Routledge
Taylor & Francis Group

Journal of Behavioral Finance

ISSN: 1542-7560 (Print) 1542-7579 (Online) Journal homepage: http://www.tandfonline.com/loi/hbhf20

The Use of EDGAR Filings by Investors

Tim Loughran & Bill McDonald

To cite this article: Tim Loughran & Bill McDonald (2017) The Use of EDGAR Filings by Investors, Journal of Behavioral Finance, 18:2, 231-248, DOI: 10.1080/15427560.2017.1308945

To link to this article: http://dx.doi.org/10.1080/15427560.2017.1308945

JOURNAL OF BEHAVIORAL FINANCE
2017, VOL. 18, NO. 2, 231–248
http://dx.doi.org/10.1080/15427560.2017.1308945

Routledge
Taylor & Francis Group

The Use of EDGAR Filings by Investors

Tim Loughran and Bill McDonald

University of Notre Dame

ABSTRACT

Using data from the Security and Exchange Commission's Electronic Data Gathering and Retrieval (EDGAR) server log, the authors examine the consumption of financial information in filings from 2003 to 2012. The EDGAR filings represent a first-source database for investors doing fundamental research on stock valuations. The magnitude of daily EDGAR requests for 10-Ks is surprisingly low and shows only a small difference between firms with and without publicly traded equity. The average publicly traded firm has their annual report requested only 28.4 total times by investors immediately after the 10-K filing. The lack of annual report requests suggests that investors generally are not doing fundamental research on stocks.

KEYWORDS

EDGAR; Fundamental research; Information decay; Web traffic; Form 10-K

Introduction

What sources of information do investors use in deciding which stocks to purchase? The best source for investors doing fundamental research on stock valuations are the filings on the Security and Exchange Commission's (SEC) Electronic Data Gathering and Retrieval (EDGAR) website. The EDGAR site has a complete set of annual reports (i.e., Form 10-Ks), 10-Qs, 8-Ks, and even IPO prospectuses (Form S-1 and Form 424). Although there are other sources for filings like the 10-K and 10-Q data, the EDGAR site is a first-source repository for the information. In this study, we are interested in the assimilation of financial data by investors as measured by activity on the EDGAR website.

In the finance literature, there is enormous debate about the efficiency of stock prices. One camp, led by Fama [1970, 1991], argues that at any point in time, stock prices should almost fully reflect all available information. Because all investors have access to identical information, no particular investor will have the ability to consistently outperform other investors. The other camp has argued that investor behavioral biases significantly affect the pricing of stock valuations. For example, Dreman [2012] proposed that by considering psychological insights, one can understand "why investors so often make incorrect decisions and why the market is subject to so many booms and busts" (p. 5). As noted by Barberis and Thaler [2003], irrationality by some investors can cause significant and long-lasting impact on security valuations even in the presence of rational traders.

Much of the prior debate on market efficiency has focused on the cost to investors of acquiring information. This cost may reflect the direct monetary expenditure of obtaining the information (i.e., the cost of purchasing satellites to count the number of cars in a store's parking lot to estimate Christmas sales) or the cost in terms of time. In a seminal article, Grossman and Stiglitz [1980] developed a simple model where prices of securities only partly reflect the information of investors in equilibrium. They argued that because acquiring information is costly, prices on individual stocks cannot always reflect all the information that is available. If all the available information is already reflected in the price of publicly traded stocks, investors who spend money to get the information would not be compensated. Grossman and Stiglitz [1980] stated that "there is a fundamental conflict between the efficiency with which markets spread information and the incentives to acquire information" (see p. 405).

Gabaix [2014] built a model based on a "sparse max" operator. Regarding stocks, investors must consider thousands of factors that play into the valuation. However, in Gabaix's model, the agent considers only the most important elements of the problem. That is, the agent, who is not fully rational, builds a simplified model of the world through behavioral effects like inattention or disproportionate salience. His model integrates the mental cost of processing the enormous quantity of data available for investors to analyze.

CONTACT Tim Loughran loughran.9@nd.edu Finance, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556-5646.

This article addresses the question of how investors acquire information to value stocks. We consider the corporate information environment from an entropy context, where we are also interested in the dynamics of information as it devolves from being novel to being fully priced. The data source considered is one relatively new to the literature, the server log associated with the SEC's EDGAR website. The SEC's server log provides the Internet Protocol (IP) address (anonymized), a timestamp, and the SEC accession number for every client request. The accession number can, in turn, be linked to an SEC file containing identifying information for each EDGAR filing. Our dataset covers the period from March 2003 to March 2012.

With the EDGAR data, it is critical to separate requests generated by robots from server requests by regular investors (e.g., nonrobots). For example, the activity of a computer program designed to download all 10-Ks is not representative of targeted information requests by individual investors or analysts. Lee, Ma, and Wang [2015] defined IP addresses with more than 50 unique firms' filings in a given day as being a "robot." This article follows Lee, Ma, and Wang's [2015] procedure of identifying robots except we define more than 50 requests in a single day from a particular IP address as being a "robot." We find that while robot requests for EDGAR filings have been steadily rising since 2008, nonrobot requests have been relatively flat.

The punchline of our article is the surprisingly low number of investors who access the annual reports of publicly traded companies at the time of its initial filing. The average publicly traded firm has their annual reports requested from the EDGAR site only 28.4 total times by investors on the day of the filing and the following day. On its filing date, the median publicly traded firm has only 9 10-K requests.

Consistent with Gabaix's [2014] bounded rationality model, most investors do not perceive the EDGAR filings as being an important element to consider. It does not appear that the vast majority of investors are doing fundamental research on valuing stocks as proxied by the server log activity on EDGAR. From a behavioral standpoint, the low number of investors carefully analyzing the annual report indicates that there are potential rewards for individuals who spend time reading over the 10-Ks. Interestingly, all-star value investor Warren Buffett has commented in several media stories that he enjoys reading annual reports.[1]

To trade on the information contained in the Form 10-K, investors must quickly analyze the text or grab some piece of information used to potentially take advantage of a possible mispricing. Yet, investors are slow to access the 10-K filings. Only 10.1% of all nonrobot 10-K requests over a 401-day window occur in the first week after the filing compared to almost half for S-1 filings (initial debt or equity prospectus). This could explain why measuring the stock price impact on the filing date of 10-Ks has been surprisingly elusive (see Easton and Zmijewski [1993] and Griffin [2003]). Investors do not appear to be trading on information contained in the 10-K at the time of its initial filing.

In addition, we consider 10-K investor requests of firms with and without publicly traded equity. The number of nonrobot requests for the public and nonpublicly traded samples is shockingly similar. One would have imagined that publicly traded companies, with significantly more shareholders, employees, and media attention, would have an enormous number of investor requests for their 10-K compared to requests for filings by nonpublic firms. Yet, an average publicly traded firm has its 10-K requested 28.4 total times on the day of and the day following the filing date compared to 14.4 total requests for firms not listed on the CRSP database.

Some articles have argued that the readability of the 10-K affects the investment decision by retail investors. Our documented low level of investor requests in small capitalization firms, however, challenges the notion that 10-K readability affects small investor trading behavior around the filing date. We find that the daily median number of nonrobot requests for firms in the smallest 3 size quintiles in the days around the 10-K filing is only 5.

It is critically important to mention the obvious: investors can obtain the 10-K and 10-Q information from other sources besides EDGAR. For example, company websites, Bloomberg, Capital IQ, and Yahoo! Finance can provide an alternative source of the EDGAR filings for investors. Thus, our analysis of the EDGAR server log cannot capture all the views/downloads that the entire universe of investors are conducting on company filings.

It is equally important, however, to underscore some problems/limitations with the other potential data sources. Generally, publicly traded companies have an investor relations website where company filings are available for download by investors doing fundamental research on valuations. First, it should be noted that the investor relations website for some firms like ExxonMobil link investors directly to the EDGAR site for any of their SEC filings or registration statements. For such cases where the investor relations department links the investors to the EDGAR site, these views/downloads will be captured in the SEC server log.

Second, it is questionable that investors primarily use the company website to retrieve SEC filings. General Electric (GE) is a prominent publicly traded company with significant daily media coverage, approximately

40% of its shareholders being retail investors, and a large pool of retired employees with defined benefit plans. Thus, GE should have a large number of investors/employees poring over its annual reports looking for insights. Yet, in a June 2, 2015, *Wall Street Journal* article, the CFO of GE noted that its 2013 annual report was downloaded only 800 times from GE's website during the entire year.[2] The CFO also noted that given few investors call GE's investor relations department with questions concerning the annual report and given the low number of views of its annual report, "both were signs that many investors didn't read the 10-K in depth."

Third, other sources of the company information often condense income statement and balance sheet information into prespecified bins. As an example, technology firm Amazon over the last 3 years, according to its 2015 annual report, increasingly devoted resources to the research and development of its web services division. A recent Bloomberg article describes Amazon as the top U.S. spender on research and development.[3] Yet, Yahoo! Finance reports in the same period that Amazon had no research & development expenditures. The problem lies in the fact that Amazon refers to its research development as the line item called "Technology and Content" and thus Yahoo's data parser simply fails to itemize the line.

One of the top selling valuation textbooks (*Valuation: Measuring and Managing the Value of Companies* by McKinsey & Company) advocates placing the value of operating values on the firm's balance sheet to properly compare companies with different leasing policies. Some U.S. firms that are very active with operating leases, like retailers or airlines, could have more assets off their balance sheets than on them. To value the off-balance sheet operating leases, which are becoming increasingly common in publicly traded companies, investors need to obtain the firm's forecast for next year's rental expense from the annual report's notes or footnotes. This valuable information is available from annual reports on EDGAR, not from a Bloomberg terminal or the Yahoo! Finance web page.

Last, 10-Ks and 10-Qs provide investors verbal clues about the company's present situation that data consolidators fail to capture. To evaluate sentiment tone, investors need the actual words contained in the financial filings. A number of academic articles have started associating manager's word selections in the 10-K with subsequent events. For example, Loughran and McDonald [2011] linked the tone of the annual report with 10-K filing returns, trading volume, and postfiling stock market volatility while Feldman, Govindaraj, Livnat, and Segal [2010] related changes in tone in the management discussion and analysis (MD&A) section to contemporaneous returns around SEC filing date.

Literature review

The important issue of how to gauge investor or analyst interest in particular stocks has evolved with the increased availability of novel datasets. For example, some research has used Google search volume to gauge investor interest in stocks (see Da, Engelberg, and Gao [2011], Drake, Roulstone, and Thornock [2012], and Chi and Shanthikumar [2014]) while others have utilized measures of analyst site visits to company locations to document how analysts acquire information for their earnings forecasts (Cheng, Du, Wang, and Wang [2014]). Cohen, Frazzini, and Malloy [2010] used the educational background linkage between sell-side Wall Street analysts and corporate managers to identify a source of the analyst's superior information. For discovering how investors acquire information on companies, the SEC's EDGAR server log of all filing requests appears particularly promising.

Using the EDGAR server log, Lee, Ma, and Wang [2015] identified economically related peer firms by examining the sequence of chronologically adjacent EDGAR searches by nonrobot investors. For example, if investors who accessed Priceline Group's EDGAR filings tended to next look at the filings of Expedia and Orbitz Worldwide, then Expedia and Orbitz would be classified as the search-based peers of Priceline. Lee, Ma, and Wang [2015] find their search-based peers to be better at explaining cross-sectional variation in firm characteristics like subsequent stock returns and valuation multiplies than the widely used 6-digit Global Industry Classification Standard. There is clearly a systematic logic in the way investors interact with the EDGAR filings.

Another article related to our work is Drake et al. [2015], who used web traffic on the EDGAR system to expand the literature's understanding of information acquisition by investors. For a relatively limited time period, 2008–2011, Drake et al. [2015] focused primarily on what variables and events determine web traffic on EDGAR. The three authors found that nonrobot EDGAR requests by investors are positively related to important corporate events, like restatements and earnings announcements, firm size, and weak abnormal stock performance.

When analyzing data from the EDGAR server log, the manner in which robot requests are separated from nonrobot requests is critical. We follow the procedure of Lee, Ma, and Wang [2015] except that we classify IP addresses with more than 50 filing requests (instead of

50 unique firms' filings) in a given day as being robot requests. Differently, Drake et al. [2015] defined IP addresses to be robots if the address has more than 5 filing requests in a given minute or more than 1,000 firm filing requests during a single day.

The difference in robot definitions leads to dramatically different nonrobot request counts between the articles. For example, we find a total of 137,013 nonrobot requests for Form 4 filings during March 2003 to March 2012 while Drake et al. [2015] (Table 3) report a total of 16,941,014 requests by nonrobot investors for this somewhat obscure EDGAR filing. For nonrobots, our ratio of 10-K requests to Form 4 requests is 251. That is, for every Form 4 request, nonrobot investors make 251 10-K requests. Given the immense investor focus on annual reports, this ratio seems reasonable. In contrast, Drake et al. have a ratio for nonrobots of only 2.3 10-K requests per Form 4 request. Clearly, their very high robot screen of 1,000 firm requests in a single day is allowing some web-crawlers to slip into their sample.

Data

In this section of the article, we describe in detail the server log data provided by the SEC. The original dataset produced by the SEC consists of thousands of files with more than 4 billion documented server requests and requiring more than one-half terabyte of storage. As an artifact of the empirical tests in this article, we develop a substantially condensed version of the log data with the server requests linked to the EDGAR Master File form type and filing date. Our condensed data set is publicly available at https://sraf.nd.edu/data/.

The EDGAR server log

Web servers maintain a log of all page requests. For each request, a typical server will log the client IP address, timestamp of the request, and page requested, in addition to a few other qualifying items.[4] The log files are only available from 2003 forward.

Following our FOIA request, we received 3,319 daily files from January 1, 2003, to March 31, 2012, with log files containing web requests for SEC filings. In the data we received, and consistent with the documentation provided by the SEC, the majority of files prior to February 13, 2003, contained either zero or less than a handful of data requests. The flow of data appears to ramp up and stabilize somewhere in mid-February of 2003.[5] Thus we choose to base all analysis in our research on the 3,090 daily EDGAR log files from March 1, 2003, to March 31, 2012. This excludes the September 24, 2005, through May 10, 2006, period, which are dates when the log files were labeled by the SEC as "lost or damaged."

We first remove from the daily log all file requests not relevant for our sample. This filter is based on 3 codes reported with each record. First, any request flagged as a web crawler is excluded from the sample.[6] Second, any index page request is excluded. Finally, all requests with server codes of 300 or greater are excluded.[7] These 3 filters reduce the original data from more than 4 billion requests to about 1.5 billion requests.

The next step in creating the sample allows us to reduce the record count by orders of magnitude. By aggregating the requests for each day to counts associated only with those filings having at least 1 nonrobot (NR) request, we are able to summarize the data into about 113 million observations from the original sample of more than 4 billion.

Increasingly over the past decade, some consumers of EDGAR data use automated programs—labeled robots—to download targeted sets of files from the SEC website. Using a time series of EDGAR server logs from 2008 to 2011, Lee, Ma, and Wang [2015] provide a detailed derivation for a heuristic rule that filters out robots.[8] Presumably, these requests are not representative of attempts to measure information flow for a specific firm. Lee, Ma, and Wang [2015] showed that 95% of all daily requests associated with a specific IP are for no more than 50 unique firms' filings. They use this number (50) as the threshold for labeling a request as one generated by a robot.

We further collapse the data by creating aggregate counts for each form on each day. We tabulate counts for each day and for each accession number which has at least 1 valid nonrobot request. We partition all requests for that filing initially into 2 variables, *Robot_count* and *NR_total* (nonrobot total), where *NR_total* represents counts with fifty or fewer requests on a particular day for the IP address associated with the request.

Lee, Ma, and Wang [2015] argued that server requests associated with users actually viewing the data are most likely associated with HTM file types—that is, files directly viewable in a web browser. Thus, we also maintain counts for nonrobot HTM files (labeled *NR_HTM*) and nonrobot TXT files (labeled *NR_TXT*).[9] After an initial examination of all the variables, our analysis will focus on only the nonrobot HTM files (*NR_HTM*).

The SEC assigns each unique filing an accession number. By linking the accession number back to the Master Index files maintained by the SEC, we are able to append the server data with the form type (e.g., 10-K, 8-K) and filing date of the form. We were able to successfully match 99.98% of the forms.

We now have a dataset of 113,073,168 records by day and by accession number, each with: (1) the server

record date; (2) a dummy variable set equal to 1 on trading days, else zero; (3) the Central Index Key assigned to the filer; (4) the total count for all nonrobot server requests (*NR_Total*); (5) the count for all nonrobot requests for HTM files (*NR_HTM*); (6) the count for all nonrobot requests for TXT files (*NR_TXT*); (7) the count for robot requests (*Robot_count*); (8) the form type of the filing; and (9) the filing-date for the form. Note that as a subset of the nonrobot requests, we maintain separate counts for HTM and TXT files.

Forms included in the analysis

Our analysis will consider the requesting patterns of EDGAR filings by form type during 2003–2012. More than 90% of all filings are accounted for by the top 50 of the 609 unique form types. Focusing on the most frequently filed forms provides a first cut of the data. We initially examine the form clusters reported in Table 1.[10] As can be seen in the groupings of Table 1, we separately consider appended filings, quarterly versus annual filings, and small business filings. Not surprisingly, quarterly reports appear almost 3 times more frequently than annual reports. Although Form 10KSB, the annual 10-K for small businesses, accounts for only 0.20% of all filings, it clearly provides an interesting comparison with traditional 10-Ks over the period in which it was an alternative to the 10-K.

Similarly, appended documents are less frequent, but provide a useful contrast to initial filings when considering information decay. After an initial assessment of the filings listed in Table 1, we will focus our efforts on 10-K filings (because they are a significant source of financial information of a firm). A brief description of each filing type we consider is provided in Appendix 3.[11]

What stands out in Table 1 is the large number of Form 4 filings (changes in beneficial ownership by managers and shareholders with more than 10% of equity). There are over 4 million different Form 4 filings during the 2003–2012 time period. As noted earlier, robots are the most frequent consumers of Form 4 information. The second most frequent type of filing is Form 8-K. Although 10-K filings dominate much of the analysis in the accounting and finance literature, the form and its amended filings account for only about 1.07% of all filings on EDGAR.

Table 1. EDGAR filing frequencies by form type.

	Form Type	Count	% of Total
1	10-K	88,461	0.88
2	10-K/A	19,446	0.19
3	10KSB	20,103	0.20
4	10KSB/A	7,124	0.07
5	10-Q	234,762	2.34
6	10-Q/A	20,339	0.20
7	10QSB	55,456	0.55
8	10QSB/A	9,145	0.09
9	8-K	892,420	8.90
10	8-K/A	38,211	0.38
11	S-1	10,604	0.11
12	S-1/A	25,486	0.25
13	424A | 424B1-424B8	280,755	2.80
14	3	387,194	3.86
15	3/A	26,019	0.26
16	4	4,269,940	42.57
17	4/A	155,404	1.55
18	5	129,570	1.29
19	5/A	5,819	0.06
20	SC 13D	45,684	0.46
21	SC 13D/A	97,546	0.97
22	SC 13G	167,712	1.67
23	SC 13G/A	288,933	2.88
24	13F-HR	110,364	1.10
25	DEF 14A	69,007	0.69
26	497	180,339	1.80
27	6-K	229,464	2.29
28	6-K/A	3,086	0.03
29	Other	3,358,829	33.49
	Total	11,227,222	100

Note. Total filings and percentages are based on all filings listed in the SEC EDGAR Master File Index from 2003 to 2012.

Results

Calendar results for the EDGAR server log

In this section, we provide an initial assay of the server requests by calendar groupings using the full sample of valid requests. Drake et al. [2015] provided similar calendar descriptions of the server traffic from 2008 to 2011. Figure 1 displays the count for all valid requests (robots and nonrobots) for each unique year/month in the 11-year sample. The blank space from September 2005 to May 2006 reflects the period when the SEC files were corrupted. The number of automated requests has clearly increased dramatically, while the number of nonrobot requests has remained relatively stable. This highlights a characteristic of the dataset which must be emphasized. Importantly, many vendors (e.g., EDGAR Online, Capital IQ, or subscriptions through Keane Federal Systems) repackage the original filings for distribution.

This plays a role in lowering the nonrobot requests because other sources provide the same data for investors. As a result, conclusions drawn from the EDGAR server log must be considered as simply one channel among many for distribution channels of financial disclosures. Although presumably the cross-sectional deviations in traffic are informative, Figure 1 suggests that any time series comparisons could be misleading as alternative delivery mechanisms have proliferated. Also apparent in the Figure 1 time series is a notable drop in nonrobot requests in the last 2 months of the sample. For reasons we have been unable to determine, the count of nonrobot requests drops substantially in the last 2

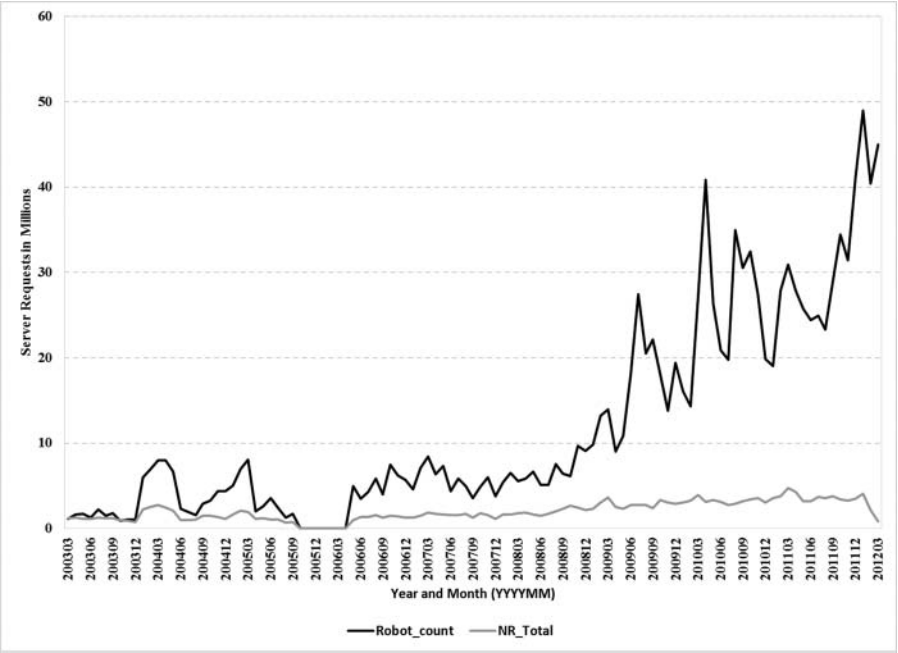

Figure 1. The robot and nonrobot counts of EDGAR server requests, in millions, for each unique year and month during March 2003 to March 2012. The blank space from September 2005 to May 2006 reflects the period when the server files were corrupted.

months of the sample, even though total requests increases.

We present the server requests by hour in Figure 2. All reported times in the article are U.S. Eastern Standard Time. As might be expected, most requests occur during trading days between 10 a.m. and 6 p.m. Our use of the Lee, Ma, and Wang [2015] threshold of 50 to categorize requests by robots appears to be quite effective. There are relatively few nonrobot server requests when most U.S. investors are asleep (1 a.m. to 7 a.m.). The quantity of nonrobot requests sharply rises once people are at work (10 a.m.) and continues strong for a few hours after the stock market's 4 p.m. close. Because automated requests frequently run over a long period with no need for human

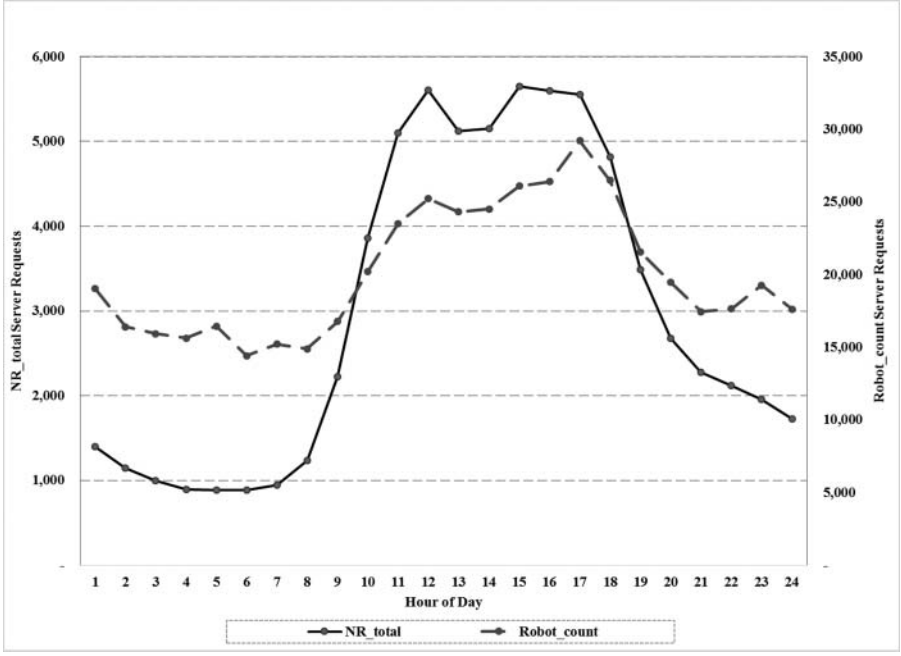

Figure 2. The nonrobot counts (*NR_total*) and robot counts for server requests by hour of day. Note that *Robot_count* is associated with the right-hand-side axis scale. All reported times in the article are U.S. Eastern Standard Time.

intervention, the robot count is less differentiated during business hours. Interestingly, robot requests peak at 5 p. m., suggesting that some targeted robot requests are harvested immediately following the close of the market.

The distribution of daily filing requests

We now focus on the compressed sample where each daily count represents the number of file requests for a given EDGAR filing on that day for all forms with at least 1 valid nonrobot server request. Table 2 provides selected percentiles for the distribution of the various counts. Lee, Ma, and Wang [2015] emphasize the "power law" nature of the server request data. That is, the number of requests for a specific firm and form type on a given day will be dominated by very low counts, with occasional bursts of extraordinary activity. This produces a distribution of counts that is much like market capitalization, where the vast majority of firms are of fractional size compared to a few extremely large firms. It is also similar to document word counts where a large number of less common words have very low counts while a small number of stop words (words like *the, and*, or *for*) have extraordinarily large counts.

The power law nature of these counts is apparent in Table 2, where the median daily count total is 1 or zero for each count classification, the 99th percentile is less than 48 for all counts, and the maximum is over 100,000 for all counts except *NR_TXT*. As noted by Lee, Ma, and Wang [2015], the small extreme value for *NR_TXT* is attributable to the fact that most text file requests are likely to be associated with robot requests. Given that *NR_HTM* and *NR_TXT* are subsets of *NR_total*, the increasing mean value for the counts across this spectrum is not surprising. Recall that all of the count statistics are conditional on a form having at least 1 nonrobot request (of any file type) on a given day. Subsequently, we will mostly focus our analysis on the count represented by *NR_HTM*, which, consistent with the observations of Lee, Ma, and Wang [2015], should best represent the consumption of disclosure information by individuals (vs. programmed robots).

To provide a meaningful measure of *N_days*—the number of days between a form's filing date and the day of the request—we must limit observations included in the sample when calculating the summary statistics for the following reasons. First, *N_days* is available only for sever log observations we could match to the EDGAR master index (more than 99.9% of the original sample). Second, the day count is truncated by the end of the sample and some filings could occur as early as 1994, thus the measure is substantively impacted by the original filing date and the termination of the sample.

Because we cannot compare decay rates for filing requests across observations with differing potential time spans, we will define the observation interval as approximately 1 calendar year (365 days). To allow all forms to have equal potential time spans, we terminate the observations 1 year prior to the end of the sample and include only those with a form filing date equal to or greater than the beginning of the server log data (March 1, 2003). Finally, to account for the September 24, 2005, to May 10, 2006, period when the log files are missing, we exclude observations with filing dates occurring after September 24, 2004, and before May 11, 2006. This modification insures that each filing has an opportunity to have a valid request entry for the 365-day postfiling period. Because each observation in the dataset we have created provides summary counts, the *N_days* statistics reported in Table 2 are weighted by the frequency count for *NR_HTM*.

In the last row of Table 2, the value of 1 for the 10th percentile of *N_days* indicates that about one-tenth of a form's requests in the year following its initial filing occur on the day of or day after the filing date (day 0). Half of the filings within a year take place well within the first two months following the initial filing (median = 56 calendar days). In our subsequent analysis, we will look

Table 2. Distributional statistics for EDGAR server requests for all form types.

Variable	Mean	Minimum	1%	5%	10%	25%	Median	75%	90%	95%	99%	Maximum
NR_HTM	1.308	0	0	0	0	0	1	1	3	4	11	115,558
NR_TXT	0.419	0	0	0	0	0	0	1	1	2	3	3,069
NR_total	1.948	1	1	1	1	1	1	2	3	5	13	117,781
Robot_count	2.668	0	0	0	0	0	0	1	3	7	47	121,340
N_days	99.223	0	0	0	1	15	56	166	282	327	358	365

Note. This table reports selected percentiles for the distribution of each daily count of server requests for a given filing by count type. Each variable is measured as the count for a given EDGAR filing on a given day in the March 2003 to December 2012 sample, excluding the period of September 24, 2005, to May 10, 2006, when the log files were not available. *NR_HTM* and *NR_TXT* are nonrobot server request counts for HTM files and TXT files, respectively. *NR_total* is the count of each filing for all nonrobot requests having at least 1 nonrobot count. *Robot_count* is the count of all robot requests (more than 50 document requests for a given IP on a given day) for a filing with at least 1 nonrobot count. *N_days* is the weighted average number of days between the filing date for a given form and the day of the counts, weighted by *NR_HTM*. *N_days* limits the sample to cases where $N\_days \leq 365$ for a given observation. In addition, as detailed in the text, each filing included must have an opportunity to be downloaded for the full 365 days within the constraints of the sample. The sample size for *NR_HTM, NR_TXT, NR_total*, and *Robot_count* is 113,073,168. The sample size for *N_days* is 94,510,378.

at these one-year decay rates in the context of specific form types and other variables.

Measuring information decay using the EDGAR server log data

One focus of this research is to examine how the consumption of financial disclosures decays over time. In this section we will consider how we measure decay. We first consider just one example to highlight the choices made in selecting a measure. Figure 3 charts *NR_HTM* for the 10-K filing of GE on February 2, 2007. Note that the plot begins 22 days after the filing date to exclude higher values (maximum of 393) occurring during the immediate filing period and thus focus the scale of the chart on the longer period of lower frequency requests. There is little reason to believe that information for financial disclosures will decay in a manner that would suggest using some sort of smooth hazard function to represent the process. Figure 3 supports this contention.

In Figure 3, we have also overlaid the filing dates of GE's 10-Q quarterly reports (diamond) and 10-K annual reports (square) in the time series. From this single case, and contrary to what might be hypothesized, it is not clear that 10-K consumption spikes with subsequent releases, given that the information could have value as a comparative basis for such reports. We will test this proposition more formally later in a regression context. Notice that once GE's 10-K filing on February 20, 2008, is available, there is a sharp dropoff in requests for its February 2007 10-K. In the next section we will consider the decay rates for various form groups. Because we expect the decay to occur rapidly but then spike occasionally with news-related demand, we will examine the proportion of requests falling within selected calendar periods following the initial filing.

Information decay rates by EDGAR form type

For clarity, we do not report in Table 3 all of the form group variants itemized in Table 1. Also, in reporting the sequence of time categories, we exclude the third quarter. Similar to our previous description for *N_days*, we tabulate the counts for the nonrobot HTM files (*NR_HTM)* using only observations whose form filing date allows for 400 subsequent calendar days within the server log sample. We use days [0,400] to capture a full calendar year plus an additional few weeks to allow for usage that might be associated with backward looking comparisons of annual filings (e.g., 10-Ks).

In columns 1–6 of Table 3, we tabulate the percent of *NR_HTM* requests into approximate calendar groupings within the [0,400] day postfiling time period using the following time windows: (1) the filing date (days [0,1]), (2) the first week (days [0,7]), (3) the first month (days [0,30]), (4) the first quarter (days [0,90]), (5) the second quarter (days [0,180]), and (6) the fourth quarter (days

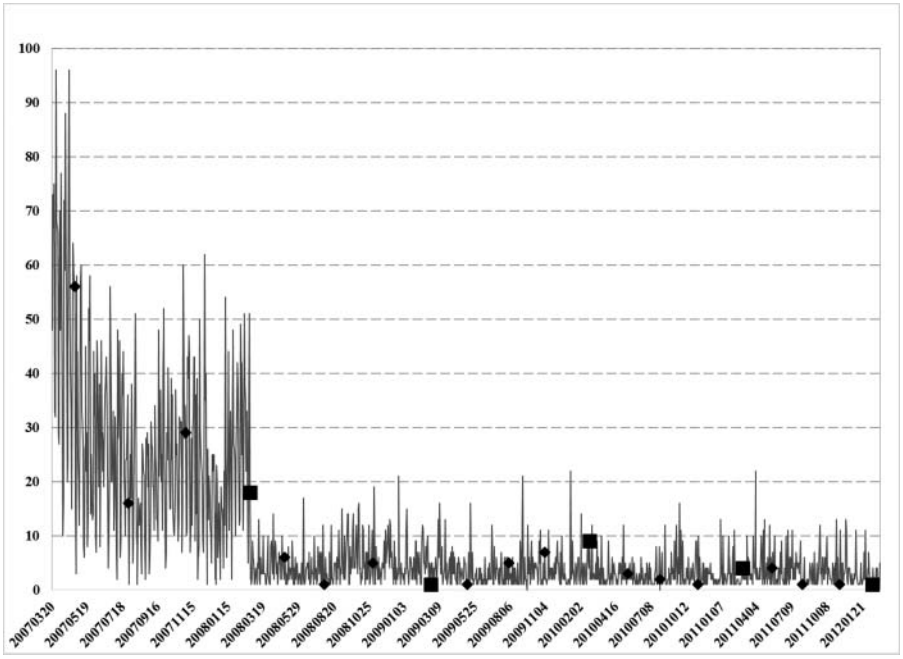

Figure 3. The nonrobot count for HTM type server requests (*NR_HTM*) for General Electric's 10-K filing of February 2, 2007. The first 21 days (including the filing date) are excluded from the chart due to their larger numbers and an attempt to focus the scale of the chart on a longer postfiling period. The dates of subsequent 10-Q filings are denoted by a diamond and subsequent 10-K filings by a square.

Table 3. Distribution of EDGAR file requests within 400 days of filing and ratio of *NR_HTM* file requests to form filings.

	Percentage of Days [0,400] File Requests							Full Sample		
Form Type	(1) Day [0,1]	(2) First Week	(3) First Month	(4) First Quarter	(5) Second Quarter	(6) Fourth Quarter	(7) Total *NR_HTM* Count ≤ 400 days	(8) Total *NR_HTM* File Requests	(9) Total Filed 2003–2012	(10) Ratio of *NR_HTM* File Requests to Filings
10-K	4.7	10.1	21.6	41.1	60.0	97.8	24,290,309	34,450,841	88,461	389.45
10-K/A	9.0	17.7	31.8	52.7	72.7	98.4	1,384,279	1,992,423	19,446	102.46
10KSB	10.4	17.8	31.1	48.3	65.7	96.9	405,518	719,538	20,103	35.79
10-Q	8.1	16.3	33.3	71.8	88.0	98.7	20,598,503	25,063,871	234,762	106.76
8-K	19.0	34.0	54.5	75.4	87.0	98.6	23,657,516	30,210,880	892,420	33.85
S-1	29.5	47.8	62.9	78.0	89.2	98.9	2,883,443	3,579,307	10,604	337.54
424	19.7	33.4	49.7	68.5	82.7	98.1	3,876,358	5,778,612	280,755	20.58
4	25.5	41.3	58.3	74.1	84.7	98.2	68,278	137,013	4,269,940	0.03
SC 13D	25.9	42.6	59.0	76.1	87.1	98.5	513,312	668,144	45,684	14.63
SC 13G	18.8	34.4	54.5	73.8	85.4	98.5	504,293	629,384	167,712	3.75
13F-HR	16.7	33.3	50.0	92.9	100.0	100.0	42	152	110,364	0.00
DEF 14A	6.8	15.9	30.8	47.3	64.5	98.0	5,452,271	7,807,698	69,007	113.14
497	17.9	30.4	46.9	67.8	83.5	98.9	546,872	655,903	180,339	3.64
6-K	12.9	27.3	50.3	74.6	87.4	98.7	3,360,208	4,028,823	229,464	17.56
Other	15.7	29.8	47.1	65.1	79.9	98.3	17,484,283	23,672,056	3,358,829	7.05

Note. Columns 1–6 report the percentage of EDGAR file requests—for a given filing and by form type—occurring from the filing day through the subsequent 400 days, by calendar period. Columns 8–10 compare the number of form requests to the number of filings for the entire period. Form groups are described in Table 1 and form types are defined in Appendix 3.

[0,365]). Note that the 400th day percent will, by definition, equal 100. Column 7 reports the total count of *NR_HTM* file requests occurring in the entire [0,400] day interval for each form type.

In columns 8–10 of Table 3, we also report a series of numbers that provide a sense of the total form requests relative to the number of forms filed. Column 8 reports the unconditional total of *NR_HTM* file requests for a given form. Column 9 uses the EDGAR Master Index files to report the total number of filings for each form over the interval 2003–2012. Column 10 then provides the ratio of *NR_HTM* file requests to the total number of filings for a given form. This ratio provides a useful measure of the relative consumption of the various form types.

One general observation is immediately clear across the various form types. For all but the 10-K (annual reports), 10KSB (annual report filed by a small business), and DEF 14A (filed when shareholder vote is required), the median request over the 401 day window takes place before the first quarter following the filing date. This is not surprising, but definitively documents that investor and analyst interest in required filings decays rapidly after their initial disclosure. Most of the various forms have 80% or more of their filing requests from the first 400 days occurring before the end of the second quarter.

The filing groups reporting the highest initial interest percentage are consistently either the S-1 filings or forms reporting beneficial ownership (Forms 4, SC 13D), with their days [0,1] requests all exceeding 25%, and easily more than half of their total days [0,400] requests occurring within the first month after filing. In our Table 3, we show that the rate of information decay is clearly different for 10-Ks compared to other form types. Interestingly, Form 10-Ks have the lowest percent of filing date requests (4.7%), and have only 60% of the total requests by the second quarter. Certainly beyond the initial payload of new information in a 10-K filing, the document serves as an intertemporal and industry comparable after its initial filing.

We can only comment on the number of requests by relating them to the number of filings. This comparison highlights the important but differential nature of 10-K versus S-1 (initial debt or equity prospectus) filings. Both the 10-K and S-1 forms are dominant in terms of the number of requests per filing, with both having more than 337 requests for a given filing. At the same time, the S-1 has almost half (47.8%) its days [0,400] requests occurring within the first week, compared with only about 10% for the 10-K. Although the information for these 2 forms decays at very different rates, they are clearly both considered important by consumers of financial information.

Also notice in the Table 3 results that 10-Ks attract much more interest from investors than 10-Q filings. There are 389.45 nonrobot requests per 10-K filing compared to only 106.76 nonrobot requests per 10-Q. This is consistent with filing returns evidence of Griffin [2003]. He finds a stronger response from investors surrounding the filings of 10-Ks than for 10-Qs.

Given the cost of producing financial disclosures, the other end of the usage spectrum should also be of interest. The average number of requests per filing for those forms not explicitly separated out in Table 3 (category "Other"), and accounting for approximately 70% of all filings, is about 7. That is, each filing, on average, is

viewed only 7 times during this sample interval. Four of the fourteen form-specific categories in Table 3 have fewer than 10 file requests per filing over the entire 401-day interval.

To the extent that some filings serve primarily as a document of record, the actual usage may not be critical. However, given the costs of producing financial disclosures and the potential for information overload, the SEC should carefully consider the requirements for those documents whose usage is minimal. In cases where the value of the information seems unquestionable, the SEC should consider marketing efforts to inform investors of the accessibility of these data. The low requests of certain filing types might be due to investors being unaware of their availability.

To provide the reader some sense of the most frequently accessed filings, Table 4 lists the top 25 server requests for nonrobot HTM files. As expected, the most commonly requested filings by investors tend to be of widely followed companies. Interestingly, only rarely do firms have one of their filings requested more than 6,000 times by investors in a given day.

Six of the top 15 download counts, including the top 2 of 115,558 and 111,490 requests, are the day of Facebook's S-1 (initial public offering prospectus) filing and the subsequent 5 days. The third and fourth largest form requests are associated with American International Group as the firm was going through bankruptcy, with 29,191 10-K filings requested on April 24, 2009, and 30,789 DEF 14C filings on December 15, 2010, a few days after the form describing their recapitalization was filed. The sixth and seventh most frequent requests are associated with the day of the S-1 filings for Groupon (28,442) and Zynga (27,674).

Table 4. Top 25 server requests for nonrobot HTM files (*NR_HTM*).

	Company Name	Form Type	Filing Date	Server Date	*NR_HTM*
1	Facebook	S-1	20120201	20120202	115,558
2	Facebook	S-1	20120201	20120201	111,490
3	AIG	Def 14DC	20101210	20101215	30,789
4	AIG	10-K	20090302	20090424	29,191
5	Facebook	S-1	20120201	20120203	28,924
6	Groupon	S-1	20110602	20110602	28,442
7	Zynga	S-1	20110701	20110701	27,674
8	Google	S-1	20040429	20040429	18,452
9	AOL	10-Q	20101103	20110122	17,273
10	EMC	8-K	20110317	20110318	15,599
11	Google	S-1	20040429	20040430	15,486
12	Groupon	S-1	20110602	20110603	14,635
13	Facebook	S-1	20120201	20120206	12,422
14	Facebook	S-1	20120201	20120204	11,931
15	Facebook	S-1	20120201	20120205	10,451
16	Google	10-K	20070301	20070305	9,320
17	Facebook	S-1	20120201	20120207	8,330
18	Kosmos Energy	S-1/A	20110303	20110428	7,933
19	LinkedIn	S-1	20110127	20110127	7,729
20	SCO Group	8-K	20091019	20091019	6,885
21	Zynga	S-1	20110701	20110702	6,883
22	LinkedIn	S-1	20110127	20110128	6,706
23	Clearwire Corp	DEF 14C	20111219	20111229	6,580
24	Harley Davidson	DEFA14A	20090415	20090424	6,436
25	Yelp	S-1	20111117	20111117	6,416

The top ten 10-K filing requests over the 401 day window are for Apple (2), American International Group, Google (3), Microsoft (3), and Motors Liquidation Company (formerly General Motors). Impressively, Apple's October 26, 2011, 10-K filing generated approximately 390 requests/day in the first quarter following the filing. Given the power-law nature of requests, the firms with the highest number of requests are clearly extraordinary outliers.

Requests categorized by public and private firms around the time of the 10-K filing

In this section, we will focus on annual 10-K reports and separate the data into firms with and without stock market data reported in the Center for Research in Security Prices (CRSP) data files. We will assume that the dichotomy of firms with and without CRSP data roughly corresponds to separating the sample into firms with actively traded public equity and those large enough to be required to file (500 shareholders and $10 million in assets) but not having actively traded public equity.[12] This division of the sample is interesting because, to the extent investors are using 10-Ks requested from the SEC website to assess a firm's stock, we would expect the request activity of the sample with actively traded public equity (i.e., the CRSP sample) to exhibit substantively higher request activity. In addition, most prior research has focused primarily on the equity investors of publicly traded companies (see Kothari [2001]).

Table 5 reports the summary statistics for nonrobot requests around the 10-K filing date categorized by publicly traded (Panel A) and private firms (Panel B). The table clearly contradicts the notion that individual investors and analysts are actively downloading 10-Ks for valuing stocks. On the day of the 10-K filing, the average publicly traded firm has only 16.1 requests. On Day +1, the average drops slightly to 12.3. Thus, on the filing day and the following day, there are, on average, only 28.4 total 10-K requests. Not surprisingly given the skewness of requests, the median number of requests is even smaller for the days immediately around the 10-K filing date (i.e., 9, 6, and 3). The typical publicly traded firm has only a handful for requests for the annual report immediately after its filing.

Panel B of Table 5 reports that an average firm without publicly traded equity has its 10-K requested about 14.4 times on the day of and day following the filing date (half of the requests as for firms with publicly traded equity). This is a difference not conditioned on firm size,

Table 5. Number of *NR_HTM* count (nonrobot HTM file requests) for firms with and without public equity around the 10-K filing date.
Panel A: Number of requests around the 10-K filing date for publicly traded companies ($n = 29{,}012$)

	Day 0	Day +1	Day +2	Day +3	Day +4
Average	16.1	12.3	6.5	6.9	6.1
Minimum	0	0	0	0	0
Median	9	6	3	3	2
Maximum	1,683	2,065	1,051	839	9,320

Panel B: Number of requests around the 10-K filing date for nonpublicly traded companies ($n = 21{,}604$)

	Day 0	Day +1	Day +2	Day +3	Day +4
Average	8.4	6.0	2.8	2.7	2.2
Minimum	0	0	0	0	0
Median	4	2	1	1	0
Maximum	3,866	930	255	223	318

Note. This table reports the number of nonrobot requests categorized by firms with and without publicly traded equity in the days around the 10-K filing date. Day 0 represents the 10-K filing date. The sample period is March 2003 to March 2012.

which is presumably much smaller for the non-CRSP sample. The median private firm has a trivial number of requests in days around the 10-K filing (i.e., only 4 on Day 0 and 2 on Day +1).

Small investor trading behavior

Miller [2010] examined the trading behavior of large and small investors in the days around the 10-K filing date. He defined trades to be made by small investors if (1) the trade dollar amount is less than or equal to $5,000 or (2) for firms with a stock price more than $50 the number of shares purchased is less than or equal to 100. Miller found that the readability of the annual report, defined using the Fog Index and number of words in the 10-K, affects small investor trading behavior. More complex annual reports (i.e., high Fog Index values or lengthy Form 10-Ks) are found to significantly reduce small investor trading behavior. Using actual investor stock portfolio holdings during 1994–1996, Lawrence [2013] found that investors hold shares of companies with better written 10-Ks. However, Lawrence reported that financially literate individuals (e.g., accountants, business managers, lawyers) who might be expected to actually read the annual reports are found to be unaffected by the 10-K's readability or document length.

Yet, how frequently are small investors actually accessing 10-Ks immediately after the filing? Are the counts of retail investor requests large enough to measurably affect trading volume in a short window around the filing date? Although we realize that some investors might go to the company's investor relations site to obtain the Form 10-K filing, EDGAR counts should provide a reasonable sense of investor interest in the filings. Retail investors would not be using robots to parse the annual reports.

Retail investors disproportionally hold shares of smaller market capitalization firms. For example, Ivković, Sialm, and Weisbenner [2008], using actual household stock holdings, found that small investors overweight stocks not included in the large capitalization S&P 500 index. To proxy for stocks traded by small investors, Table 6 reports the mean and median total number of times nonrobot investors (i.e., *NR_HTM*) examine 10-Ks on the days around the filing date (Day 0 through Day +4) categorized by annually determined market value quintiles. Not surprisingly, large capitalization firms have significantly more nonrobot requests than small firms. The mean number of total investor requests in the 5 days around the 10-K filing date is 96.2 for firms in the largest size quintile compared to only 28.9 total requests for publicly traded firms in the smallest size quintile.

The table reports that firms in the bottom 3 size quintiles are infrequently examined in the days surrounding the 10-K filing date by any type of investor. The median number of total requests for 10-Ks in the lowest 3 size quintiles ranges from only 20 to 27 during the 5 days around the filing date. About five 10-K requests per day

Table 6. Mean and median number of nonrobot 10-K requests (*NR_HTM*) during Day 0 through Day +4 by market value quintiles.

Size Quintile	Mean	Median
Small	28.9	20
2	33.0	22
3	36.5	27
4	44.8	32
Large	96.2	56

Note. Table 6 reports the mean and median number of total nonrobot (*NR_HTM*) 10-K requests during the 5 days around the 10-K filing date (Day 0 to Day +4) by market value quintiles, 2003–2012. The market value quintiles are created on an annual basis using all firms with available data.

is not much activity. Further, just because an investor accessed the 10-K on the EDGAR site does not mean they actually read the document and thus were affected by its readability. Because retail investors, by definition, trade relatively few shares, the low counts of EDGAR requests for the 10-Ks of smaller firms cast substantial doubt on notion that retail investors are being affected by the readability of the annual report in the days around the filing date.

Lehavy, Li, and Merkley [2011] and Loughran and McDonald [2014] examined how 10-K readability affects professional analysts. Both articles find that as the readability of the annual report decreases, Wall Street analysts have more difficulty incorporating valuation relevant information into their earnings forecasts. Given that Lehavy et al. [2011] found the average firm is covered by only 6.14 analysts and that the average analyst takes 18.77 days to issue their first report following the 10-K filing, the EDGAR server requests in Table 6 are certainly large enough to suppose that analysts could be in the pool of nonrobots looking at the annual report.

Tobit regressions of firms with and without CRSP data

To quantify some of our analysis, we use Tobit regressions to examine a panel with 401 postfiling day counts in a multivariate setting. In the Tobit regressions, we examine, for a given 10-K filing, the impact of subsequent periodic filings on the request counts and compare the results across the 2 samples. In addition, we use the multivariate setting for the public equity sample, where market data is readily available, to append additional control variables in the analysis and estimate a regression where the dependent variable is the days [0,1] request counts.

Because we know from Table 6 that firm size will be a central determinate in the 10-K request count for publicly traded companies, we will consider a simple bivariate view of the sample of firms with CRSP data before comparing the CRSP and non-CRSP samples. For the full sample, the median CRSP firm experiences 152 requests during the first quarter, or about 1.68 requests per day.

We next create a panel of data, for both the CRSP and non-CRSP samples, where for each firm's 10-K disclosure a time series of *NR_HTM* counts for the days [0,400] relative to the 10-K filing date is generated. This produces a public equity sample of 11,517,522 observations and a smaller nonpublic equity sample of 8,779,494 observations. The distribution of *NR_HTM* is highly skewed due to the occasional spikes in requests, thus we use the natural logarithm of (1+*NR_HTM*) as the dependent variable.

Also, given the distributional data we have already examined, not surprisingly about two-thirds (66.68%) of the *NR_HTM* observations have a value of zero. As a result of this characteristic in the data, we use a Tobit model for the first 3 regressions in columns 1–3 of Table 7 where the dependent variable is the 10-K *NR-HTM* file requests for each day [0,400] relative to the form filing date. We assume in the Tobit regressions that *NR_HTM* is a measure of investor and analyst interest, and this latent variable only causes *NR_HTM* to take on a nonzero value beyond a certain threshold.

The first 3 columns in Table 7 estimate a Tobit model on the panel data where dummy variables demark subsequent filing events and a logarithmic trend captures the obvious decay we expect in the level of interest. Because we are interested in comparing firms with and without readily accessible market and industry data, the first 2 Tobit models in columns 1 and 2 do not include any market-related control variables or industry dummies. The time series variables over the days [0,400] are as follows. *Trading-day dummy* is set equal to 1 if the date of the observation is on a CRSP trading day, else 0. The *10-K(t+1) dummy* is set equal to 1 on days [0,1] when the firm's next 10-K is filed (else 0), with a similar logic defining the *10-Qs(t+1) dummy*.

Thus, the typical observation will have a 2-day window with a subsequent 10-K filing and three 2-day windows with subsequent 10-Q filings. In addition, a *Post 10-K(t+1) dummy* variable is included which is set equal to 1 for all days in the 401-day interval following the 2 day window for the subsequent 10-K filing, else 0. The *Post 10-K(t+1) dummy* variable provides for a drop off in requests anticipated after a new 10-K is filed.

We also include *Log(trend)* which is the natural log of (1+t), where t is each of the day [0,400] observations.[13] In our interpretation of columns 1–3, where the sample sizes are all greater than 8 million, we will focus on the magnitude of the results, because we are nearing the fictitious "population" of classical inference, and the *t* statistics will almost always be statistically significant.

Because the dependent variable in the Tobit regressions is the log transform of *NR_HTM*, the coefficients on the dummy variables can be interpreted as the percentage change in the dependent variable relative to when the dummy is equal to 1. Also recall that in a Tobit analysis, the interpretation of the coefficient should be in the context of the latent variable. As previously noted, with the magnitude of the sample, the level of significance in all the variables is not surprising and we need to consider carefully what the coefficients mean in the context of our analysis.

Interestingly, the coefficient on *Trading-day dummy* is actually larger for the nonpublic equity sample versus the public equity sample (0.955 vs. 0.909), which suggests that the importance of the trading day dummy is primarily attributable to trading days being "work" days. The magnitude of the coefficients for both samples indicates that investor and analyst interest in financial disclosures (as measured by requests) is almost double on business days versus weekends and holidays. This quantifies a pattern that most researchers/investors would expect to exist.

If the information in a 10-K is used as a year-to-year benchmark, we would expect to see an increase in requests when a new 10-K is filed. This is the case for the nonpublic equity sample, with a coefficient of 0.267 on the *10-K(t+1) dummy* variable, but is not reflected in the public equity sample where the coefficient is negative (−0.112). Whether the initial 10-K has already been downloaded, and thus does not require an update on the date of the subsequent filing, or whether this form of benchmarking across annual reports is simply not predominant in public equity firms is not discernable from the sample. However, it is surprising that some benchmarking activity is not observed for the CRSP sample.

The *Post 10-K(t+1) dummy* indicates that there is a substantial drop in requests following the release of a new 10-K across both samples, with a coefficient of −0.643 for the nonpublic equity sample and −0.823 for the public equity sample. Similarly for the subsequent 10-Q filings, the nonpublic equity firms show an increase in requests associated with a subsequent 10-K (with a coefficient of 0.222), while the public equity firms show very little increase (0.047).

As expected, the log trend variable appearing in the first 2 columns has a negative coefficient, and given the log-log relation, can be interpreted as an elasticity. Although the nonpublic equity sample reflects a stronger decay in requests (−0.400 −vs. −0.321), both clearly decay quickly following the initial filing.

Column 3 of Table 7 focuses on the CRSP sample where we can now consider additional control variables derived from market-related data. Larger firms are expected to have higher levels of investor interest, which is consistent with the positive and notably significant coefficient on *Log(market capitalization)*. *Abs(filing date return [0,1])* is included in the Tobit analysis as a proxy for the information content of the 10-K filing. The positive coefficient confirms the expectation that filings with high information content also experience more investor interest, as measured by request activity. The coefficient on *Log(pre-alpha)* provides evidence on the asymmetry of interest, with negative prefiling performance producing more interest among investors (consistent with the evidence of Drake et al. [2015]). If we assume that *Log (pre_RMSE)* captures the uncertainty of the information environment for a firm, then the Tobit results indicate that more volatile information environments heighten the demand for financial information. *Nasdaq*, with a larger number of small firms listed, produces a slight reduction in the level of interest (about 6%). The specific sources and measures for each control variable are defined in Appendix 1.

In column 4 of Table 7, the dependent variable is now a single observation for each filing—the *NR_HTM* counts summed for days [0,1]. The filing date return has more impact on investor attention in this case, however the other variables are consistent with the time series results in column 3. Although not tabulated, the Fama and French [1997] 48 industry dummies included in column 4 provide an interesting breakdown of investor interest. The 5 most negative industry coefficients in

Table 7. Regressions for 10-K filings with *Log(1+NR_HTM)* as the dependent variable.

	Days [0,400] Sample			Days[0,1] Sample
Variable	Nonpublic Equity Sample (1)	Public Equity Sample (2)	Public Equity Sample (3)	Public Equity Sample (4)
Trading-day dummy	0.955	0.909	0.881	
	(484.72)	(886.44)	(1,005.39)	
10-K(t+1) dummy	0.267	−0.112	−0.134	
	(20.54)	(−16.67)	(−23.56)	
Post 10-K(t+1) dummy	−0.643	−0.823	−0.836	
	(−164.88)	(−408.92)	(−480.44)	
10-Qs(t+1) dummy	0.222	0.047	0.037	
	(35.31)	(13.18)	(12.60)	
Log(trend)	−0.400	−0.321	−0.312	
	(−540.02)	(−724.85)	(−838.66)	
Log(market capitalization)			0.329	0.230
			(1,302.19)	(68.72)
Abs(filing date return [0,1])			0.507	1.456
			(84.93)	(17.46)
Log(pre_alpha)			−48.840	−52.676
			(−261.94)	(−21.11)
Log(pre_RMSE)			7.271	9.154
			(327.48)	(30.43)
Nasdaq			−0.064	−0.083
			(−72.03)	(−6.92)
Industry Dummies	No	No	Yes	Yes
Year Dummies	Yes	Yes	Yes	Yes
Constant	Yes	Yes	Yes	Yes
Pseudo/Adjusted R^2	6.14%	8.69%	18.62%	36.49%
Sample Size	8,779,494	11,517,522	11,517,522	28,722
% *NR_HTM*=0	81.73%	55.20%	55.20%	5.08%

Note. Columns 1–3 are Tobit regressions, where the dependent variable is the *10-K Log(1+NR_HTM)* file requests for each day [0,400] relative to the form filing date. *Trading-day dummy* is zero for a CRSP trading day, else zero. *10-K(t+1) dummy* and *10-Qs(t+1) dummy* are set equal to 1 on the filing day, and day after, of a subsequent 10-K or 10-Q for a given firm, else zero. *Post 10-K(t+1) dummy* is equal to 1 for all days after day +1 of the subsequent 10-K filing. *Trend* is a linear trend variable. The other variables are defined in Appendix 1. Industry dummies are based on the Fama and French (1997) 48-industries. Column 4 is an ordinary least squares regression where the time series is collapsed so that there is 1 observation for each 10-K filing and the dependent variable is the total file requests (*NR_HTM*) for days [0,1].

column 4 were Gold, Utilities, Banks, Finance, and Insurance, while the 5 most positive were Beer, Meals, Soda, Clothes, and Retail. This suggests that while exhibiting lower interest in more opaque firms such as utilities and banks, investors tend to be most interested in consumer-related firms they are familiar with.

Conclusions

Information is of central importance in the operation and efficiency of financial markets. What sources of information do investors use in valuing stocks? By viewing investors' demand for financial disclosures by filing type and across time, our results expand this perspective to provide insights into the actual consumption of information. Additionally, from a regulatory standpoint, our results provide an initial view on the usage rate of documents that are typically prepared at substantial costs.

Using the SEC's EDGAR server log, many investor patterns are identified. First, the magnitude of nonrobot investor requests is surprisingly low. The average 10-K filing of a publicly traded firm is requested only 28.4 total times on the day of and the day after the initial filing. The median CRSP firm experiences only 152 requests for its 10-K during the first quarter after the filing (about 1.68 requests per day). The low requests totals for the median firm in the smallest 3 size quintiles (only 5 requests per day around the 10-K filing date), challenges the notion that annual report readability affects the trading behavior of retail investors.

Second, we find that the request quantity and rate over time differ substantially by disclosure type. As might be expected, annual reports are digested slowly, whereas IPO filings are quickly consumed. For example, only 10.1% of all 10-K requests by nonrobots occurring in a 401-day window happen in the first week after the filing date compared to almost 50% for S-1 filings (initial equity or debt prospectus). Last, the pattern for public and nonpublicly traded companies is surprisingly similar. An average firm not on CRSP has its 10-K requested 14.4 times on the day of and day following the filing date compared to 28.4 requests for publicly traded firms.

Arguably, some of the relatively low request rates we have identified could be attributable to investors and analysts using secondary sources to access firms' financial disclosures. The EDGAR filings, however, represent a first-source repository. As noted in the introduction, nonEDGAR sources like Bloomberg and Yahoo! Finance often condense income statement and balance sheet information into prespecified bins. This significantly decreases the worth of such sources to investors doing fundamental research of equity valuations.

From a policy perspective, if the request counts accurately reflect investor interest, the SEC should consider carefully those forms that are rarely accessed. If their information potential is very high, then the public needs to be made aware of the content of these forms. If not, the production costs would suggest the elimination of such forms. Conversely, if the relative low request activity is simply an artifact of investors and analysts using secondary sources, the SEC should make the public more aware of EDGAR's potential as a free website with first-source materials.

Acknowledgments

The authors thank Mohamed Al Guindy, Brad Badertscher, Robert Battalio, Scott Bauguess, Pengjie Gao, Jeremy Griffin, Kathleen Hanley, Peter Kelly, S. P. Kothari (discussant), Kelvin Law, Usha Mittoo, Jay Ritter, Darren Roulstone (discussant), Elvira Sojli, Chuck Trzcinka, an anonymous referee, and seminar participants at the University of Notre Dame, Tilburg University, Erasmus University, Case Western Reserve University, UT-Dallas, FSU SunTrust Beach Conference, FMA Europe conference, Frontiers in Finance conference, and the UC Davis conference on Information and Asset Prices for helpful comments.

Notes

1. See the CNBC blog entitled "How to read a 10-K like Warren Buffett" (http://www.cnbc.com/2014/01/27/how-to-read-a-10-k-like-warren-buffet.html) or the *Wall Street Journal* MoneyBeat blog entitled "It's Time for Investors to Re-Learn the Lost Art of Reading" (http://blogs.wsj.com/moneybeat/2016/04/01/its-time-for-investors-to-re-learn-the-lost-art-of-reading/).
2. See the June 2, 2015 *Wall Street Journal* article (http://www.wsj.com/articles/the-109-894-word-annual-report-1433203762).
3. See the Bloomberg article from April 29, 2016 (https://www.bloomberg.com/view/articles/2016-04-29/amazon-and-facebook-are-big-spenders-on-r-d).
4. Most servers automatically create a log of all activity, usually recorded in a file using the Common Log Format (or some extension of that) endorsed by the World Wide Web Consortium (W3C).
5. A full description of the sample derivation and the verbatim SEC data description is provided in Appendix 2. The SEC's documentation indicates the data begins on February 14, 2003.
6. Note a request not labeled as a web crawler can still be a robot. The user agent record sent as part of the client/server handshake allows web crawlers to self-identify. Although large search firms such as Google have an incentive to self-identify, programs written to download SEC data have no reason to make this declaration, and most robots likely do not.
7. See http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html for a list of server codes. Typically, 300-level codes

indicate a file has been relocated while 400-level codes indicate a client error in the request (e.g., URL not found).

8. Lee, Ma, and Wang's [2015] screens differ slightly from ours. They require the daily IP addresses to be associated with information requests for at least 2 unique firms in a given day, which facilitates their paper's focus on cosearches. Additionally, Lee, Ma, and Wang [2015] restricted the sample to searches for firms in the S&P 1500 and to 10-K, 10-Q, 8-K, 14-A, and S-1 filings. Given the focus of our article, we do not include these additional constraints beyond their initial requirement of an IP having more than 50 requests to be classified as a robot.
9. It is important to note that *NR_HTM* + *NR_TXT* does not equal *NR_Total*. There are other document types, XML for example, which are included in the tabulation of *NR_Total*.
10. Garcia and Norli [2012] documented similar frequencies of EDGAR filings during the 1993–2011 time period.
11. An SEC-provided description of all forms is provided at http://www.sec.gov/info/edgar/forms/edgform.pdf.
12. Companies traded on a national market exchange are required to file pursuant to section 12(b) of the exchange act. Companies with $10 million in assets and 500 shareholders of record are required to file pursuant to section 12(g). The specifics of firms required to file can be found in Sections 12(b), 12(g), and 15(d) of the Securities Exchange Act of 1934, as amended by the Jumpstart Our Business Startups Act (Titles V and VI).
13. We experimented with various functional forms for the trend, including appending additional moments. The logarithmic trend variable provided the best fit and most concise expression for the trend. The specific trend specification had only a very minor impact on the other coefficient estimates.

References

Barberis, N. and R. Thaler. "A Survey of Behavioral Finance." *Handbook of the Economics of Finance*, 1, (2003), pp. 1053–1128.

Cheng, Q., F. Du, X. Wang, and Y. Wang. "Seeing Is Believing: Do Analysts Benefit from Site Visits?" University of Hong Kong, Working Paper, 2014.

Chi, S. and D. Shanthikumar. "The Geographic Dispersion of Google Search and the Market Reaction to Earnings Announcements." University of California, Irvine, Working Paper, 2014.

Cohen, L., A. Frazzini, and C. Malloy. "Sell-side School Ties." *Journal of Finance*, 65, (2010), pp. 1409–1437.

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

Drake, M., D. Roulstone, and J. Thornock. "Investor Information Demand: Evidence from Google Searches Around Earnings Announcements." *Journal of Accounting Research*, 50, (2012), pp. 1001–1040.

Drake, M., D. Roulstone, and J. Thornock. "The Determinants and Consequences of Information Acquisition via EDGAR." *Contemporary Accounting*, 32, (2015), pp. 1128–1161.

Dreman, D. *Contrarian Investment Strategies: The Psychological Edge*. New York: Simon and Schuster, 2012.

Easton, P. and M. Zmijewski. "SEC Form 10K/10Q Reports and Annual Reports to Shareholders: Reporting Lags and Squared Market Model Prediction Errors." *Journal of Accounting Research*, 31, (1993), pp. 113–129.

Fama, E. "Efficient Capital Markets: A Review of Theory and Empirical Work." *Journal of Finance*, 25, (1970), pp. 383–417.

Fama, E. "Efficient Capital Markets: II." *Journal of Finance*, 46, (1991), pp. 1575–1617.

Fama, E. and K. French. "Industry Costs of Equity." *Journal of Financial Economics*, 43, (1997), pp. 153–193.

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), pp. 915–953.

Gabaix, X. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 129, (2014), pp. 1661–1710.

Garcia, D. and O. Norli. "Crawling EDGAR." *Spanish Review of Financial Economics*, 10, (2012), pp. 1–10.

Griffin, P. "Got Information? Investor Response to Form 10-K and Form 10-Q EDGAR Filings." *Review of Accounting Studies*, 8, (2003), pp. 433–460.

Grossman, S. J. and J. E. Stiglitz. "On the Impossibility of Informationally Efficient Markets." *The American Economic Review*, 70, (1980), pp. 393–408.

Ivković, Z., C. Sialm, and S. Weisbenner. "Portfolio Concentration and the Performance of Individual Investors." *Journal of Financial and Quantitative Analysis*, 43, (2008), pp. 613–655.

Kothari, S. P. "Capital Markets Research in Accounting." *Journal of Accounting and Economics*, 31, (2001), pp. 105–231.

Lawrence, A. "Individual Investors and Financial Disclosure." *Journal of Accounting and Economics*, 56, (2013), pp. 130–147.

Lee, C., P. Ma, and C. Wang. "Search-based Peer Firms: Aggregating Investor Perceptions Through Internet Cosearches." *Journal of Financial Economics*, 116, (2015), pp. 410–431.

Lehavy, R., F. Li, and K. Merkley. "The Effect of Annual Report Readability on Analyst Following and the Properties of their Earnings forecasts." *Accounting Review*, 86, (2011), pp. 1087–1115.

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

Loughran, T. and B. McDonald. "Measuring Readability in Financial Disclosures." *Journal of Finance*, 69, (2014), pp. 1643–1671.

Miller, B. P. "The Effects of Reporting Complexity on Small and Large Investor Trading." *Accounting Review*, 85, (2010), pp. 2107–2143.

Appendix 1

Definitions of the variables used in the article

Abs(filing date return [0,1])	The cumulative return for a firm's stock on days [0,1] relative to the 10-K filing.
Log(market capitalization)	The natural logarithm of a firm's stock price times the number of shares outstanding on the day before the 10-K filing date.
Log(pre_alpha)	The natural logarithm of the intercept in a market model regression on days [−252,−1], with a minimum of 66 observations.
Log(pre_RMSE)	The natural logarithm of the root-mean-square error of a market model regression on days [−252,−1], with a minimum of 66 observations.
Log(trend)	The natural logarithm of a trend variable set equal to 1 to 401 for the 400 days following a 10-K filing.
Nasdaq	A dummy variable set equal to 1 if the firm's stock trades on the Nasdaq stock exchange, else zero.
NR_HTM	The total nonrobot server request count for HTM files.
NR_total	The total nonrobot server request count for all file types.
NR_TXT	The total nonrobot count for TXT files.
Robot_count	The total robot server request count. Robots are defined as more than 50 document requests for a given IP on a given day for a filing with at least 1 nonrobot count.
10-K(t+1) dummy	The dummy variable is set to 1 on the filing day, and day after, of a subsequent 10-K for a given firm, else zero
10-Qs(t+1) dummy	The dummy variable is set to 1 on the filing day, and day after, of a subsequent 10-Q for a given firm, else zero.
Post 10-K(t+1) dummy	The dummy variable is set to 1 for all days after day t+1 of the subsequent 10-K filing, else zero.
Trading-day dummy	A dummy variable equal to 1 on a CRSP trading day, else zero.

Appendix 2

EDGAR FOIA request data description

A.1. The FOIA request

In response to our FOIA request, the SEC provided 3,378 data files for the calendar days from January 1, 2003, through March 31, 2012. Their documentation (below) indicates that compilation began on February 14, 2003. The files in the later part of February 2003 are notably smaller than those appearing in the first weeks of March 2003. Thus we initialize all of our analysis beginning with the March 1, 2003, file, exclude some of the clearly corrupted files from September 24, 2005, through May 10, 2006, and conclude with the March 31, 2012, file, resulting in a sample of 3,090 days. The verbatim data description provided by the SEC in response to our FOIA request is provided in the next section of this appendix.

A.2. Data description provided by the SEC

The Division of Economic and Risk Analysis has now assembled information on internet requests for EDGAR filings through sec.gov covering the period February 14, 2003, through March 31, 2012. The information was extracted from Apache log files that record and store user access statistics for the sec.gov website. However, there is incomplete coverage from 09/24/2005 through 05/10/2006 due to damaged/missing log files. There may be additional lost or damaged files during all periods, so that the information assembled by DERA is not necessarily the complete picture of all sec.gov website traffic.

The processed files assembled by DERA are organized by year, month, and day (e.g., log20081231) so that there are 1,782 total files between February 14, 2003, and December 31, 2007, and are SAS formatted. Each file contains an entry for all user requests to sec.gov with the string "GET Archives/edgar/data" in any part of the request line. This indicates that the user is requesting EDGAR-specific filing information. For each user request, 5 of the Apache log file fields (date, time, zone, code and filesize) are directly extracted and recorded. The processed files also contain 10 derived measures, including an obfuscated IP address, 7 measures that capture characteristics of the requested file (cik, accession, extension, and idx), 3 measures that capture user agent information (noagent, browser, and crawler), 2 measures that capture referrer information (norefer and find). The full request, user agent, and referrer fields in the Apache log files are not included due to file size (storage) limitations. Full definitions are below.

1. ip: with ###.###.###.xxx—first 3 octets of the IP address with the fourth octet obfuscated with a 3 character string that preserves the uniqueness of the last octet without revealing the full identity of the IP. For example, all fourth octets of 150 will have the same 3 character string across all files.
2. date: apache log file date
3. time: apache log file time
4. zone: apache log file zone
5. cik: SEC central index key associated with the document requested
6. accession: SEC document accession number associated with the document requested
7. extension: document file type (e.g., html, txt, etc.)
8. code: Apache log file status code for the request
9. filesize: document file size
10. idx: takes on a value of 1 if the requester has landed on the index page of a set of documents (e.g., -index.htm), and zero otherwise.
11. norefer: takes on a value of 1 if the Apache log file referrer field is empty, and zero otherwise
12. noagent: takes on a value of 1 if the Apache log file user agent field is empty, and zero otherwise.
13. find: Numeric values from 0 to 10, that correspond to whether the following character strings /[$string]/ were found in the referrer field—this could indicate how the document requester arrived at the document link (e.g., internal EDGAR search):
    a. $find=0;
    b. if($referrer=~m/.*(action/=getcompany)/) {$find=1};
    c. if($referrer=~m/.*(action/=getcurrent)/) {$find=2};
    d. if($referrer=~m/.*(Find/+Companies)/) {$find=3};
    e. if($referrer=~m/.*(cgi/-bin//srch/-edgar)/) {$find=4};
    f. if($referrer=~m/.*(EDGARFSClient)/) {$find=5};
    g. if($referrer=~m/.*(cgi/-bin//current)/) {$find=6};
    h. if($referrer=~m/.*(Archives//edgar)/) {$find=7};
    i. if($referrer=~m/.*(cgi/-bin//viewer)/){$find=8};
    j. if($referrer=~m/.*(.*/-index)/){$find=9};
    k. k.if($find==0 && $referrer ne "-" && $referrer ne "")){$find=10};
14. crawler: takes on a value of 1 if the user agent self-identifies as 1 of the following webcrawlers or has a user code of 404. Below are the actual Perl regular expressions used.
    a. if($agent=~m/ (wget|Googlebot|polybot|Yahoo/ !/s*Slurp|spider|robot|perl|python|l wp|crawler)/i){$crawl=1};
    b. if($code==404){$crawl=1};
15. browser: 3 character string that identifies potential browser type by analyzing whether the user agent field contained the following /[text]/. Below are the actual Perl regular expressions used.
    a. if($agent=~m/MSIE/){$browser="mie"};
    b. if($agent=~m/Firefox/){$browser="fox"};
    c. if($agent=~m/Safari/){$browser="saf"};
    d. if($agent=~m/Chrom/){$browser="chr"};
    e. if($agent=~m/Seamonk/){$browser="sea"};
    f. if($agent=~m/Opera/){$browser="opr"};
    g. if($agent=~m/(DoCoMo|KDDI|Cricket|Vodaphone)/){$browser="oth"};
    h. if($agent=~m/Windows/s*NT/) {$browser="win"};
    i. if($agent=~m/Mac/s*OS/i) {$browser="mac"};
    j. if($agent=~m/Linux/i){$browser="lin"};
    k. if($agent=~m/iPhone/){$browser="iph"};
    l. if($agent=~m/iPad/){$browser="ipd"};
    m. if($agent=~m/Android/){$browser="and"};
    n. if($agent=~m/(BB10|PlayBook|BlackBerry)/) {$browser="rim"};
    o. if($agent=~m/(IEMobile|Windows/ s*CE|Windows/s*Phone)/){$browser="iem"};

Appendix 3

EDGAR form descriptions for Table 1

This appendix provides a brief description of the specific forms by group that are used in Table 1. In all cases a "/A" suffix implies an amended filing.

	Group label	Forms included	Description
1	10-K	10-K	Annual report. The 405 designation was a check box for filers whose officer or director failed to file a timely Form 4. Due to inconsistency in its use by companies the form type was discontinued in 2002.
2	10-K/A	10-K/A	
3	10KSB	10KSB	10-K filed by small business. This form was eliminated in March of 2009 with all filers now required to use the form 10-K.
4	10KSB/A	10KSB/A	
5	10-Q	10-Q	Quarterly report.
6	10-Q/A	10-Q/A	
7	10QSB	10QSB	Quarterly report filed by small business. This form was eliminated in March of 2009. The hyphenated version is actually a typographical error of a small number of filers.
8	10QSB/A	10QSB/A	
9	8-K	8-K	Current report filing used to notify investors of a material event.
10	8-K/A	8-K/A	
11	S-1	S-1	Registration statement associated with a securities offering.
12	S-1/A	S-1/A	
13	424	424A | 424B1-424B8	Prospectus typically filed immediately after a security offering.
14	3	3	Forms 3, 4, and 5 all pertain to the beneficial ownership of securities and are required for directors, officers, and shareholders with more than 10% of equity. Form 3 is filed as an initial statement, form 4 is a statement of changes, and form 5 is an annual statement of changes.
15	3/A	3/A	
16	4	4	
17	4/A	4/A	
18	5	5	
19	5/A	5/A	
20	SC 13D	SC 13D	SC 13 type filings are required for reporting beneficial ownership of 5% or more of equity securities. To file a "13G" versus "13D", the filing party must own between 5 and 20% of the company and acquire the shares only as a passive investor.
21	SC 13D/A	SC 13D/A	
22	SC 13G	SC 13G	
23	SC 13G/A	SC 13G/A	
24	13F-HR	13F-HR	Initial quarterly holdings report filed by institutional managers.
25	DEF 14A	DEF 14A	Form filed when a shareholder vote is required, typically linked to the annual meeting.
26	497	497	Investment companies are required to file all definitive materials such as proxy statements or prospectuses.
27	6-K	6-K	Any information foreign companies report to local regulators must also be reported as a 6-K.
28	6-K/A	6-K/A	