Information Decay and Financial Disclosures

Tim Loughran
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.8432 *voice*
Loughran.9@nd.edu

Bill McDonald
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.5137 *voice*
mcdonald.1@nd.edu

October 18, 2014

ABSTRACT

Using data from the SEC's EDGAR server log, we examine the consumption of financial information in filings from 2003 through 2012. We document how quickly download counts for EDGAR forms decay over the year following their initial filing. We find that different types of disclosures experience very different information flows as measured by download activity. After examining a variety of the most common forms, we focus on annual reports (10-K) and registration statements (S-1). The magnitude of the daily EDGAR requests for 10-Ks is relatively low and shows a surprisingly small difference between firms with and without publicly-traded equity.

Key words: EDGAR; information decay; web traffic; Form 10-K; Form S-1.

JEL Classifications: G14; G18; M41

We thank Brad Badertscher, Robert Battalio, Scott Bauguess, Pengjie Gao, Jeremy Griffin, Kathleen Hanley, Jay Ritter, Sophie Shive, Chuck Trzcinka, and seminar participants at the University of Notre Dame and UT-Dallas for helpful comments.

The production and consumption of information plays a central role in all things financial. While accounting researchers contribute more generously to the literature on information production, a cross-section of disciplines examines its consumption under the event study microscope. In most event studies, the actual consumption of information is simply proxied by the associated price effects. In this study, we are interested in the assimilation of financial data by investors as measured by activity on the Security and Exchange Commissions' (SEC) web site.

Short-term event studies generally find that the bulk of content in an information release is quickly incorporated into stock prices. Longer term event studies, along with sorts based on alternative attributes, commonly reveal anomalies relative to a null hypothesis of market efficiency, where information is fully and instantaneously impounded. Setting aside the unresolvable joint hypothesis problem, these anomalies imply, at least in some cases, an assimilation of information over time. Perhaps this is the case because the information is complex, thus requiring additional analysis, or possibly because the information is being used as a benchmark for subsequent information flows. For example, an upcoming financial disclosure by Home Depot could cause a concomitant demand for historical disclosures by Home Depot or recent vintage disclosures by Lowe's.

We consider the corporate information environment from an entropy context, where we are interested in the dynamics of information as it devolves from being novel to being fully priced. While many others have focused on the market impact of information, we consider how its consumption decays over time, a factor presumably correlated with the value of information over time. The data source considered is one relatively new to the literature, the server log associated with the SEC's Electronic Data Gathering and Retrieval (EDGAR) web site.

The SEC's server log provides the Internet Protocol (IP) address (anonymized), a timestamp, and the SEC ascension number for every client request. The ascension number can, in turn, be linked to an SEC file containing identifying information for each EDGAR filing. Our dataset covers a period from March, 2003 to March, 2012. The raw data files contain more than four billion file requests and require more than a half terabyte of storage. At this point, the dataset is not publicly available and can be obtained only through a time consuming Freedom of Information Act request. As an artifact of the analysis in this research, we exploit the characteristics of the data to produce a substantially condensed version of the data which we make available for download on the Internet.

After providing some descriptive views of the data, many of which are similar to other studies using a shorter time series, we consider, for the most common EDGAR disclosures, the decay rate for server requests over the year following a form's filing. We then focus on firms' 10-K annual reports and, separately, S-1 filings associated with initial public offerings (IPOs). The S-1 filings provide specific examples where we can explore how the rate of decay relates to other financial variables and how download activity might impact IPO subsequent volatility.

With the EDGAR data, it is critical to separate requests generated by robots from server requests by regular investors (e.g., non-robots). 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. This paper follows the Lee, Ma, and Wang (2014) procedure of defining 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, non-

robot requests have been relatively flat. As might be expected, most requests occur during trading days between 10 a.m. and 6 p.m.[1]

The consumption of financial information by investors varies dramatically across the filing types. The ratio of non-robot requests relative to total filings is highest for the S-1 (security offering registration), 10-K (annual report), and DEF 14A (filed when a shareholder vote is required) forms. Some form types, like Form 4 (changes in beneficial ownership of securities by large shareholders and officers), have millions of total filings available on EDGAR, yet non-robots rarely request the information. Given the broad availability of data services providing information on "insider trading," the relatively low request counts for Form 4 filings suggests that some documents are primarily digested by information intermediaries before being consumed by the public.

We report that 10-K non-robot requests are more temporally diffuse, relative to other form types, over the year following their initial filing, as they are likely used to gain insights into a company's operations or are used for comparison with competitors. In contrast, Form S-1 requests, associated with IPOs, decay quickly after the initial filing (even if we aggregate downloads across subsequent S-1/A filings). Given the request patterns, investors clearly are not accessing the 10-K filings only for information trades on the filing date. For 10-Ks, only 10.1% of all non-robot requests over a 400-day window occur in the first week after the filing compared to almost half for S-1 filings. 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)).

In addition, we consider regressions for the 10-Ks of firms with and without CRSP data as a division that provides a comparison of firms with and without publicly traded equity. The

[1] All reported times in the paper are US Eastern Standard Time.

download patterns surrounding the subsequent 10-Q and 10-K filings differ in ways that might not be expected. Most notably, we do not observe substantially higher download activity for firms with public equity. This might reflect the use of secondary sources by investors and analysts or could suggest that the SEC should make the public more aware of EDGAR as a free distributor of first source disclosures.

For S-1 filings by IPOs like Facebook and Groupon, investors are rapid consumers of the new information on EDGAR. In contrast, investor 10-K requests over the year following the initial filing are relatively diffuse, albeit with some decline in the data requests the farther it is from the filing date. For IPOs, we find that the higher is the number of non-robot requests, the higher is the firm's subsequent stock return volatility.

In section I of the paper we discuss some of the extant research using limited subsets of the EDGAR server data. Section II of the paper describes the data in detail. Section III presents descriptive results for the full sample and for focused subsets of the sample. Section IV concludes.

I. 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 et al. (2011), Drake et al. (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 et al. (2014)). Cohen et al. (2010) even use 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 et al. (2014) identify economically-related peer firms by examining the sequence of chronologically adjacent EDGAR searches by non-robot 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 et al. (2014) 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 six digit Global Industry Classification Standard. There is clearly a systematic logic in the way investors interact with the EDGAR filings.

The two papers most related to our work are Drake et al. (2014) and Bauguess et al. (2013). Both of these papers use 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. (2014) focus primarily on what variables and events determine web traffic on EDGAR. The three authors find that non-robot EDGAR requests by investors are positively related to important corporate events, like restatements and earnings announcements, firm size, and weak abnormal stock performance. Drake et al. also find that EDGAR requests have a positive influence on the price discovery process associated with earning announcements.

When analyzing data from the EDGAR server log, the manner in which robot requests are separated from non-robot requests is absolutely critical. We follow the procedure of Lee et al. (2014) in classifying IP addresses with more than 50 unique firm filing requests in a given day as being robot requests. Differently, Drake et al. (2014) define 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 non-robot request counts between the papers. For example, we find a total of 137,013 non-robot requests for Form 4 filings during March, 2003 to March, 2012 while Drake et al. (2014, Table 3) report a total of 16,941,014 requests by non-robot investors for this somewhat obscure EDGAR filing. For non-robots, our ratio of 10-K requests to Form 4 requests is 251. That is, for every Form 4 request, non-robot 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 non-robots 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. In an internet appendix, we report the top 25 downloads by robots. Interestingly, more than half of the most common robot server requests are for Form 4 filings. Robots, not humans, are the frequent consumers of the information contained in Form 4 filings.

In another important difference between the papers, we examine EDGAR requests on all days while Drake et al. (2014) specifically exclude holidays and weekends from their empirical analyses. Although weekends and holidays have significantly lower EDGAR requests than occur on trading days, there is still substantial activity by investors on non-trading days. We find that the average combined total EDGAR requests on Saturday and Sunday is more than 630,000.

Like our paper, Bauguess et al. (2013) examine investor usage of Form S-1 filings before a company goes public. They find that higher EDGAR search traffic scaled by IPO proceeds is linked with higher IPO first-day returns, larger absolute offer price revisions, and greater

probability of completing the IPO. In contrast with Bauguess et al., we focus on linking investor requests with the IPO's subsequent stock return volatility.

II. Data

In this section of the paper, we will 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 four billion documented server requests and requiring more than one-half terabyte of storage. As an artifact of the empirical tests in this paper, 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 http://---.---.---.

A. 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.[2] Although the SEC EDGAR server log data is not publicly available at this time, it can be obtained through a Freedom of Information Act (FOIA) request to the SEC. The log files are only available from 2003 forward. The time series provided by the SEC is incomplete in some instances, thus we are faced with an important, yet imperfect dataset. In the following paragraphs we will detail the process of building the dataset we examine in this study.

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

[2] 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).

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.[3] 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 three codes reported with each record. First, any request flagged as a web crawler is excluded from the sample.[4] Second, any index page request is excluded. Finally, all requests with server codes of 300 or greater are excluded.[5] These three filters reduce the original data from more than four 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 one non-robot (NR) request, we are able to summarize the data into about 113 million observations from the original sample of more than four billion. How we identify robot requests follows.

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 2008-2011, Lee, Ma, and Wang (2014) (LMW hereafter) provide a detailed derivation for a heuristic rule that filters out robots. Presumably, these downloads are not representative of attempts to measure information flow for a specific firm.

[3] A full description of the sample derivation and the verbatim SEC data description is provided in Appendix B. The SEC's documentation indicates the data begins on February 14, 2003.

[4] 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.

[5] 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).

LMW show that 95% of all daily downloads associated with a specific IP are for no more than 50 unique 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 one valid non-robot request. We partition all requests for that filing initially into two variables, *Robot_count* and *NR_total* (non-robot total), where *NR_total* represents counts with fifty or fewer requests on a particular day for the IP address associated with the request.

LMW argue 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 non-robot HTM files (labeled *NR_HTM*) and non-robot TXT files (labeled *NR_TXT*).[6] After an initial examination of all the variables, most of our analysis will focus on the non-robot HTM files (*NR_HTM*). Drake et al. (2014) do not distinguish between HTM and TXT requests.

Each record in the original SEC data provides the IP address associated with the request. Such data could be useful in examining geographic-related patterns, however, all of the IP addresses have been obfuscated in a way that maintains the uniqueness of the address, but masks its specific identity. Thus, although we use the IP address to identify robots, we do not retain this information in the data summarization process. Although collapsing the data by firm/day and not retaining individual IP-related records allows us to shrink the data by orders of magnitude, note that those studies, like Lee et al. (2014), wishing to trace the downloads of a given user across time will have to revert to the full SEC dataset.

[6] 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*.

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, etc.) 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 one on trading days, else zero; 3) the Central Index Key (CIK) assigned to the filer; 4) the total count for all non-robot server requests (*NR_Total*); 5) the count for all non-robot requests for HTM files (*NR_HTM*); 6) the count for all non-robot 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 non-robot requests, we maintain separate counts for HTM and TXT files. As mentioned before, the specific strategy used to produce our final dataset results in a file of reasonable size which can be readily distributed and analyzed by others. Although our initial description of the calendar/time characteristics of the sample will include all valid requests from the original SEC sample, our subsequent analysis relating to forms will utilize the compressed data.

B. Forms Included in the Analysis

Our analysis will consider the downloading patterns of EDGAR filings by form type during 2003 through 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 I. As can be seen in the groupings of Table I, we separately consider appended filings, quarterly versus annual filings, and small business filings. Not surprisingly, quarterly reports appear almost three 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 I, we will focus our efforts on 10-K filings (since they are a significant source of financial information of a firm) and S-1 filings for a sample of IPOs. A brief description of each filing type we consider is provided in Appendix C.[7]

What stands out in Table I 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.

III. Results

A. 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, Roulstone, and Thornock (2014), in a contemporaneous paper, provide similar calendar descriptions of the server traffic from 2008-2011. As discussed earlier, their sample uses a very different approach for filtering robots. Note that the counts discussed in this section are not aggregated by form type and represent the count of valid server requests by time series slice. That is, Figure 1 presents the total counts by year

[7] An SEC-provided description of all forms is provided at: http://www.sec.gov/info/edgar/forms/edgform.pdf.

and month, while Figures 2, 3, and 4 present the average downloads per calendar period of interest (e.g., month or hour).

Our results focus on counts of how many times SEC forms are downloaded. As a benchmark of the pool that these requests target, the average number of SEC forms filed per day ramped up from about 1,000 in the first full year requiring electronic submission (1997) to a peak of about 4,500 in 2007, with the 2013 levels at approximately 3,900.

Figure 1 displays the count for all valid requests (robots and non-robots) for each unique year/month in the eleven 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 non-robot 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 non-robot requests since other sources provide the same data for investors. Also, many retail investors receive their 10-K filings through the U.S. mail instead of accessing them on EDGAR. 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 non-robot requests in the last two months of the sample. For reasons we have been unable to determine, the count of non-robot

downloads files drops substantially in the last two months of the sample, even though total downloads increases.

In Figure 2, we display the server requests by month. In this chart, we report the *NR_total*, *NR_HTM*, and *NR_TXT* counts. Both the *NR_HTM* and *NR_TXT* counts track the *NR_total* count across months. With the predominance of firms whose fiscal year end is December 31, there appears to be an increase in the demand for information as new disclosures become available in February and March. The summer months produce lower levels of demand for information. Although the robot count is not plotted in the chart, it produces a similar pattern with a low of about 372 thousand counts in June and peaking at approximately 632 thousand requests in February.

Figure 3 presents the *NR_total* counts and *Robot_counts* by week day. In addition, the counts are reported for non-weekend holidays (non-trading days). Not surprisingly, for both the *Robot_count* and *NR_total*, weekends generate less traffic. Similarly non-weekend holidays exhibit a lower level of requests, however, not quite to the extent of weekends. Clearly the counts are differentiated for non-business days; yet they are not dropping to an extent where these periods can be ignored. As noted earlier, Drake et al. (2014) only focus their analysis on trading days.

Finally, for completeness we also present the server requests by hour in Figure 4. As should be expected, the *NR_total* peaks during the business day. Our use of the LMW (2014) threshold of 50 to categorize requests by robots appears to be quite effective. There are relatively few non-robot server requests when most US investors are asleep (1 a.m. to 7 a.m.). The quantity of non-robot 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. Since automated requests frequently run over a long

period with no need for human intervention, the robot count is less differentiated during business hours. Interestingly, robot requests peak at 5 p.m., suggesting that some more targeted robot downloads are harvested immediately following the close of the market.

B. The Distribution of Daily Filing Downloads

We now focus on the compressed sample where each daily count represents the number of file downloads for a given EDGAR filing on that day for all forms with at least one valid non-robot server request. Table II provides selected percentiles for the distribution of the various counts. LMW emphasize the "power law" nature of the server request data. That is, the number of downloads 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 II where the median count total is one or zero for each count classification, the 99th percentile is less than 48 for all counts, while the maximum is over 100,000 for all counts except *NR_TXT*. As noted by LMW, the small extreme value for *NR_TXT* is attributable to the fact that most text file downloads are likely to be associated with robot downloads. 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 one non-robot download (of any file type) on a given day.

Subsequently in this section and in the rest of the paper, we will focus our analysis on the count represented by *NR_HTM*, which, consistent with the observations of LMW, should best represent the consumption of disclosure information by individuals (versus programmed robots).

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 downloaded more than 6,000 times by investors in a given day. Table A.2 in the internet appendix lists the top 25 server downloads for non-robot HTM files. Six of the top 15 download counts, including the top two of 115,558 and 111,490 requests, are the day of Facebook's S-1 (initial public offering prospectus) filing and the subsequent five days. The third and fourth largest form downloads are associated with American International Group as the firm was going through bankruptcy, with 29,191 10-K filings downloaded 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 downloads are associated with the day of the S-1 filings for Groupon (28,442) and Zynga (27,674).

To provide a meaningful measure of *N_days* – the number of days between a form's filing date and the day of the download – 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 downloads across observations with differing potential time spans, we will define the observation interval as approximately one

calendar year (365 days). To allow all forms to have equal potential time spans, we terminate the observations one 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 download entry for the 365-day post-filing period. Because each observation in the dataset we have created provides summary counts, the *N_days* statistics reported in Table II are weighted by the frequency count for *NR_HTM*.

In the last row of Table II, the value of one for the 10th percentile of *N_days* indicates that about one-tenth of a form's downloads 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 at these one-year decay rates in the context of specific form types and other variables.

C. Measuring Information Decay using the EDGAR Server Log Data

The central 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 5 charts *NR_HTM* for the 10-K filing of General Electric (GE) on February 2, 2007. We chose this date so that the subsequent time interval does not include the segment of missing data. 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 downloads. 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 5 supports this contention.

In Figure 5, 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 drop-off 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 downloads falling within selected calendar periods following the initial filing.

D. Information Decay Rates by EDGAR Form Type

For clarity, we do not report in Table III all of the form group variants itemized in Table I. Also, in reporting the sequence of time categories, we exclude the third quarter. Table A.1 in the internet appendix provides a complete itemization of all of the Table I forms.

Similar to our previous description for *N_days*, we tabulate the counts for the non-robot 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) through (6) of Table III, we tabulate the percent of *NR_HTM* downloads into approximate calendar groupings within the [0,400] day post-filing 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 the fourth quarter (days [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) through (10) of Table III, we also report a series of numbers that provide a sense of the total form downloads 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 download over the 400 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 downloads from the first 400 days occurring before the end of the second quarter. From the table, we can also calculate the weighted average percentile by calendar period where the weight is determined by the total

number of filings for each group. The weighted average across all forms for days [0,1] is 17.2%, for one week is 31.4%, for 2nd quarter 81.1%, and 3rd quarter 90.4%.

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] downloads all exceeding 25%, and easily more than half of their total days [0,400] downloads occurring within the first month after filing. Although some of this decay could be attributable to subsequent S-1/A filings replacing the initial S-1 filing, later in the paper when we aggregate across the various S-1 and amended filings, we arrive at a similar conclusion.

The rate of information decay is clearly different for the 10-K, which has the lowest percent of filing date downloads (4.7%), and has only 60% of the total downloads 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 downloads by relating them to the number of filings. This comparison highlights the important but differential nature of 10-K versus S-1 filings. Both the 10-K and S-1 forms are dominant in terms of the number of downloads per filing, with both having more than 337 downloads for a given filing. At the same time, the S-1, associated with an initial filing for a public offering of debt or equity, has almost half (47.8%) its days [0,400] downloads occurring within the first week, compared with only about 10% for the 10-K. Although the information for these two forms decays at very different rates, they are clearly both considered important by consumers of financial information.

Also notice in the Table III results that 10-Ks attract much more interest from investors than 10-Q filings. There are 389.45 non-robot requests per 10-K filing compared to only 106.76

non-robot 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 downloads per filing for those forms not explicitly separated out in Table III (category "Other"), and accounting for approximately 70% of all filings, is about seven. That is, each filing, on average, is viewed only seven times during this sample interval. Four of the fourteen form-specific categories in Table III have fewer than 10 file downloads per filing over the entire 401 day interval. Surprisingly, Form 13F-HR, (quarterly holdings filed by institutional managers), has a total of only 152 *NR_HTM* downloads for 110,364 filings.

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 downloads of certain filing types might be due to investors being unaware of its availability.

E. 10-K Filings

In the prior sections, we have considered the server downloads without linking the data to any other sources, except the EDGAR Master File. We present those results as representative of all firms impacted by SEC mandates. In this section, we will focus in more detail on the annual 10-K report 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.[8] This division of the sample is interesting because, to the extent investors are using 10-Ks downloaded from the SEC website to assess a firm's stock, we would expect the download activity of the sample with actively traded public equity (i.e., the CRSP sample) to exhibit substantively higher download activity. In addition, most prior research has focused primarily on the equity investors of publicly-traded companies (see Kothari (2001)).

We will first consider some descriptive results for 10-K downloads and then use Tobit regressions to examine a panel with 400 post-filing 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 download counts and compare the results across the two 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] download counts.

The top ten 10-K filing downloads 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 downloads per day in the first quarter following the filing. Given the power-law nature of downloads, the firms with the highest number of downloads are clearly extraordinary outliers.

Because we know that firm size will be a central determinate in the 10-K download count

[8] 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).

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. Figure 6 presents the median downloads for the first post-filing quarter and [0,400] day interval for 10-K filings by size (stock price x number of outstanding shares) quintiles. Not surprisingly, the magnitude of download activity increases directly with size. For the full sample, the median CRSP firm experiences 152 downloads during the first quarter, or about 1.68 downloads per day. The smallest quintile of firms has a median of one download per day, while the largest firms in the top quintile of the New York Stock Exchange have a median download greater than six per day during the first quarter. The proportion of downloads occurring in the first quarter relative to the day [0,400] window decreases directly with size, with 44.4% of the total downloads for the full period occurring in the first quarter for the smallest firms and only 37.0% for the largest firms. Thus, usage rates persist at higher levels for larger firms, while smaller firms experience a more substantial decay in 10-K activity.

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 non-public equity sample of 8,779,494 observations. The distribution of *NR_HTM* is highly skewed due to the occasional spikes in downloads, 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 three regressions in columns (1)-(3) of Table IV where the dependent variable is the 10-K *NR-HTM* file downloads 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 non-zero value beyond a certain threshold.

The first three columns in Table IV 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 two 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 one if the date of the observation is on a CRSP trading day, else zero. The *10-K(t+1) dummy* is set equal to one on days [0,1] when the firm's next 10-K is filed (else zero), with a similar logic defining the *10-Qs(t+1) dummy*.

Thus, the typical observation will have a two-day window with a subsequent 10-K filing and three two-day windows with subsequent 10-Q filings. In addition, a *Post 10-K(t+1) dummy* variable is included which is set equal to one for all days in the 401-day interval following the two day window for the subsequent 10-K filing, else zero. The *Post 10 K(t+1) dummy* variable provides for a drop off in downloads 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.[9] In our interpretation of columns (1) through (3), where the sample sizes are all greater than 8 million, we will focus on the magnitude of the results, since we are nearing the fictitious "population" of classical inference, and the *t*-statistics will almost always be statistically significant.

[9] 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.

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 one. 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 non-public equity sample versus the public equity sample (0.955 versus 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 downloads) 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 downloads when a new 10-K is filed. This is the case for the non-public 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. In addition, the *Post 10-K(t+1) dummy* indicates that there is a substantial drop in downloads following the release of a new

10-K across both samples, with a coefficient of -0.643 for the non-public equity sample and -0.823 for the public equity sample.

Similarly for the subsequent 10-Q filings, the non-public equity firms show an increase in downloads associated with a subsequent 10-K (with a coefficient of 0.222), while the public equity firms show very little increase (0.047). The difference in download activity surrounding subsequent filings across the two samples would suggest that the use of past 10-Ks as a benchmark is much different for firms with and without public equity.

Regardless of the sample partition, given that the number of daily downloads is small, the coefficients do not imply a substantial change in downloads on the date of a subsequent 10-K filing. We can further consider the window of time when a subsequent 10-K for the firm is likely to be filed by examining server requests during the event window of one year (365 days) plus or minus 30 days. During this sixty-one day interval, the total median (mean) number of downloads across the sample is 15 (40.5), and 50 (97.9) for the largest NYSE decile. Thus, even for the largest firms, the expected number of downloads per day during this period of time is somewhat less than two, which from the prior results suggests only a modest change in downloads on the critical dates. In sum, although there is a measurable impact of subsequent filings, and this impact seems to differ across firms with and without public equity, the magnitude of this effect is not sizeable. This result is consistent with the single observation of GE reported before in Figure 5 (i.e., we do not observe notable spikes on subsequent filing dates).

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

Column (3) of Table IV focuses on the CRSP sample where we can now consider additional control variables derived from market-related data. As expected from the descriptive results in Figure 6, 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 download activity. The coefficient on *Log(pre-alpha)* provides evidence on the asymmetry of interest, with negative pre-filing performance producing more interest among investors (consistent with the evidence of Drake et al. (2014)). 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 the appendix.

In column (4) of Table IV, 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 48 industry dummies included in column (4) provide an interesting breakdown of investor interest. The five most negative industry coefficients in column (4) were Gold, Utilities, Banks, Finance, and Insurance, while the five 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.

One could argue that the specific patterns surrounding downloads on the date of the subsequent 10-K filing simply are not captured by the particular combination of time dummies we employ in the regressions. To address this issue, Figure 7 plots the mean and 95th percentile for each sample for days [0,30] relative to the filing date and days [-30, 30] relative to the filing of the subsequent 10-K.[10] Figure 7 clearly shows average daily downloads dropping to less than a handful for both samples in the first few days following the initial filing, and a clear drop in interest following the filing of a new 10-K.

More notably, Figure 7 also contradicts the notion that individual investors and analysts are actively downloading 10-Ks for valuing stocks. Although the difference in downloads between the public and non-public equity samples is evident in the figure, it is not overwhelming. An average firm without publicly traded equity would have its 10-K downloaded about 15 times on the day of and day following the filing date, while an average firm whose equity is publicly traded would have their 10-K downloaded on these days approximately 27 times. This is a difference not conditioned on firm size, which is presumably much smaller for the non-CRSP sample. If the majority of investors and analysts are using secondary sources to obtain financial data, then we cannot draw strong conclusions on usage. However, if first source materials are considered important, then the SEC needs to increase the public's awareness of this free source of financial disclosures.

F. Investor S-1 Filing Requests of IPO Firms

In this section, we will focus on S-1 and S-1/A (amended) filings associated with initial

[10] Note that due to the days [0, 400] event window, not all firms will have 30 post-filing days available following the subsequent 10-K.

public offerings of stock. We append the EDGAR server downloads to the IPO dataset of Loughran and McDonald (2013), which is an extension of Professor Ritter's IPO database.[11]

There are primarily three dates of interest in the IPO sample – the S-1 filing date, the download date for a given measure of *NR_HTM*, and the IPO offer date. The S-1, however, is frequently revised in the process of going public (see Hanley and Hoberg, 2012). Thus, we also consider the amended filings (S-1/A), since we would expect download activity to shift with the most recent version of the amended S-1. In our IPO sample, the mean and median number of S-1/A filings is between 4 and 5.

Given that the number of days between the S-1 filing and offer date in the original Loughran-McDonald (2013) sample ranges from 21 to 1,525, Figure 8 plots the download activity (via *NR_HTM*) for IPOs in the days immediately following the first S-1 filing (days [0, 10]) and, counts for the final S-1/A filing occurring at least 10 days before the offer date. The darker bars in Figure 8 present the median percentage of downloads, relative to the total downloads for a given form, overlaid with the raw *NR_HTM* count (in grey, right-hand-side scale). In the left segment of the figure, the S-1 downloads from day [0, 10] are displayed, while the right segment displays the percentage and counts for last S-1/A filing occurring at least 10 days before the offer date. Clearly for S-1 filings, much of the interest is focused on the period immediately following the filing, with the first two days accounting for almost 17% of the total downloads. Because most S-1s are superseded by subsequent S-1/A filings, we do not present in the figure the S-1 activity immediately preceding the offer date, although we can estimate from the data that this number is less than 1% for all days [-10, 0].

If we tabulate from the sample the total number of *NR_HTM* downloads for each IPO in

[11] The original Loughran-McDonald sample is 1,887 IPOs, and only 550 of those occur after the beginning date of our EDGAR server data and have available server data from their S-1 filing date to the IPO date. The sample expands slightly to 552 when the analysis focuses on the nine days around the offering.

days [0, 10] and divide that by the total *NR_HTM* downloads from filing date to offer date, the median is 51.2%, with a mean of 50.2%. Thus, for the typical IPO, approximately half of the S-1 downloads take place within the first 10 days following the initial filing. For the S-1/A filing preceding the offer date, the percent of total download activity peaks again around 10%, however notice that the median count is substantially lower than the initial S-1 activity. Clearly the S-1 plays a critical role in the initial dissemination of information with amended filings supplementing the original disclosure.

G. Subsequent IPO Volatility

The number of EDGAR server requests should proxy well for investor interest in the IPO. More requests from different investors should signify higher divergence of opinion on the IPO's valuation which should be positively related to subsequent stock return volatility. In the regressions presented in Table V, all columns have subsequent stock return volatility (root-mean-square-error (RMSE)) as the dependent variable. RMSE is from a market model using trading days [+5, +60] relative to the IPO date.

We include eight explanatory variables known as of the IPO date which potentially could explain subsequent volatility. The eight variables are (1) Log(proceeds) – global gross proceeds before the over-allotment option; (2) Log(1+age) – the natural logarithm of one plus the age of the firm at the time of the IPO using updated data from Loughran and Ritter (2004); (3) First-day returns – the percentage change from the offer price to the closing price on the first day of trading; (4) VC dummy – a dummy variable equal to one if the IPO is backed by a venture capitalist, else zero; (5) Top-tier dummy – a dummy variable set to one if the lead underwriter has an updated Carter and Manaster (1990) and Carter, Dark, and Singh (1998) ranking of 8 or

more, else zero; (6) Positive EPS dummy – a dummy variable set to one if trailing earnings per share at the time of the IPO are positive, else zero; (7) Prior Nasdaq 15-day returns – the buy-and-hold returns for the CRSP Nasdaq value-weighted index in the 15 days prior to the offering; and (8) Up revision – the upward adjustment in the IPO offer price relative to the mid-point of the filing range if the offer price is greater than the mid-point, else zero. In all the Table V regressions, we include an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with the standard errors clustered by year and industry.

For the IPO sample, we tabulate *NR_HTM* (the cumulative number of S-1 and S-1/A server requests) over two non-overlapping windows: (1) from the S-1 filing date until 5 days before the IPO date and (2) +/- 4 days around the IPO date. The first two columns of Table V use the longer window while the last two columns focus the event window on the nine days around the initial offering. Column (1) reports that larger firms (in terms of gross proceeds) and older firms have lower subsequent stock return volatility. It is important to note that IPO first-day returns have no significant linkage with subsequent volatility. When *Log(NR_HTM)* is added to the column (2) regression, the coefficient on *NR_HTM* is positive and statistically significant (*t*-statistic of 2.53) in the presence of the eight control variables. When more non-robot investors request information on the firm, the IPO is associated with higher subsequent volatility.

The last two columns of Table V report regression results when *NR_HTM* is tabulated only over days +/- 4 days around the IPO date. In the presence of the control variables, the coefficient on *NR_HTM* is positive and significant (t-statistic of 6.36) in column (4). As more investors access the S-1 and S-1/A filings at the time of the IPO, the larger is the number of investors who accumulate information on the newly public company. As noted by Miller (1977, p. 1151), “in

practice, the very concept of uncertainty implies that reasonable men may differ in their forecasts." More investors with differing opinions on the IPO valuation should be associated with higher subsequent volatility for the IPO.

IV. Conclusions

Information is of central importance in the operation and efficiency of financial markets. Much of what we know about the assimilation of information is based on the indirect observation of market price responses around information events. 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. Since 2008, robot server requests have spiked upwards while non-robot requests have been generally flat. Weekends experience nontrivial counts from both robots and non-robots. During the day, robot requests peak at 5 p.m. while non-robot server requests sharply rise once people are at work (10 a.m.) and continue strong for a few hours after the stock market's 4 p.m. close.

We find that the download 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 occurring in a 401-day window happen in the first week after the filing date compared to almost 50% for S-1 filings. We find that, controlling for other factors, there are almost twice as many non-robot requests on trading days

as on non-trading days (i.e., weekends or holidays). However, even for the largest decile of NYSE firms, this only implies a doubling of one request to two for the average filing.

It is not clear from the data that prior 10-K filings are being used as a definitive benchmark for current filings. The magnitude of the daily EDGAR requests for 10-Ks is relatively low and, importantly, shows a surprisingly small difference between firms with and without publicly traded equity.

Many of the required filings are not actively downloaded. There are more than four million Form 4 filings on EDGAR, yet the number of non-robot requests totals only 137,013 over the March, 2003 to March, 2012 period. Form 13F-HR is even more neglected with only 152 non-robot requests during the entire time period.

For an IPO sample, we find that non-robot requests have a positive linkage with subsequent stock return volatility. This is true for a window from the S-1 filing date to five days prior to the IPO and for a period +/- 4 days around the firm's IPO date. More attention from investors is associated with wider stock return fluctuations.

Arguably, some of the relatively low download rates we have identified could be attributable to investors and analysts using secondary sources to access firms' financial disclosures. From a policy perspective, if the download 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 download 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.

Appendix A. Definitions of the variables used in the paper.

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 one if the firm's stock trades on the Nasdaq stock exchange, else zero.
NR_HTM	The total non-robot server download count for HTM files.
NR_total	The total non-robot server download count for all file types.
NR_TXT	The total non-robot count for TXT files.
Robot_count	The total robot server download count. Robots are defined as in Lee et al. (2014).
10-K(t+1) dummy	The dummy variable is set to one 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 one 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 one for all days after day t+1 of the subsequent 10-K filing, else zero.
Trading-day dummy	A dummy variable equal to one on a CRSP trading day, else zero.
<u>IPO Variables</u>	
Log(1+age)	The natural logarithm of one plus the age of the firm at the time of the IPO using updated data from Loughran and Ritter (2004).

First-day return	Defined as the percentage change from the offer price to the closing price.
Positive EPS dummy	Dummy variable set to one if trailing EPS is positive at the time of the IPO, else zero.
Prior Nasdaq 15-day returns	The buy-and-hold returns of the CRSP Nasdaq value weighted Index on the 15-trading days prior to the IPO date, ending on day t-1.
Log(proceeds)	The natural logarithm of the global gross proceeds before the over-allotment option.
RMSE	RMSE is from a market model estimated using trading days [+5, +60] relative to the IPO date. At least 30 observations of daily returns must be available on CRSP to enter the sample.
Top-tier dummy	Dummy variable set equal to one if the lead underwriter of the IPO has an updated Carter and Manaster (1990) rank of eight or more, else zero.
Up revision	Defined as the upward adjustment in the IPO offer price relative to the mid-point of the filing range if the offer price is greater than the mid-point, else zero.
VC dummy	Dummy variable set to one if IPO is backed by venture capital, else zero.

Appendix B: 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, five of the Apache log file fields (date, time, zone, code and filesize) are directly extracted and recorded. The processed files also contain ten derived measures, including an obfuscated IP address, seven measures that capture characteristics of the requested file (cik, accession, extension, and idx), three measures that capture user agent information (noagent, browser, and crawler), two 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 three 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 three 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 one if the Apache log file referrer field is empty, and zero otherwise
12. noagent: takes on a value of one 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)/){$find=9};
    k. if($find==0 && $referrer ne "-" && $referrer ne ""){$find=10};
14. crawler: takes on a value of one if the user agent self-identifies as one 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: three 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 C: EDGAR form descriptions for Table I

This appendix provides a brief description of the specific forms by group that are used in Table I. 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
8	10QSB/A	10QSB/A	
9	8-K	8-K	Current report filing used to notify investors of a material events.
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 ten percent 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	

References

Bauguess, S., Cooney, J., Hanley, K., 2013, Investor demand for information in newly issued securities, Working paper, University of Maryland.

Carter, R., Dark, F., Singh, A., 1998, Underwriter reputation, initial returns, and the long-run performance of IPO stocks, *Journal of Finance* 53, 285–311.

Carter, R., Manaster, S., 1990, Initial public offerings and underwriter reputation, *Journal of Finance* 45, 1045–1068.

Cheng, Q., Du, F., Wang, X., Wang, Y., 2014, Seeing is believing: Do analysts benefit from site visits?, Working paper, The University of Hong Kong.

Chi, S., Shanthikumar, D., 2014, The geographic dispersion of Google search and the market reaction to earnings announcements, Working paper, University of California, Irvine.

Cohen, L., Frazzini, A., Malloy, C., 2010, Sell-side school ties, *Journal of Finance* 65, 1409-1437.

Da, Z., Engelberg, J., Gao, P., 2011, In search of attention, *Journal of Finance* 66, 1461-1499.

Drake, M., Roulstone, D., Thornock, J., 2012, Investor information demand: Evidence from Google searches around earnings announcements, *Journal of Accounting Research* 50, 1001-1040.

Drake, M., Roulstone, D., Thornock, J., 2014, The determinants and consequences of information acquisition via EDGAR, *Contemporary Accounting Research*, forthcoming.

Easton, P., Zmijewski, M., 1993, SEC form 10K/10Q reports and annual reports to shareholders: Reporting lags and squared market model prediction errors, *Journal of Accounting Research* 31, 113-129.

Fama, E., French, K., 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153-193.

Griffin, P., 2003, Got information? Investor response to Form 10-K and Form 10-Q EDGAR filings, *Review of Accounting Studies* 8, 433-460.

Hanley, K., Hoberg, G., 2012, Litigation risk, strategic disclosure and the underpricing of initial public offerings, *Journal of Financial Economics* 103, 235-254.

Kothari, S.P., 2001, Capital markets research in accounting, *Journal of Accounting and Economics* 31, 105-231.

Lee, C., Ma, P., Wang, C., 2014, Search based peer firms: Aggregating investor perceptions through internet co-searches, *Journal of Financial Economics*, forthcoming.

Loughran, T., McDonald, B., 2013, IPO first-day returns, offer price revisions, volatility, and form S-1 language, *Journal of Financial Economics* 109, 307-326.

Loughran, T., Ritter, J., 2004, Why has IPO underpricing changed over time?, *Financial Management* 33, 5–37.

Miller, E. M., 1977, Risk, uncertainty, and divergence of opinion, *Journal of Finance* 32, 1151-1168.

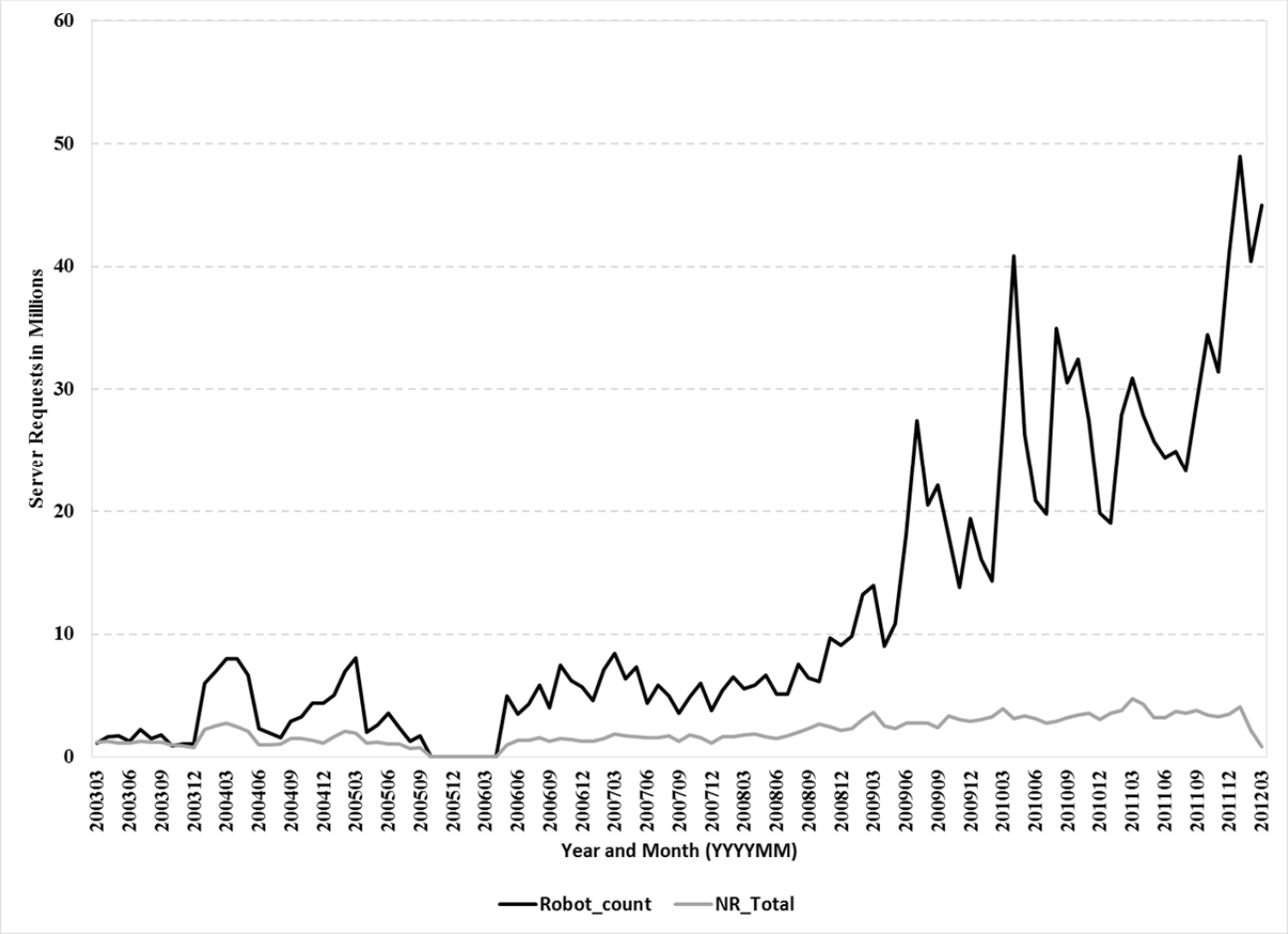

Figure 1. A plot of the robot and non-robot count of EDGAR server requests, in millions, for each unique year and month during March, 2003 to March, 2012. The blank space from September, 2005 to May, 2006 reflects the period when the server files were corrupted.

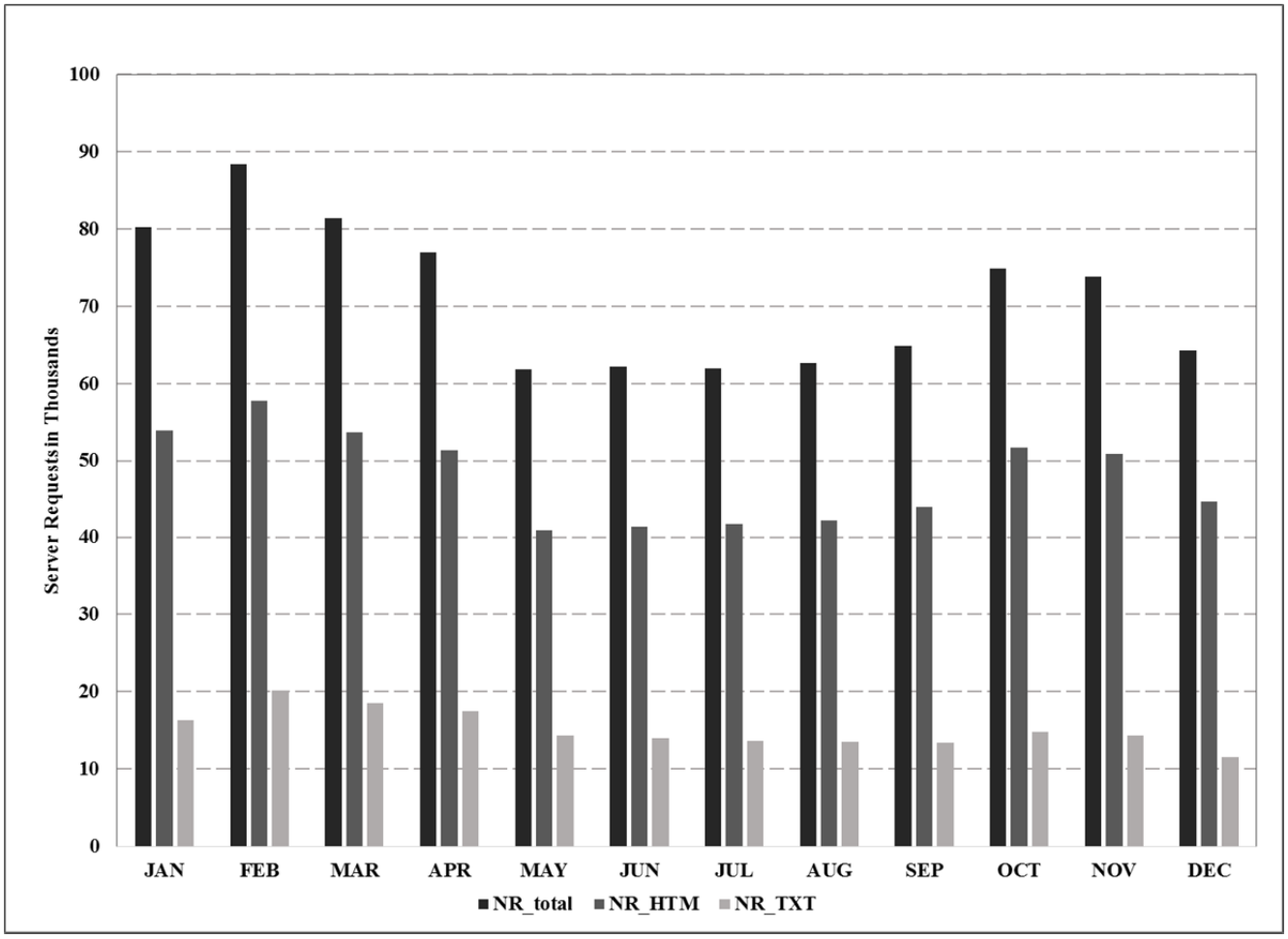

Figure 2. A plot of the total non-robot server request, non-robot HTM file request, and non-robot TXT file request averages by month for March, 2003 through March, 2012.

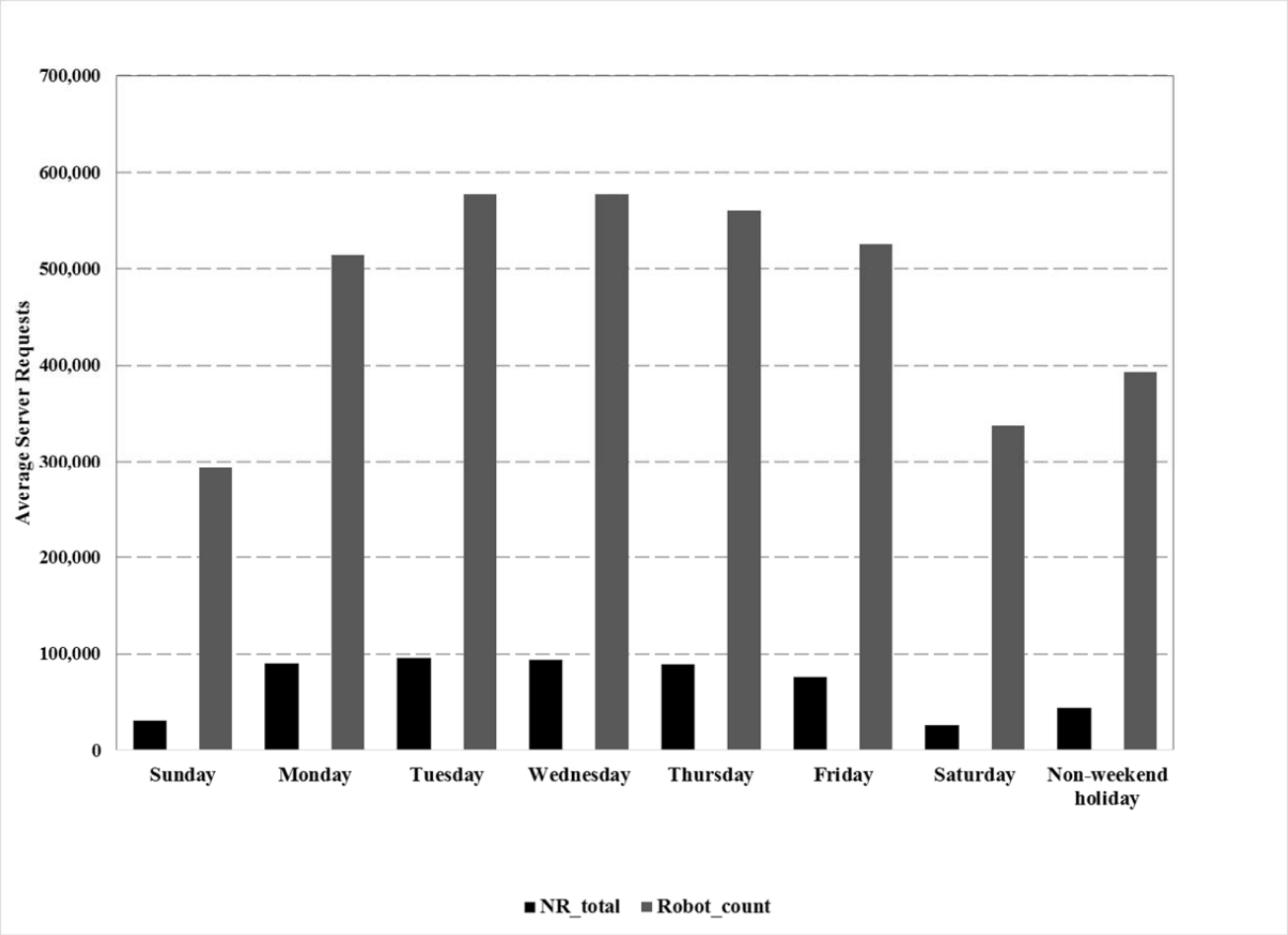

Figure 3. Non-robot counts (*NR_total*) and robot counts for server requests by day-of-week. "Non-weekend holiday" reports the average counts for all non-trading days not occurring on a Saturday or Sunday.

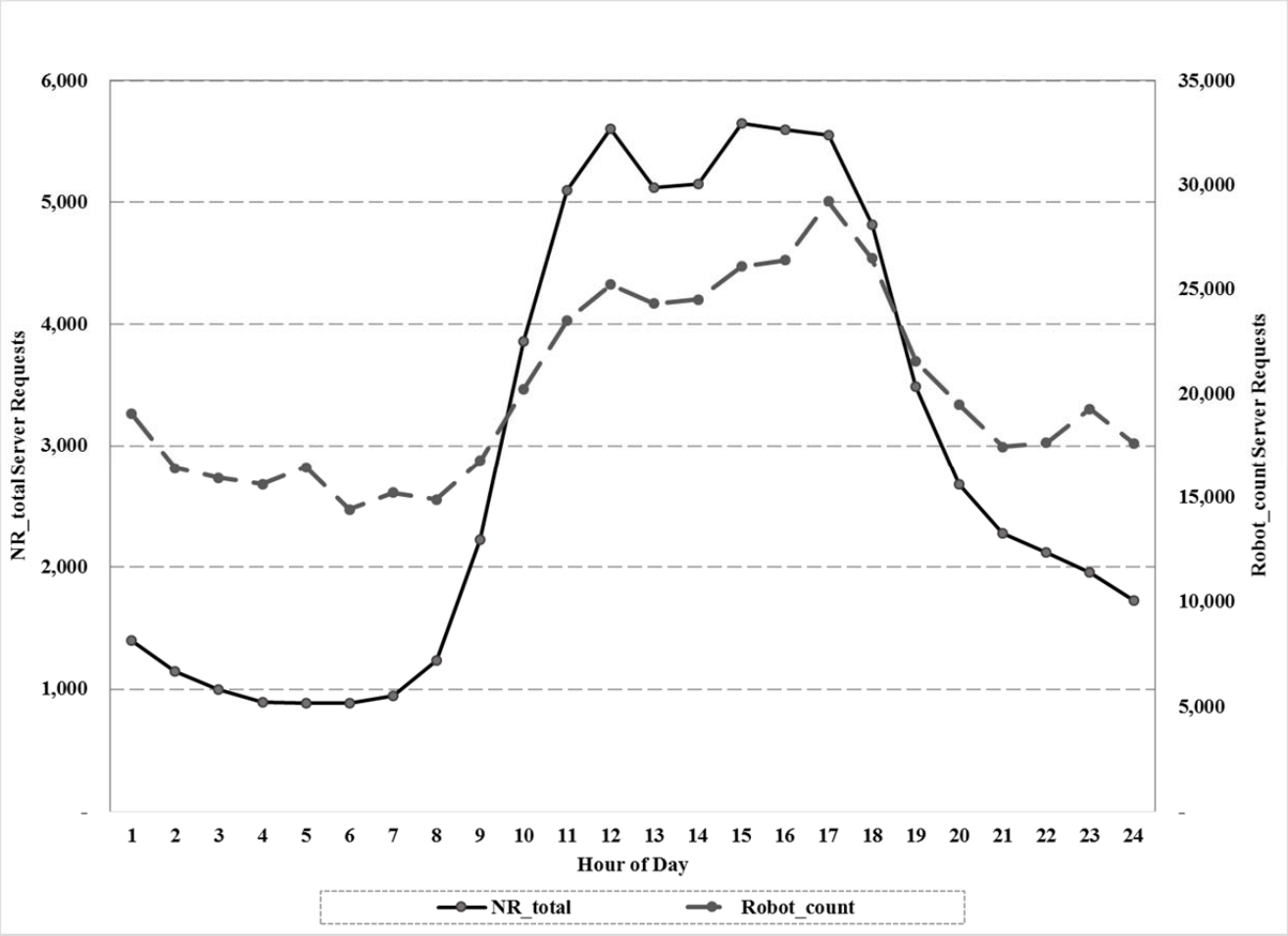

Figure 4. A plot of the non-robot counts (*NR_total*) and robot counts for server requests by hour-of-day. Note that *Robot_count* is associated with the right-hand-side axis scale. All reported times in the paper are US Eastern Standard Time.

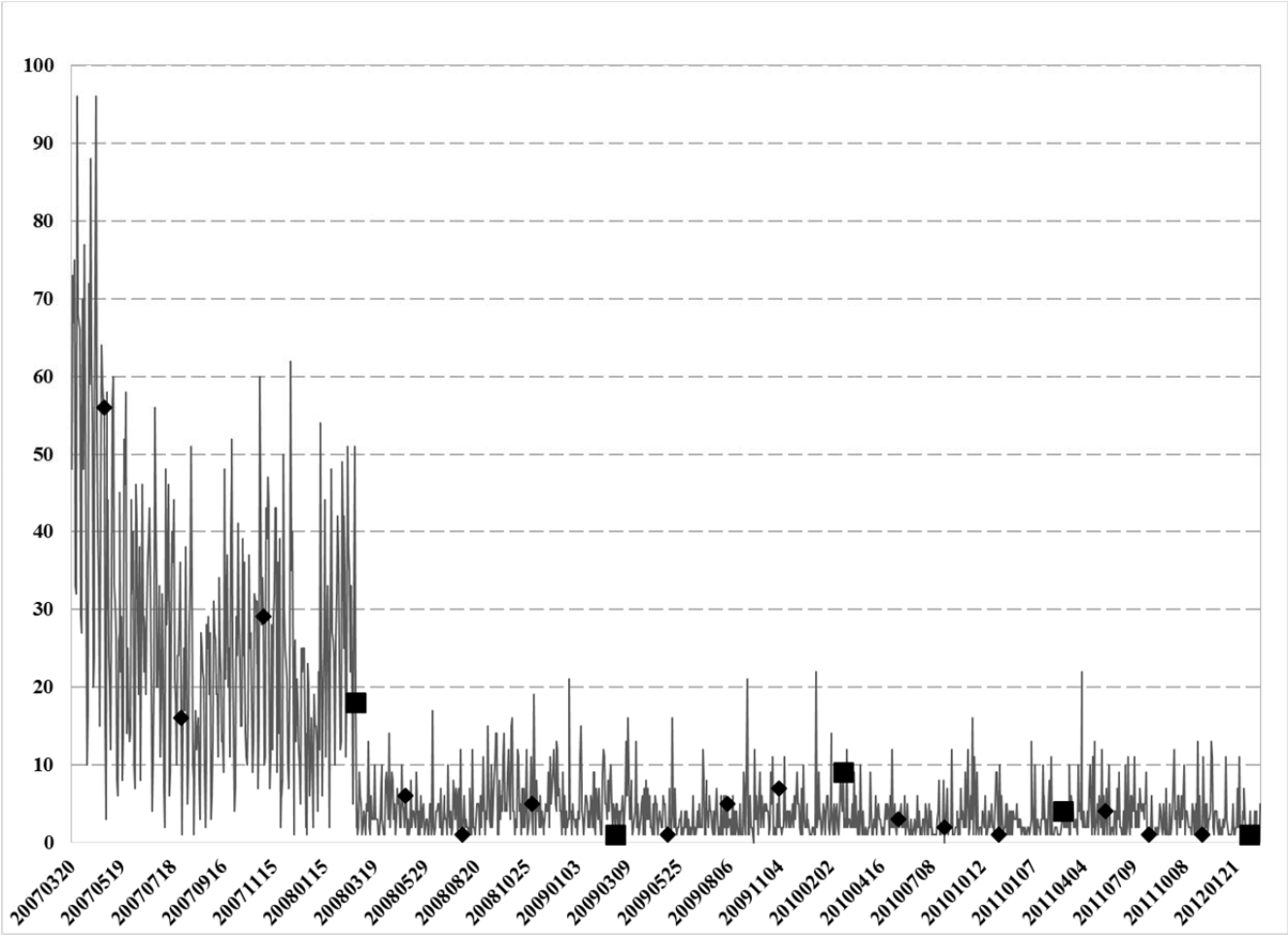

Figure 5. The non-robot count for HTM type server requests (*NR_HTM*) for General Electric's 10-K filing of February 2, 2007. The first 21 days (including the filing date) are excluded from the chart due to their larger numbers and an attempt to focus the scale of the chart on a longer post-filing period. The dates of subsequent 10-Q filings are denoted by a diamond and subsequent 10-K filings by a square.

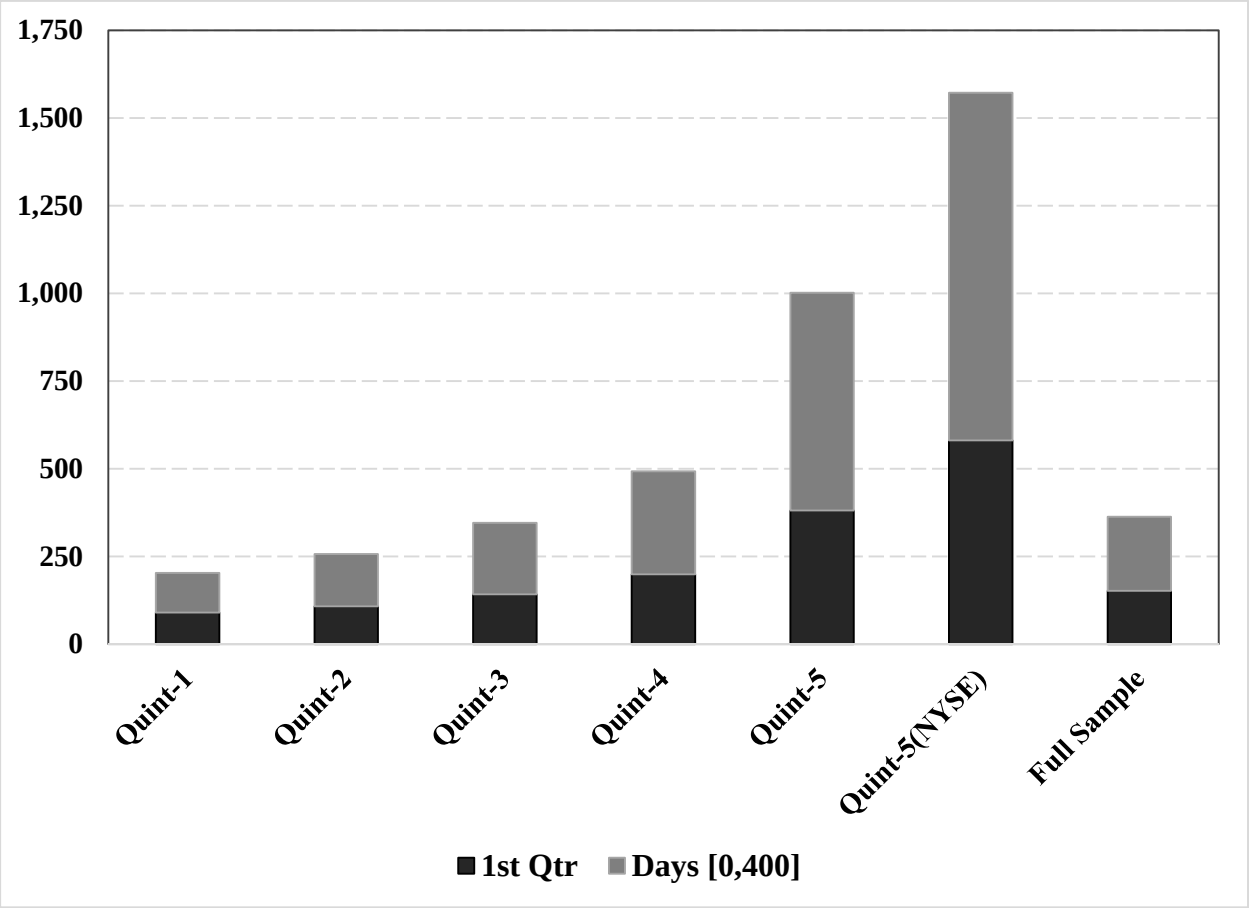

Figure 6. The median total number of *NR_HTM* downloads per 10-K filing by market capitalization quintile. The darker portion of the bar represents the first post-filing quarter median number of downloads for each category, with the full length of the bar representing the median total number of downloads for days [0-400]. The quintiles for Quint-1 through Quint-5 are estimated from the full sample. Quint-5(NYSE) is the largest quintile of those firms listed on the New York Stock Exchange. The last column summarizes the median total downloads across the full sample.

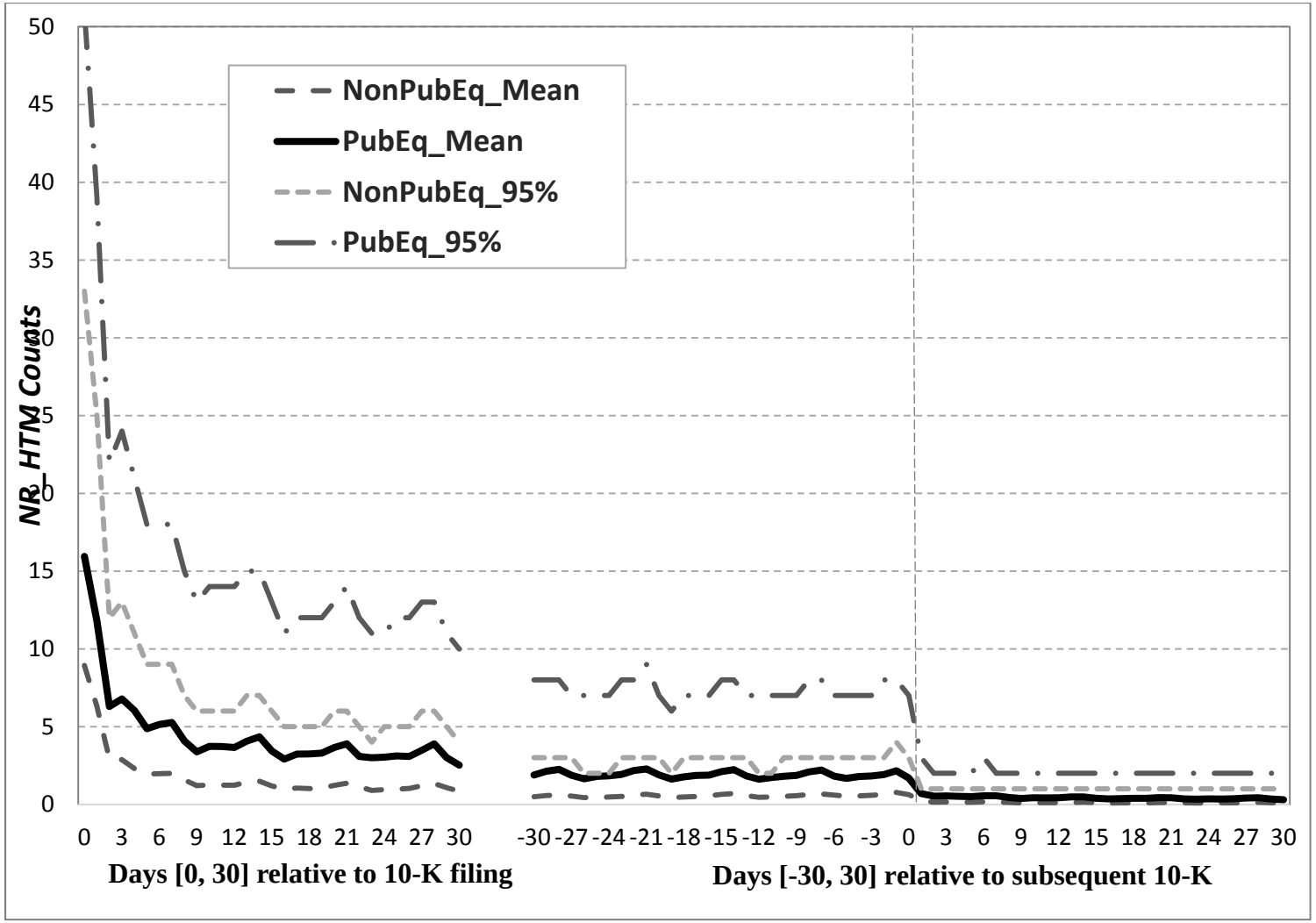

Figure 7. A plot of the 10-K *NR_HTM* count (non-robot htm file downloads) for firms with and without public equity. Days [0, 30] on the x-axis are relative to a given firm's 10-K filing date. Days [-30,30] are relative to the filing of the firm's subsequent 10-K.

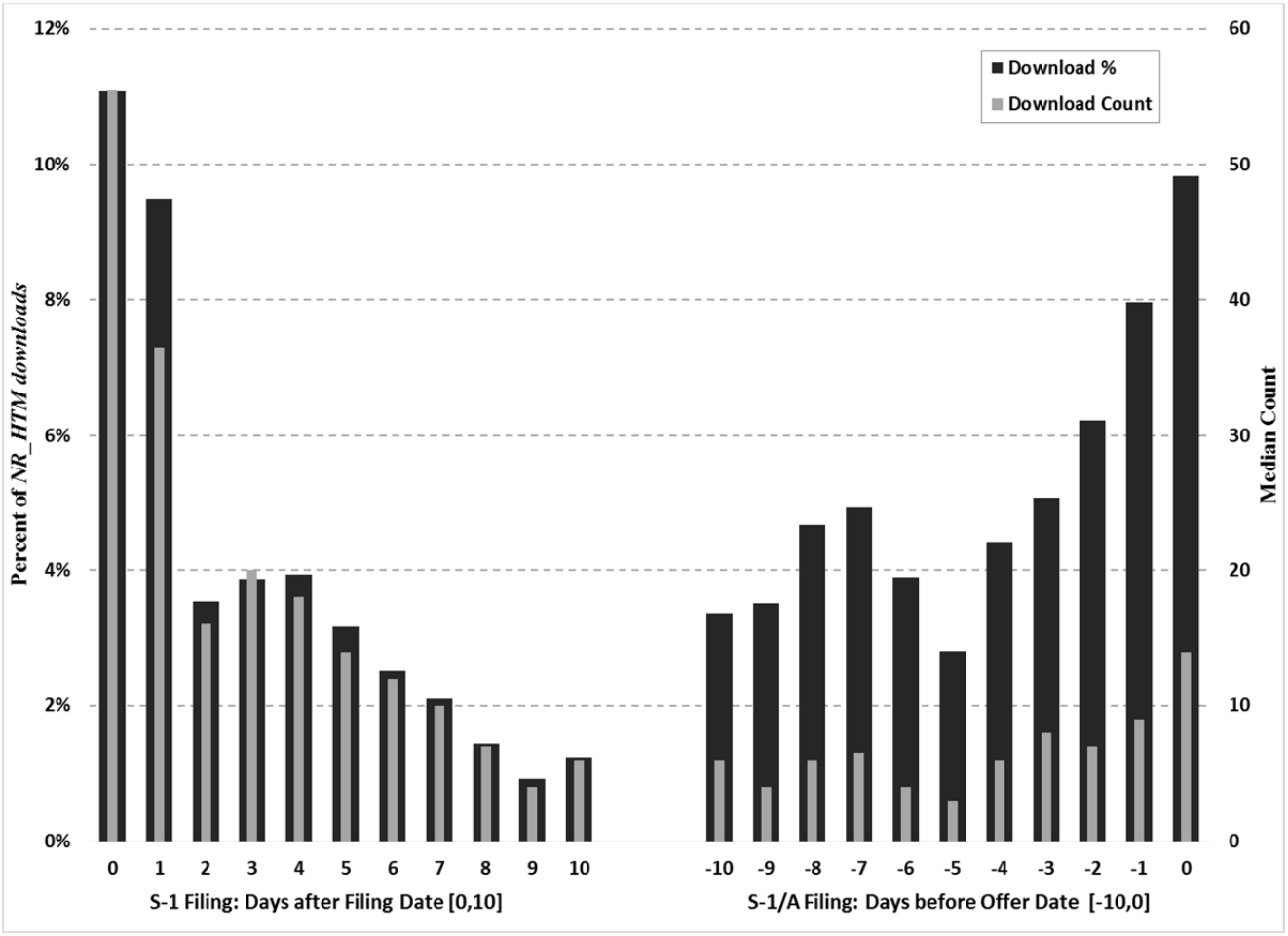

Figure 8. The percent of total *NR_HTM* downloads between filing date and offer date for S-1 days [0, 10] and S-1/A days [-10, 0] with the median count overlaid in grey. The first segment of the chart presents the percent of *NR_HTM* downloads for days [0, 10] relative to the filing date and total downloads for that form, with the median count overlaid in grey. The second segment presents the percent of *NR_HTM* downloads for the S-1/A filing occurring at least 10 days before the offer date, with the median count overlaid in grey.

Table I
EDGAR Filing Frequencies by Form Type

This table reports the form groupings used in the initial analysis. Total filings and percentages are based on all filings listed in the SEC EDGAR Master File Index from 2003-2012.

	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%

Table II

Distributional Statistics for EDGAR Server Requests for all Form Types

The 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 non-robot server request counts for HTM files and TXT files, respectively. *NR_total* is the count of each filing for all non-robot requests having at least one non-robot 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 one non-robot 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* <= 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.

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

Table III

Distribution of EDGAR File Requests within 400 days of Filing and Ratio of *NR_HTM* File Requests to Form Filings

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 I and form types are defined in Appendix C. The amended forms, with the exception of the 10-K/A and 10-QSB group, and Forms 3 and 5 are not reported here but are available in a complete table presented in the internet appendix.

	Percentage of Days [0,400] File Requests							Full Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Form Type	Day [0,1]	1st Week	1st Month	1st Qtr	2nd Qtr	4th Qtr	Total *NR_HTM* Count <= 400 days	Total *NR_HTM* File Requests	Total Filed 2003-2012	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

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

Columns (1) through (3) are Tobit regressions, where the dependent variable is the 10-K Log(1+*NR_HTM*) file downloads 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 one 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 one 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 A. Industry dummies are based on the Fama and French (1997) 48-industries. Column (4) is an OLS regression where the time-series is collapsed so that there is one observation for each 10-K filing and the dependent variable is the total file downloads (*NR_HTM*) for days [0,1].

	Days [0,400] sample			Days[0,1] sample
	Non-public equity sample	Public equity sample	Public equity sample	Public equity sample
Variable	(1)	(2)	(3)	(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%

Table V
Regressions with Subsequent Stock Return Volatility (RMSE) as the Dependent Variable for IPOs with Available Data during 2003-2012

All columns have root-mean-square-error (RMSE) as the dependent variable. RMSE is from a market model estimated using trading days [+5, +60] relative to the IPO date. For *NR_HTM*, columns (1) and (2) use the summation of EDGAR S-1 and S-1/A server requests from the S-1 filing date until five days prior to the IPO date. In columns (3) and (4), the summation for *NR_HTM* is in the window of +/- four days around the IPO date. See Appendix A for definitions of all other variables. Included are an intercept, Fama and French (1997) 48-industry dummies, and calendar year dummies. The *t*-statistics are in parentheses with the standard errors clustered by year and industry.

Sample	S-1 filing date to five days before offer date		Offer date +/- four days	
	(1)	(2)	(3)	(4)
Log(NR_HTM)		0.110		0.197
		(2.53)		(6.36)
Log(proceeds)	-0.198	-0.231	-0.195	-0.267
	(-3.83)	(-4.08)	(-3.75)	(-5.64)
Log(1+age)	-0.225	-0.211	-0.227	-0.220
	(-3.29)	(-3.13)	(-3.35)	(-3.47)
First-day returns	0.002	0.002	0.002	0.001
	(0.78)	(0.78)	(0.74)	(0.35)
VC dummy	0.064	0.041	0.071	0.033
	(0.83)	(0.57)	(0.95)	(0.50)
Top-tier dummy	-0.168	-0.158	-0.165	-0.162
	(-1.46)	(-1.42)	(-1.47)	(-1.33)
Positive EPS dummy	0.023	0.036	0.026	0.052
	(0.44)	(0.79)	(0.50)	(0.93)
Prior Nasdaq 15-day returns	-0.014	-0.014	-0.014	-0.010
	(-1.18)	(-1.20)	(-1.16)	(-0.86)
Up revision	0.005	0.004	0.005	0.003
	(0.85)	(0.79)	(0.83)	(0.50)
Industry dummies	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes
Constant	Yes	Yes	Yes	Yes
R^2	32.60%	33.11%	32.53%	33.88%
Sample Size	550	550	552	552

Internet Appendix

Table A.1
Distribution of EDGAR file requests within 400 days of filing and ratio of *NR_HTM* file requests to form filings

This table is the complete version of Table III from the manuscript. Columns (1) – (7) 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 (9) – (10) compare the number of form requests to the number of filings for the entire period. Form groups are descried in Table I and form types are defined in Appendix C.

	Percentage of Days [0,400] File Requests								Full Sample		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Form Type	Day [0,1]	1st Week	1st Month	1st Qtr	2nd Qtr	3rd Qtr	4th Qtr	Total *NR_HTM* Count <= 400 days	Total *NR_HTM File Requests*	Total Filed 2003-2012	Ratio of *NR_HTM* File Requests to Filings
10-K	4.7%	10.1%	21.6%	41.1%	60.0%	79.2%	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%	87.9%	98.4%	1,384,279	1,992,423	19,446	102.46
10KSB	10.4%	17.8%	31.1%	48.3%	65.7%	80.2%	96.9%	405,518	719,538	20,103	35.79
10KSB/A	12.5%	21.4%	36.0%	55.9%	74.8%	87.7%	97.9%	91,658	147,903	7,124	20.76
10-Q	8.1%	16.3%	33.3%	71.8%	88.0%	93.6%	98.7%	20,598,503	25,063,871	234,762	106.76
10-Q/A	15.3%	27.4%	47.9%	78.4%	90.3%	95.4%	99.0%	643,815	789,536	20,339	38.82
10QSB	10.7%	18.6%	34.2%	66.0%	85.0%	91.9%	98.5%	503,813	755,480	55,456	13.62
10QSB/A	14.2%	24.8%	42.9%	69.6%	84.7%	92.3%	98.4%	48,846	75,402	9,145	8.25
8-K	19.0%	34.0%	54.5%	75.4%	87.0%	93.5%	98.6%	23,657,516	30,210,880	892,420	33.85
8-K/A	17.6%	32.9%	53.2%	73.9%	86.1%	93.3%	98.5%	834,218	1,103,829	38,211	28.89
S-1	29.5%	47.8%	62.9%	78.0%	89.2%	95.0%	98.9%	2,883,443	3,579,307	10,604	337.54

S-1/A	20.4%	39.6%	62.3%	78.4%	88.5%	94.6%	98.8%	2,681,703	3,447,224	25,486	135.26
424	19.7%	33.4%	49.7%	68.5%	82.7%	91.7%	98.1%	3,876,358	5,778,612	280,755	20.58
3	25.5%	40.3%	56.4%	72.7%	84.4%	91.9%	98.1%	26,180	43,379	387,194	0.11
3/A	21.4%	35.3%	51.5%	68.6%	82.0%	88.2%	97.6%	1,479	2,446	26,019	0.09
4	25.5%	41.3%	58.3%	74.1%	84.7%	91.1%	98.2%	68,278	137,013	4,269,940	0.03
4/A	17.9%	35.1%	52.1%	67.3%	80.3%	87.0%	97.1%	3,512	7,127	155,404	0.05
5	16.6%	26.5%	43.4%	60.4%	72.7%	82.8%	96.7%	1,618	11,818	129,570	0.09
5/A	10.6%	21.7%	44.9%	58.1%	71.7%	80.8%	99.0%	198	532	5,819	0.09
SC 13D	25.9%	42.6%	59.0%	76.1%	87.1%	93.8%	98.5%	513,312	668,144	45,684	14.63
SC 13D/A	26.3%	42.8%	60.8%	77.9%	88.6%	94.8%	98.8%	1,020,036	1,237,687	97,546	12.69
SC 13G	18.8%	34.4%	54.5%	73.8%	85.4%	92.6%	98.5%	504,293	629,384	167,712	3.75
SC 13G/A	16.2%	31.0%	51.7%	71.1%	82.5%	90.4%	98.0%	670,540	837,954	288,933	2.90
13F-HR	16.7%	33.3%	50.0%	92.9%	100.0%	100.0%	100.0%	42	152	110,364	0.00
DEF 14A	6.8%	15.9%	30.8%	47.3%	64.5%	81.1%	98.0%	5,452,271	7,807,698	69,007	113.14
497	17.9%	30.4%	46.9%	67.8%	83.5%	92.9%	98.9%	546,872	655,903	180,339	3.64
6-K	12.9%	27.3%	50.3%	74.6%	87.4%	93.7%	98.7%	3,360,208	4,028,823	229,464	17.56
6-K/A	15.9%	29.9%	50.4%	73.5%	86.9%	93.3%	98.6%	34,459	42,802	3,086	13.87
Other	15.7%	29.8%	47.1%	65.1%	79.9%	90.0%	98.3%	17,484,283	23,672,056	3,358,829	7.05

Table A.2
Top 25 server downloads for non-robot 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

Table A.3
Top 25 server downloads by firm and day for *Robot_count*, where *NR_HTM* is greater than zero

	Company Name	CIK	Form Type	Filing Date	Server Date	Robot count
1	Superconductor Technologies	895665	4	20100507	20100508	121,340
2	Superconductor Technologies	895665	4	20100507	20100509	120,332
3	Superconductor Technologies	895665	4	20100507	20100510	108,706
4	ClubCorp Club Operations	1515382	S-4	20110328	20110724	59,705
5	Allied Waste Industries	848865	S-4	20100506	20100706	55,241
6	Abax Upland Fund	1464111	4	20091110	20091111	49,194
7	American Medical Response	888675	S-4	20110927	20110927	48,993
8	Mana Holdings	1456822	4	20110715	20110716	44,029
9	Telenetics Corp	810018	4	20050204	20101227	38,811
10	Medical Properties Trust	1287865	S-4	20111006	20111006	35,483
11	Caesars New Jersey	276310	S-4	20081029	20100416	34,997
12	Province Healthcare	1044942	S-4	20110506	20110510	34,691
13	Allie Waste Industries	848865	S-4	20100506	20110731	33,840
14	Warburg Pincus X Partners	1451560	4	20120110	20120111	33,582
15	Manna Holdings	1456822	4	20110715	20110717	32,523
16	OCM Spirit Holdings II	1521309	4	20110603	20110604	21,774
17	Burlington Coat Factory	718916	S-4	20061010	20100416	31,475
18	Spirit Airlines	1498710	4	20110603	20110605	30,577
19	Telenetics Corp	810018	4	20050204	20101227	30,172
20	Telenetics Corp	810018	4	20060214	20101227	29,973
21	Facebook Inc	1326801	S-1	20120201	20120201	29,784
22	OCM Princ. Opportunities Fund IV	1394232	4	20091218	20091219	29,552
23	OCM Crimson Holdings	1478590	4	20091218	20091220	28,965
24	Gotham Partners Intl	1178174	4	20030702	20030703	27,187
25	Bessemer Venture Partners VI	1362891	3	20110603	20110604	25,765