance* 65, 1409-1437.

Cole, C. J., and C. L. Jones. "Management Discussion and Analysis: A Review and Implications for Future Research." *Journal of Accounting Literature* 24 (2005): 135–74.

Cooper, Rick, Theodore Day, and Craig Lewis, 2001, Following the leader: A study of individual analysts' earnings forecasts, *Journal of Financial Economics* 61, 383-416.

Degeorge, Francois, Jayendu Patel, and Richard Zeckhauser, 1999. Earnings management to exceed thresholds, *Journal of Business* 72, 1-33.

DellaVigna, Stefano, and Joshua Pollet, 2009, Investor inattention and Friday earnings announcements, *Journal of Finance* 64, 109-749.

Diether, Karl B, Anna Scherbina, and Christopher Malloy, 2002, Differences of opinion and the cross-section of stock returns, *Journal of Finance* 57, 2113-2141.

Feldman, R.; S. Govindaraj; J. Livnat; and B. Segal. "Management's Tone Change, Post Earnings Announcement Drift and Accruals." *Review of Accounting Studies* 15 (2010): 915–53.

Frazzini, Andrea, and Owen Lamont, 2006, The earnings announcement month premium and trading volume, Working paper, Yale University.

Gintschel, A. and S. Markov. 2004. The Effectiveness of Regulation FD. *Journal of Accounting and Economics* 37: 293-314.

Groysberg, Boris, Paul Healy, Craig Chapman, Devin Shanthikumar, and Yang Gui (2007), Do buy-side analysts out-perform the sell-side?, Working paper, Harvard University.

Hanley, K. W., and G. Hoberg. "The Information Content of IPO Prospectuses." *Review of Financial Studies* 23 (2010): 2821–64.

Hoberg, G., and G. Phillips. "Product Market Synergies in Mergers and Acquisitions: A Text Based Analysis." Working paper, University of Maryland, 2010.

Hobson, Jessen, William J. Mayew, and Mohan Venkatachalam. 2012. Analyzing Speech to Detect Financial Misreporting, *Journal of Accounting Research*, 50 (2): 349-392.

Hollander, Stephan, Maarten Pronk, and Erik Roelofsen, 2010, Does Silence Speak? An Empirical Analysis of Disclosure Choices During Conference Calls, *Journal of Accounting Research*, Vol. 48, Issue 3, pp. 531-563.

Hirshleifer, David, Kewei Hou, and Siew Hong Teoh, 2012, The accruals anomaly: Risk or mispricing? *Management Science* 58, 320-335.

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

Hong, Harrison, and Jeffrey D. Kubik, 2003, Analyzing the analysts: Career concerns and biased forecasts, *Journal of Finance* 58, 313-351.

Kinney, William, David Burgstahler and Roger Martin, 2002, Earnings Surprise "Materiality" as Measured by Stock Returns, *Journal of Accounting Research* 40, 1297-1329.

Kothari, SP, Susan Shu and Peter Wysocki, 2009, Do managers withhold bad news?, *Journal of Accounting Research* 47, 241-276.

Larcker, David, and Anastasia Zakolyukina, 2012, Detecting Deceptive Discussions in

Conference Calls, *Journal of Accounting Research*, Vol. 50, Issue 2, pp. 495-540.

Li, E., and K. Ramesh. "Market Reaction Surrounding the Filing of Periodic SEC Reports." *The Accounting Review* 84 (2009): 1171–208.

Li, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–47.

Li, F. "The Determinants and Information Content of the Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010): 1049–102.

Li, F. "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?" Working paper (2010b), University of Michigan.

Li, F. "Textual Analysis of Corporate Disclosures: A Survey of the Literature," *Journal of Accounting Literature* (2011), Vol. 29, 143-165.

Lin, Hsiou-wei and Maureen F. McNichols, 1998, Underwriting relationships, analysts' earnings forecasts and investment recommendations, *Journal of Accounting and Economics* 25, 101-127.

Lin, Hsiou-wei, Maureen F. McNichols, and Patricia O'Brien, 2005, Analyst impartiality and investment banking relationships, *Journal of Accounting Research* 43, 623-650.

Loughran, T., and B. McDonald. "Plain English." Working paper, University of Notre Dame, 2008.

Kumar, Alok, 2010, Self-selection and the forecasting abilities of female equity analysts, *Journal of Accounting Research*, 48, 393-435, 2010.

Matsumoto, Dawn, Maarten Pronk, and Erik Roelofsen, 2011, What Makes Conference Calls Useful? The Information Content of Managers Presentations and Analysts Discussion Sessions. *The Accounting Review*: July 2011, Vol. 86, No. 4, pp. 1383-1414.

Mayew, W. 2008. Evidence of Management Discrimination Among Analysts During Earnings Conference Calls. *Journal of Accounting Research* 46 (3): 627-659.

Mayew, William J., and Mohan Venkatachalam. 2012, The Power of Voice: Managerial Affective States and Future Firm Performance, *Journal of Finance*, 67 (1): 1-43.

Mayew, W., N. Sharp, and M. Venkatachalam, 2012, Using Earnings Conference Calls to Identify Analysts with Superior Private Information, *Review of Accounting Studies* (forthcoming).

Michaely, Roni, and Kent L. Womack, 1999, Conflict of interest and the credibility of

underwriter analyst recommendations, *Review of Financial Studies* 12, 653-686.

Mohanran, Partha and Shyam V. Sunder, 2006, How has Regulation FD affected the operations of financial analysts?, *Contemporary Accounting Research* 23, 491-52.

Muslu, V.; S. Radhakrishnan; K. R. Subramanyam; and D. Lim. "Forward-Looking Disclosures in the MD&A and the Financial Information Environment." Working paper, University of Texas at Dallas and University of Southern California, 2009.

Nelson, K. K., and A. C. Pritchard. "Litigation Risk and Voluntary Disclosure: The Use of Meaningful Cautionary Language." Working paper, Rice University and University of Michigan, 2007.

Niessner, Marina, 2013, Strategic disclosure timing and insider trading, Working paper, Yale University.

Rogers, R. K., and J. Grant. "Content Analysis of Information Cited in Reports of Sell-Side Financial Analysts." *Journal of Financial Statement Analysis* 3 (1997): 17–30.

Sloan, R., 1996, Do stock prices fully reflect the information in accruals and cash flows about future earnings?, *Accounting Review* 71, 289-316.

Stickel, Scott, 1992, Reputation and performance among security analysts, *Journal of Finance* 47, 1811-1836.

Zhou, Dexin, 2014, The blame game, Working paper, Emory University.

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

This table reports the summary statistics of 10-Ks and 10-Qs from 1995 to 2014. Document Size is the number of words. Size of Change is the number of words in the Change. 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.

	Count	Mean	SD	Min	Max
Document Size	353735	159873.7	159873.7	20357	5.24e+07
Document Size - 10K	90198	308633	282473	34660	2.43e+07
Document Size - 10Q	263537	114848.4	286663.9	18824	3.14e+07
Sentiment of Change	353735	-.0003371	.0011069	-.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

Table II: Summary Statistics on Similarity Measures

Panel A reports the summary statistics of four different measures of document similarity. Panel B reports the correlation between the four similarity measures. 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

	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 B: 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 III: Main Results – Portfolio

This Table reports the calendar-time portfolio returns. For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores 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. 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). Panel A reports equal-weight portfolio returns and Panel B reports value-weight portfolio returns. ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

Panel A: Equally Weighted

	Sim_Cosine						Sim_Jaccard						
	Q1	Q2	Q3	Q4	Q5	Q5 – Q1		Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0065*	0.0076**	0.0072**	0.0090***	0.0094***	0.0029***	Excess	0.0062	0.0064*	0.0074**	0.0091***	0.0102***	0.0040***
Return	(1.7399)	(2.0505)	(2.1098)	(2.7231)	(2.8340)	(3.0098)	Return	(1.6054)	(1.7128)	(2.1018)	(2.7459)	(3.2587)	(2.9151)
3-Factor	-0.0013*	-0.0004	-0.0005	0.0014*	0.0020***	0.0033***	3-Factor	-0.0017**	-0.0017**	-0.0005	0.0015**	0.0029***	0.0046***
Alpha	(-1.8818)	(-0.5930)	(-0.6561)	(1.7571)	(2.7378)	(4.2597)	Alpha	(-2.1802)	(-2.2939)	(-0.6491)	(2.1239)	(3.7939)	(4.9218)
5-Factor	-0.0011	-0.0001	-0.0004	0.0014*	0.0023***	0.0034***	5-Factor	-0.0015*	-0.0015**	-0.0003	0.0018**	0.0030***	0.0045***
Alpha	(-1.6121)	(-0.1529)	(-0.6093)	(1.8892)	(3.4372)	(4.2956)	Alpha	(-1.9647)	(-2.0640)	(-0.4713)	(2.5846)	(4.1026)	(4.7260)
	Sim_MinEdit						Sim_Simple						
	Q1	Q2	Q3	Q4	Q5	Q5 – Q1		Q1	Q2	Q3	Q4	Q5	Q5 – Q1
Excess	0.0064*	0.0073*	0.0070*	0.0089***	0.0097***	0.0033**	Excess	0.0072*	0.0079**	0.0082**	0.0090***	0.0090***	0.0018
Return	(1.6890)	(1.9512)	(1.9389)	(2.6620)	(3.2834)	(2.4480)	Return	(1.8671)	(2.1185)	(2.3413)	(2.7340)	(3.0359)	(1.2038)
3-Factor	-0.0016**	-0.0007	-0.0009	0.0013*	0.0027***	0.0043***	3-Factor	-0.0008	-0.0002	0.0003	0.0014**	0.0020**	0.0028***
Alpha	(-2.2247)	(-0.9848)	(-1.4068)	(1.7767)	(3.6470)	(5.4785)	Alpha	(-1.0934)	(-0.2075)	(0.3834)	(2.0139)	(2.5730)	(3.2194)
5-Factor	-0.0013*	-0.0005	-0.0006	0.0013*	0.0028***	0.0041***	5-Factor	-0.0006	0.0003	0.0004	0.0016**	0.0021***	0.0027***
Alpha	(-1.9301)	(-0.7328)	(-0.9783)	(1.9446)	(3.8337)	(5.1380)	Alpha	(-0.8898)	(0.3700)	(0.6345)	(2.3037)	(2.6774)	(3.0117)

Panel B: Value Weighted

Sim_Cosine

	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0040	0.0044	0.0051	0.0079**	0.0078**	0.0038***
Return	(1.2095)	(1.3085)	(1.6391)	(2.5627)	(2.3629)	(2.7547)
3-Factor	-0.0018**	-0.0019**	-0.0007	0.0018**	0.0019*	0.0037***
Alpha	(-2.0280)	(-2.1017)	(-0.7910)	(1.9748)	(1.7411)	(2.7024)
5-Factor	-0.0013	-0.0021**	-0.0009	0.0021**	0.0021*	0.0034**
Alpha	(-1.4101)	(-2.2624)	(-1.0640)	(2.3542)	(1.9115)	(2.3996)

Sim_Jaccard

	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0015	0.0055	0.0072**	0.0072**	0.0076**	0.0061***
Return	(0.4459)	(1.6504)	(2.2260)	(2.3058)	(2.5168)	(3.9898)
3-Factor	-0.0046***	-0.0005	0.0012	0.0013	0.0018*	0.0063***
Alpha	(-4.8741)	(-0.4956)	(1.1990)	(1.3893)	(1.6714)	(4.4578)
5-Factor	-0.0044***	-0.0004	0.0014	0.0012	0.0019*	0.0063***
Alpha	(-4.5642)	(-0.3962)	(1.4451)	(1.2487)	(1.8656)	(4.4351)

Sim_MinEdit

	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0036	0.0043	0.0068**	0.0077**	0.0077***	0.0041**
Return	(1.0609)	(1.2900)	(2.0867)	(2.5586)	(2.6093)	(2.4051)
3-Factor	-0.0025***	-0.0018*	0.0007	0.0020**	0.0020*	0.0045***
Alpha	(-2.8874)	(-1.8498)	(0.7883)	(2.1000)	(1.8087)	(3.0695)
5-Factor	-0.0021**	-0.0016	0.0009	0.0020**	0.0012	0.0033**
Alpha	(-2.4416)	(-1.6325)	(1.1168)	(2.1022)	(1.0502)	(2.2778)

Sim_Simple

	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
Excess	0.0024	0.0061*	0.0077**	0.0078**	0.0074**	0.0050***
Return	(0.6879)	(1.8821)	(2.4476)	(2.5284)	(2.4775)	(2.6924)
3-Factor	-0.0039***	0.0002	0.0018*	0.0019*	0.0019	0.0058***
Alpha	(-3.8893)	(0.1802)	(1.8704)	(1.8797)	(1.4452)	(3.5865)
5-Factor	-0.0036***	0.0005	0.0018*	0.0018*	0.0015	0.0051***
Alpha	(-3.4960)	(0.6607)	(1.7835)	(1.7139)	(1.1461)	(3.1419)

Table IV: Main Results – Fama MacBeth Regression

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our 4 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, and Ret(-12, -1) is the cumulative return from month -12 to month -1. SUE is the standardized unexpected earning and computed as actual earning per share minus average analyst forecast earnings per share, divided by the standard deviation of forecasts. ***, **, and * denote significance at 1%, 5%, and 10% levels, 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 reports potential mechanism of 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. ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)
			Sim_Simple		
Sentiment	3.5595***				
	(90.4767)				
Uncertainty		-3.5497***			
		(-33.1870)			
Litigiousness			-0.1264***		
			(-13.2670)		
Change CEO				-0.0076***	
				(-10.4748)	
Change CFO					-0.0086***
					(-8.0932)
Constant	0.1875***	0.1841***	0.1827***	0.1836***	0.1831***
	(28.8477)	(28.0393)	(27.7814)	(27.9144)	(27.8337)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes
R-Squared	0.0816	0.0624	0.0606	0.0610	0.0604
N	338138	338138	338138	338138	338138

Table VI: Portfolio Sort - By Section

This Table reports the calendar-time portfolio returns for various sections of firms' financial reports. 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 scores 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. ***, **, and * denote significance at 1%, 5%, and 10% levels, 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)
Quantitative and Qualitative 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)
Quantitative and Qualitative 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)
Quantitative and Qualitative 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)
Quantitative and Qualitative 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 VII: Post Sarbanes Oxley (2003 - 2014) for the Risk Factors Section.

This Table reports the calendar-time portfolio returns and the risk factors post Sarbanes Oxley (2003-2014). For each of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores 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 Return (return minus risk free rate), Fama-French 3-factor Alphas (market, size, and value), and 5-factor Alphas (market, size, value, momentum, and liquidity) and risk-factor loadings of the top minus bottom quintile portfolio (Q5 – Q1). ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

	Equally Weighted				Value Weighted			
	Sim_Cosine	Sim_Jaccard	Sim_MinEdit	Sim_Simple	Sim_Cosine	Sim_Jaccard	Sim_MinEdit	Sim_Simple
	Excess Return				Excess Return			
Constant	0.0044	0.0111***	0.0062*	0.0060**	0.0091**	0.0086**	0.0064	0.0041
	(1.2723)	(3.1530)	(1.7965)	(1.9816)	(2.3904)	(2.1179)	(1.6268)	(1.1600)
	3-Factor				3-Factor			
Constant	0.0054	0.0115***	0.0073**	0.0070**	0.0101**	0.0096**	0.0080**	0.0059*
	(1.5554)	(3.2078)	(2.0967)	(2.3317)	(2.6119)	(2.3857)	(2.0484)	(1.7184)
MKTRF	-0.1217	-0.0552	-0.1596*	-0.1439*	-0.1512	-0.1808	-0.2621**	-0.2637***
	(-1.3195)	(-0.5811)	(-1.6960)	(-1.7807)	(-1.4511)	(-1.6497)	(-2.4622)	(-2.8154)
SMB	-0.0973	-0.0918	-0.0763	-0.1073	0.0615	0.0426	0.0529	0.0350
	(-0.5783)	(-0.5155)	(-0.4510)	(-0.7380)	(0.3236)	(0.2170)	(0.2777)	(0.2092)
HML	-0.0674	-0.0256	0.0887	0.0474	-0.0621	-0.0886	0.0161	-0.1338
	(-0.4443)	(-0.1634)	(0.5736)	(0.3567)	(-0.3640)	(-0.4931)	(0.0925)	(-0.8694)
	5-Factor				5-Factor			
Constant	0.0056	0.0111***	0.0071**	0.0071**	0.0092**	0.0101**	0.0075*	0.0056
	(1.5894)	(3.0614)	(1.9960)	(2.3269)	(2.3817)	(2.4599)	(1.9307)	(1.6511)
MKTRF	-0.1293	-0.0652	-0.1497	-0.1291	-0.1414	-0.1923*	-0.1863*	-0.1876**
	(-1.3461)	(-0.6623)	(-1.5284)	(-1.5368)	(-1.3164)	(-1.6849)	(-1.7289)	(-1.9978)
SMB	-0.0859	-0.1090	-0.0912	-0.1041	0.0177	0.0274	0.0138	0.0056
	(-0.5013)	(-0.6057)	(-0.5284)	(-0.7030)	(0.0923)	(0.1372)	(0.0733)	(0.0340)
HML	-0.0984	0.0208	0.1263	0.0509	0.0498	-0.1129	0.1349	-0.0351
	(-0.5935)	(0.1206)	(0.7500)	(0.3523)	(0.2699)	(-0.5770)	(0.7301)	(-0.2172)
UMD	-0.0257	-0.0271	0.0331	0.0489	0.0355	0.0831	0.2472***	0.2488***
	(-0.3176)	(-0.3270)	(0.4001)	(0.6893)	(0.3937)	(0.8680)	(2.7340)	(3.1614)
PS_VWF	-0.0282	0.0939	0.0332	-0.0363	0.1526	-0.0414	-0.0169	-0.0520
	(-0.3161)	(1.0185)	(0.3670)	(-0.4680)	(1.5348)	(-0.3932)	(-0.1704)	(-0.6009)

Table VIII: Portfolio Sort – Document Characteristics

This Table reports calendar-time portfolio 5-factor alphas (market, size, value, momentum, and liquidity) 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 of the four similarity measures, we compute quintiles based on the prior year's distribution of similarity scores 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. 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 Litigiousness are the number of words categorized as uncertainty and litigiousness, respectively, normalized by the size of the Change. Sentiment category identifiers (e.g., negative, positive, uncertainty, litigious) are taken from Loughran and McDonald (2011)'s Master Dictionary. ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

		Sim_Cosine						Sim_Jaccard					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
	Low	-0.0009	-0.0049**	-0.0011	0.0001	0.0018	0.0026	-0.0045***	-0.0044***	-0.0024	0.0023	0.0009	0.0054**
Sentiment		(-0.7123)	(-2.4323)	(-0.8359)	(0.0655)	(1.5807)	(1.4798)	(-2.7913)	(-3.1639)	(-1.2370)	-1.6184	-0.6911	-2.4101
	High	0.0017	-0.0022	0.0004	0.0013	0.0021	0.0006	0.0008	0.0004	0.0013	0.0022	0.0015	0.0011
		(1.2713)	(-1.4511)	(0.2767)	(0.9940)	(1.5911)	(0.3044)	-0.6297	-0.266	-0.7833	-1.5338	-1.2704	-0.6093
	Low	-0.0003	-0.0024	0.0012	0.0014	0.0018	0.0021	-0.0023*	-0.0034**	0.002	0.0025*	0.002	0.0044**
Uncertainty		(-0.2047)	(-1.5217)	(0.8707)	(1.0239)	(1.3515)	(1.0751)	(-1.6548)	(-2.0413)	-1.2431	-1.8589	-1.4689	-2.4187
	High	-0.0022*	-0.0007	0.0006	0.0007	0.0005	0.0032*	-0.0054***	-0.001	0	0.0008	0.0013	0.0072***
		(-1.7899)	(-0.4183)	(0.4222)	(0.4518)	(0.4417)	(1.8134)	(-3.1124)	(-0.7230)	(-0.0218)	-0.5928	-1.1628	-3.5092
	Low	-0.0010	-0.0032**	0.0015	0.0018	0.0004	0.0014	-0.0029**	-0.0042***	0.0013	0.0011	0.0016	0.0047**
Litigiousness		(-0.7701)	(-2.0781)	(1.0152)	(1.2306)	(0.3863)	(0.8268)	(-1.9848)	(-2.6452)	-0.774	-0.8267	-1.0496	-2.1829
	High	-0.0023*	-0.0007	0.0010	0.0024*	0.0012	0.0040**	-0.0048***	-0.0011	0.0006	0.0024**	0.002	0.0071***
		(-1.8054)	(-0.4501)	(0.7448)	(1.8381)	(1.0190)	(2.2466)	(-2.7580)	(-0.7463)	-0.3233	-2.0542	-1.5655	-3.2909

		Sim_MinEdit						Sim_Simple					
		Q1	Q2	Q3	Q4	Q5	Q5 - Q1	Q1	Q2	Q3	Q4	Q5	Q5 - Q1
	Low	-0.0036**	-0.0022	0.0016	-0.0008	0.0013	0.0048**	-0.0047***	-0.0024	-0.0001	0.0027**	0.0010	0.0057***
Sentiment		(-2.3516)	(-1.5372)	(1.1200)	(-0.6059)	(0.9551)	(2.1460)	(-3.3643)	(-1.5296)	(-0.1041)	(2.0023)	(0.7035)	(2.6567)
	High	-0.0002	-0.0002	0.0006	0.0004	0.0026*	0.0032	0.0011	0.0006	0.0008	0.0009	0.0020	0.0012
		(-0.1464)	(-0.1844)	(0.4199)	(0.2755)	(1.6932)	(1.5618)	(0.8134)	(0.6002)	(0.5391)	(0.5091)	(1.1541)	(0.5032)
	Low	-0.0033**	0.0004	-0.0015	0.0014	-0.0003	0.0033*	-0.0017	-0.0013	-0.0001	0.0017	0.0022	0.0038*
Uncertainty		(-2.0092)	(0.2767)	(-1.1442)	(0.8347)	(-0.1981)	(1.6723)	(-1.1747)	(-1.0097)	(-0.0768)	(1.3819)	(1.4079)	(1.8473)
	High	-0.0014	-0.0021	0.0012	0.0017	0.0026*	0.0041**	-0.0041**	-0.0008	0.0030***	0.0012	0.0007	0.0051**
		(-1.0799)	(-1.5031)	(0.9572)	(1.2670)	(1.7718)	(2.0624)	(-2.2905)	(-0.6771)	(2.6108)	(0.6432)	(0.3959)	(2.1409)
	Low	-0.0005	-0.0022	-0.0005	-0.0008	0.0032**	0.0038*	-0.0023	-0.0030**	0.0019	-0.0007	0.0016	0.0039*
Litigiousness		(-0.4520)	(-1.3860)	(-0.3590)	(-0.5422)	(2.0016)	(1.9562)	(-1.6448)	(-2.2771)	(1.6493)	(-0.5575)	(1.0031)	(1.8726)
	High	-0.0032*	0.0001	-0.0004	0.0027**	0.0016	0.0051**	-0.0035**	-0.0001	0.0028**	0.0030**	0.0010	0.0049**
		(-1.9640)	(0.0807)	(-0.3698)	(1.9978)	(0.9775)	(2.2169)	(-2.0759)	(-0.1127)	(2.4679)	(2.1654)	(0.6788)	(2.0119)

Table IX: The Influence of Specific Law Firms

This Table reports the impact of law firm characteristics on firm-level similarity scores. We extract and hand-code law firm names from 10-Ks and 10-Qs. Panel A reports the differential effects of in-house versus outside lawyers on firm-level similarity scores. Panel B reports law firm fixed effects on firm-level similarity scores and the F-tests on the joint significance of law firm fixed effects. ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

Panel A

	(1) Sim_Cosine	(2) Sim_Jaccard	(3) Sim_MinEdit	(4) Sim_Simple
InHouseLawyer	-0.0370***	-0.0602***	-0.0266***	-0.0087***
	(-23.8120)	(-41.9617)	(-19.5237)	(-11.8535)
Constant	0.9107***	0.4830***	0.4514***	0.1815***
	(26.8179)	(15.7561)	(15.4419)	(28.5386)
Firm Fixed Effects	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
R-Squared	0.0620	0.1266	0.1197	0.0666
N	411023	411023	411023	415535

Panel B

	(1) Sim_Cosine		(2) Sim_Jaccard		(3) Sim_MinEdit		(4) Sim_Simple	
Law Firm Fixed Effects	No	Yes	No	Yes	No	Yes	No	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-Squared	0.1402	0.1592	0.104	0.125	0.0527	0.0711	0.1031	0.1216
N	88,024	88,024	88,024	88,024	88,024	88,024	86,359	86,359
F-test for joint significance of Law Firm fixed effects	1.2799 Prob > chi2 = 0.0000		1.4371 Prob > chi2 = 0.0000		1.4370 Prob > chi2 = 0.0000		1.3076 Prob > chi2 = 0.0000	
Number of constraints	1901		1901		1901		1885	

Table X: Robustness

This Table reports the Fama-MacBeth cross-sectional regressions of individual firm-level stock returns on our 4 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, and Ret(-12, -1) is the cumulative return from month -12 to month -1. 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). ***, **, and * denote significance at 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)
			Ret	
Sim_Cosine	0.0056***			
	(3.0429)			
Sim_Jaccard		0.0044***		
		(2.7556)		
Sim_MinEdit			0.0024*	
			(1.9725)	
Sim_Simple				0.0173**
				(1.9935)
Size	-0.0002	-0.0001	-0.0001	-0.0001
	(-0.3633)	(-0.2574)	(-0.2565)	(-0.1548)
log(BM)	0.0018	0.0018	0.0017	0.0015*
	(0.7027)	(0.7127)	(0.7017)	(1.6527)
Ret(-1,0)	-0.0299***	-0.0299***	-0.0298***	-0.0340***
	(-4.1986)	(-4.1663)	(-4.1313)	(-4.5371)
Ret(-12,-1)	0.0084**	0.0083**	0.0084**	0.0062**
	(2.4815)	(2.4526)	(2.4867)	(2.2197)
Invest	-0.0044	-0.0042	-0.0043	-0.0046*
	(-1.6277)	(-1.5790)	(-1.5971)	(-1.7823)
GrossProfit	0.0046**	0.0046**	0.0045**	0.0051**
	(2.3736)	(2.3503)	(2.3132)	(2.5002)
FreeCashflow	0.0048	0.0045	0.0049	0.0037
	(1.1198)	(1.0488)	(1.1466)	(0.8334)
Accrual	-0.0113***	-0.0113***	-0.0114***	-0.0058***
	(-2.8590)	(-2.8637)	(-2.8846)	(-2.6575)
Cons	0.0047	0.0071	0.0080	-0.0075
	(0.5067)	(0.7797)	(0.8793)	(-0.3988)
R-Squared	0.0809	0.0812	0.0812	0.0019
N	607864	607864	607864	600075

Figure 1: Example NetApp, Inc. (ticker = NTAP) Similarity Score

This figure plots the similarity score of NetApp, Inc. from 1996 to 2014.

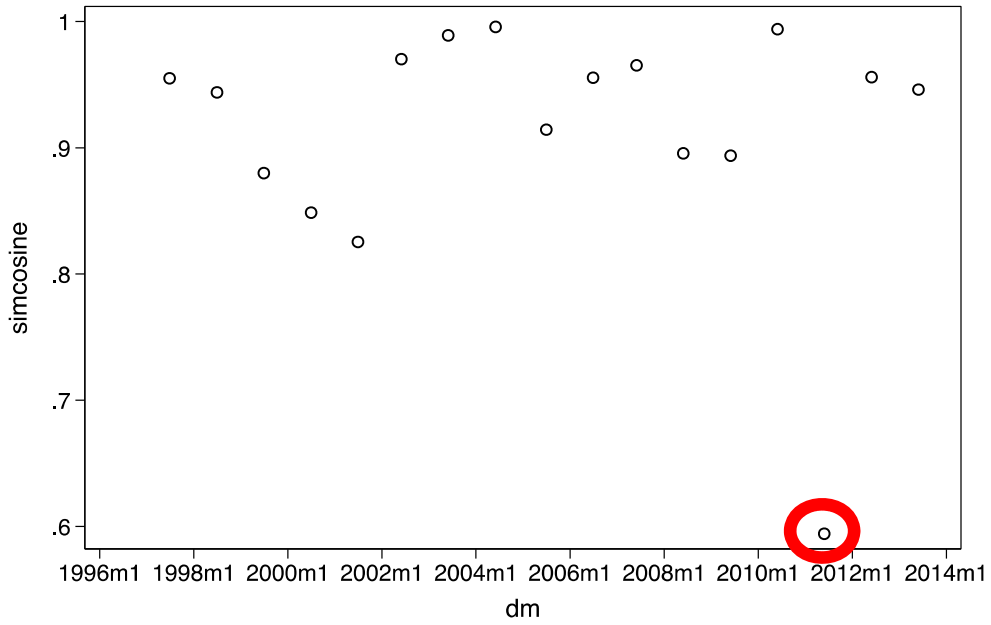

Figure 2: News articles and Congressional Open Letter regarding NetApp's involvement with Syria

1) November 3rd, 2011: "**Syria Crackdown Gets Italy Firm's Aid With U.S.-Europe Spy Gear**", reported that Syrian intelligence agents have contracted Area SpA, an Italian surveillance company, to complete a highly sophisticated system that tracks Internet activity using NetApp equipment.

   http://www.bloomberg.com/news/articles/2011-11-03/syria-crackdown-gets-italy-firm-s-aid-with-u-s-europe-spy-gear

2) November 9th, 2011: **"NetApp Role in Syria Spy Project Spurs Demands for U.S. Inquiry"**

   http://www.bloomberg.com/news/articles/2011-11-10/netapp-role-in-syria-spy-project-spurs-demands-for-u-s-inquiry

   ➔ *Senators Mark Kirk, a Republican from Illinois, and Robert Casey, a Democrat from Pennsylvania, will send a letter today to the State and Commerce departments requesting an investigation into two U.S. companies whose technology has been used to "monitor activities of Syrian citizens," according to a draft of the letter. One of the companies is NetApp, whose role in the Internet surveillance system was detailed in a Nov. 3 article by Bloomberg News.*

   In addition, Representative James McGovern, a Democrat from Massachusetts and co-chairman of the Tom Lantos Human Rights Commission in the House, said he has instructed his staff to follow up with government agencies regarding NetApp to make sure U.S. sanctions against Syria are being enforced.

   "I find it unconscionable that a U.S.-based company's technology is being sent to Syria to help spy on peaceful citizens," McGovern said.

   In their letter, Senators Kirk and Casey ask that pending conclusion of an investigation, officials consider suspending all U.S. government work with NetApp, which received more than $111 million in U.S. contracts since 2001.

3) November 14th, 2011: Senators Mark Kirk (R-IL), Robert Casey (D-PA) and Christopher Coons (D-DE) sent the following letter to the secretary of state and secretary of commerce, asking the administration officials to look into the matter:

   http://www.casey.senate.gov/newsroom/releases/casey-urges-administration-to-investigate-companies-allegedly-aiding-syrian-regime

Figure 3: Example passages and the changes made to them from NetApp, Inc.'s 10-ks in 2010 and 2011

Panel A:

2010 (Old)

The failure to comply with U.S. government regulatory requirements could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.

2011 (New)

Failure to comply with U.S. government regulatory requirements **by us or our reseller partners** could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.

Panel B:

2010 (Old)

We are a party to lawsuits in the normal course of our business, including our ongoing litigation with Sun Microsystems which was recently acquired by Oracle Corporation. Litigation can be expensive, lengthy and disruptive to normal business operations. Moreover, the results of complex legal proceedings are difficult to predict. An unfavorable resolution of a particular lawsuit could have a material adverse effect on our business, operating results, or financial condition.

2011 (New)

We may be a party to lawsuits and other claims in the normal course of our business from time to time, including intellectual property, commercial, product liability, employment, class action, whistleblower and other litigation and claims, and **governmental and other regulatory investigations and proceedings.** Litigation can be expensive, lengthy and disruptive to normal business operations. Moreover, the results of complex legal proceedings are difficult to predict. An unfavorable resolution of a particular lawsuit could have a material adverse effect on our business, operating results, or financial condition.

Panel C:

2010 (Old)

The U.S. government has contributed to our revenue growth and has become an important customer for us. Future revenue from the U.S. government is subject to shifts in government spending patterns. A decrease in government demand for our products could materially affect our revenues. In addition, our business could be adversely affected as a result of future examinations by the U.S. government.

2011 (New)

The U.S. government has contributed to our revenue growth and has become an important customer for us. Future revenues from the U.S. government are subject to shifts in government spending patterns. A decrease in government demand for our

products could materially and adversely affect our revenues. In addition, our business could be adversely affected **by claims that we or a channel partner have failed to comply with regulatory and contractual requirements applicable to sales to the U.S. government.**

Panel D:

2010 (Old)

In addition, selling our products to the U.S. government also subjects us to certain regulatory requirements. For example, in April 2009, we entered into a settlement agreement with the United States of America, acting through the United States Department of Justice ("DOJ") and on behalf of the General Services Administration (the "GSA"), under which we paid the United States $128.0 million, plus interest of $0.7 million, related to a dispute regarding our discount practices and compliance with the price reduction clause provisions of GSA contracts between August 1997 and February 2005. The failure to comply with U.S. government regulatory requirements could subject us to fines and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.

2011 (New)

Selling our products to the U.S. government, **whether directly or through channel partners**, also subjects us to certain regulatory and contractual requirements. **Failure to comply with these requirements by either us or our channel partners could subject us to investigations, fines, and other penalties, which could have a material adverse effect on our revenues, operating results and financial position.** As an example, the United States Department of Justice ("DOJ") and the General Services Administration ("GSA") have in the past pursued claims against and financial settlements with IT vendors, including us and several of our competitors and channel partners, under the False Claims Act and other statutes related to pricing and discount practices and compliance with certain provisions of GSA contracts for sales to the federal government. The DOJ and GSA continue to pursue actively such claims. **We are currently discussing contract compliance matters regarding sales made through a channel partner with the DOJ and GSA, and have produced documents and met with the DOJ and GSA on several occasions. If the DOJ determines to initiate an action against a channel partner and/or us, we would be subject to litigation, could be subjected to fines and penalties. We could also decide to pay the DOJ a settlement, either to avoid a potential action or in termination of an action. Violations of certain regulatory and contractual requirements could also result in us being suspended or debarred from future government contracting.** Any of these outcomes could have a material adverse effect on our revenues, operating results and financial position.

Figure 4: Returns of NetApp, Inc. (ticker = NTAP) in the months following the release of NTAP's 2011 10-k

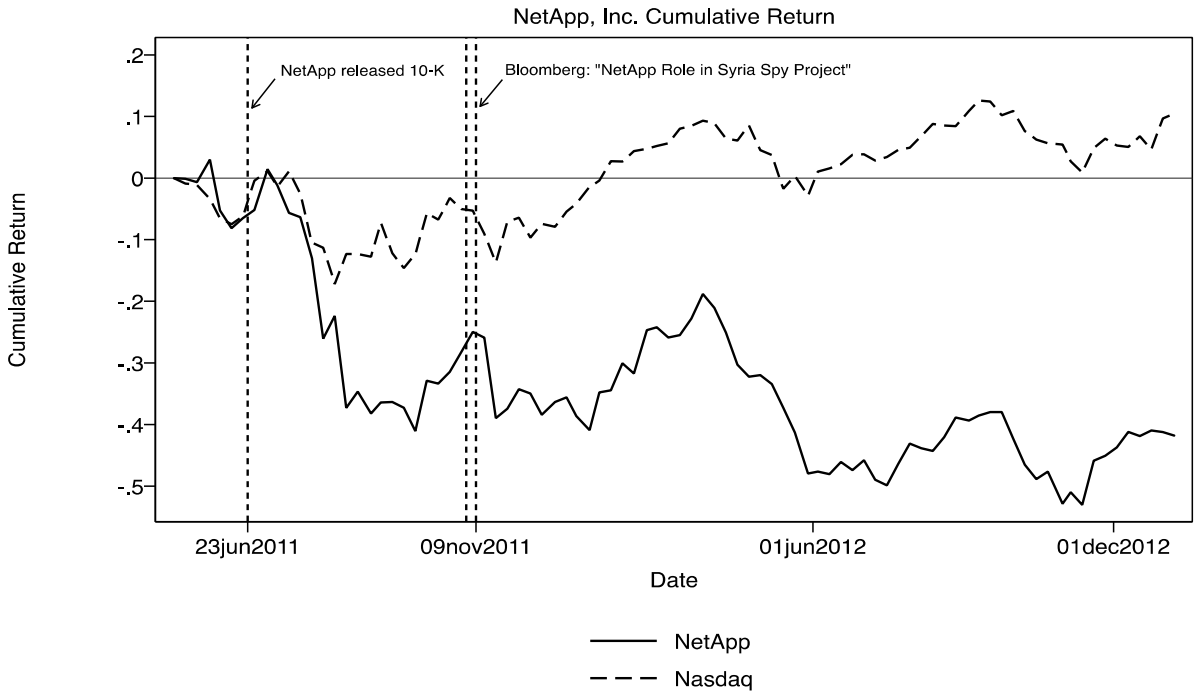

Figure 5: Long-term Event Time Returns

This figure shows the average cumulative abnormal return for each quintile portfolio sorted based on firms' similarity score, for 1 month to 12 months after portfolio formation.

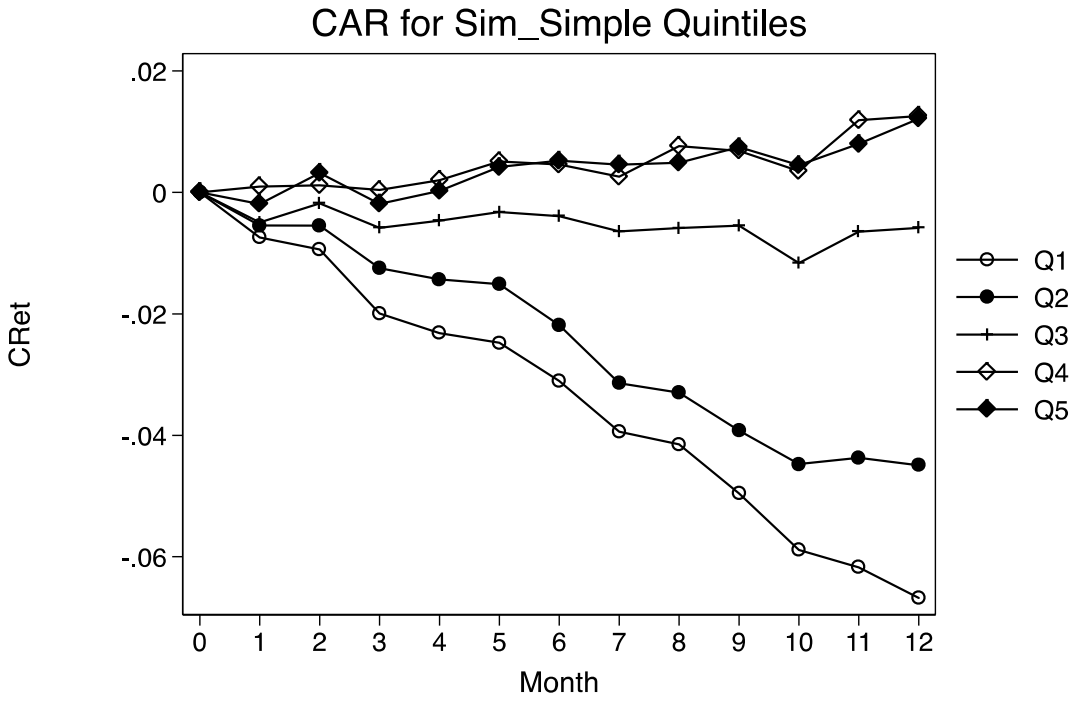

Figure 6: Item Definitions in 10Ks and 10-Qs

	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 7: Which Section is Most Important – 10K

This figure reports the average similarity score for different items of firms' 10-Ks. Item definitions can be found in Figure 5.

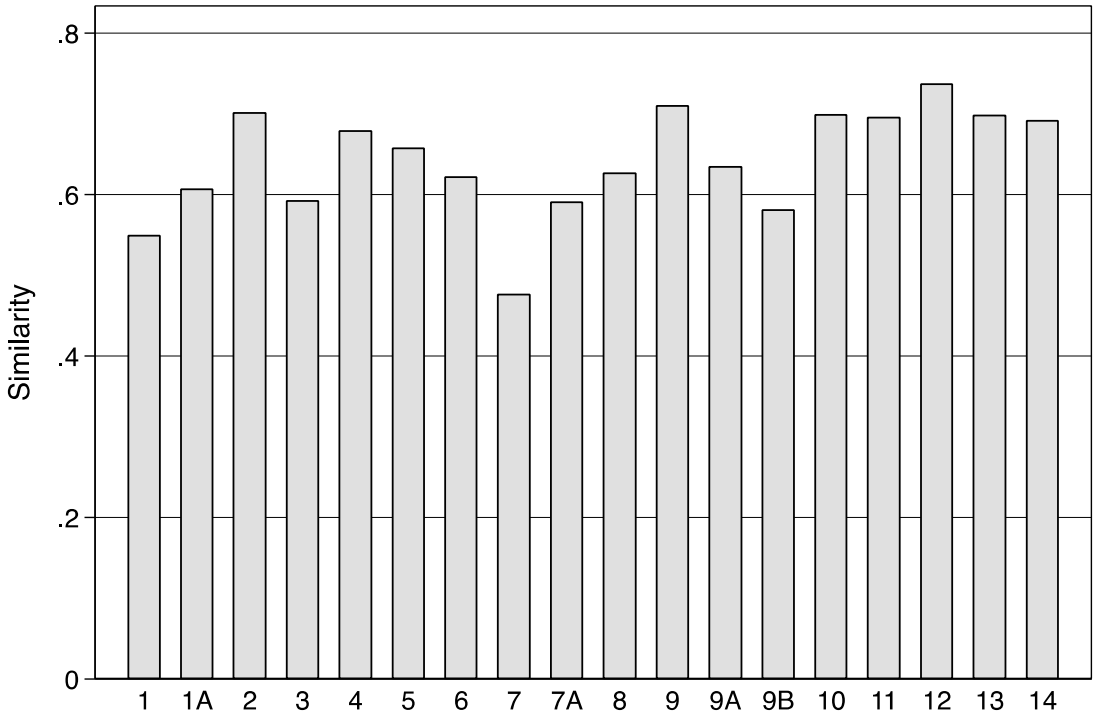

Figure 8: Which Section is Most Important – 10Q

This figure reports the average similarity score for different items of firms' 10-Qs.