Using the EDGAR Log File Data Set

James P. Ryans[1,*]

London Business School

Abstract

The SEC's EDGAR log file data set is a collection of web server log files that allow researchers to study the demand for SEC filings. This multiple terabyte data set provides researchers with a direct measure of demand for financial reports, but the log files must be filtered to remove downloads by computer programs (or *robots*), and the sheer size of the files presents big data challenges. This paper compares three methods for counting human views in the EDGAR log files and aggregates the data on a filing-day basis so that it is accessible to desktop hardware and statistical analysis software. Overall, the three methods agree on the robot-human classification for 96 percent of users, but for sample 10-K filings, they can disagree by up to 27 percent. Download counts may be biased by up to 36 percent if multiple views by the same user are counted. Ryans's (2017) method eliminates multiple download counting and appears to effectively classify robots in cases of disagreement among the measures. The choice of measure may be particularly important when studying demand for Forms 10-K, 10-Q, 4, 13F-HR, as well as SEC comment letters. The aggregated data and sample code are available from the author.

Keywords: EDGAR downloads, SEC filings, demand for financial information, investor attention, big data

JEL: M41, G14, G18

[*]The summarized EDGAR log files discussed in this paper and accompanying R code are available for noncommercial use at `http://www.jamesryans.com`.

Email address: `jryans@london.edu` (James P. Ryans)

[1]I thank Alastair Lawrence and Teri Lombardi Yohn for useful comments. I also thank the SEC for providing early versions of the EDGAR log file data by Freedom of Information Act requests, and for making updates readily accessible in recent years.

February 16, 2017

1. Introduction

This document describes the SEC's EDGAR log file data set, which records internet search traffic for EDGAR filings, and compares three methods used in the literature for summarizing this data on a filing-day basis to facilitate analysis. The accounting and finance literature has used the EDGAR logs to study the demand for financial reporting information (e.g., Drake, Roulstone, and Thornock, 2015; Dechow, Lawrence, and Ryans, 2016; Loughran and McDonald, 2016; Bozanic, Hoopes, Thornock, and Williams, 2016; Ryans, 2017). While it is possible for investors to access SEC filings from other sources, such as a firm's investor relations web site, it appears that EDGAR captures a significant fraction of financial disclosure demand. Monga and Chasan (2015) quote General Electric CFO Jeffrey Bornstein, who noted that GE's 2013 annual report was downloaded from their investor relations web site just 800 times. For the same annual report, the EDGAR logs record 21,987 (4,325) downloads in the year (two months) following its filing. Some firms, such as Google (Alphabet, Inc), forward investors directly to the EDGAR web site to obtain their SEC filings, and as a result, EDGAR may capture an even greater fraction of filing views for some firms. Although the volume of EDGAR downloads for GE's 10-K appear significant, it represents a relatively small number in comparison to GE's two million individual shareholders (Monga and Chasan, 2015). A caveat to EDGAR download data is that investors have other avenues for obtaining these filings in addition to investor relations web sites. For example, EDGARonline, Bloomberg, and FactSet provide access to these reports, and we cannot currently observe download statistics from these sources.

The SEC provides access to the EDGAR log files on their web site, with a

delay of approximately one year.[2] The data is challenging to manipulate into usable form, as it currently consists of 4,839 daily files containing a record of each EDGAR download. Each daily log file is large: the log file for a single day's downloads on March 31, 2016 contains approximately 14 million records in a 1.6 gigabyte file. Clearly, working with more than a few days' data in-memory is not feasible, so by definition researchers are presented with a big data challenge when analyzing the log files. However, when aggregated to the level of one record per filing-day, the dataset is reduced to a total of 374 million observations from 2003 to 2016, and it becomes accessible to researchers with typical desktop hardware and statistical analysis software.

Prior literature studies aspects of demand for financial disclosures in different settings and using a variety of proxies, such as index membership (e.g., Shleifer, 1986; Chen, Noronha, and Singal, 2004), shareholder composition (e.g., Lehavy and Sloan, 2008; Bushee, Core, Guay, and Hamm, 2010), and analyst following (e.g., Irvine, 2003). Demand for financial information is also associated with the concept of investor attention, as investor attention is likely to lead to consumption of firm disclosures. Studies of investor attention consider news media, trading volume and extreme returns (e.g., Gervais, Kaniel, and Mingelgrin, 2001; Barber and Odean, 2008; Engelberg and Parsons, 2011). DellaVigna and Pollet (2009) study a setting where investor demand is lower due to inattention to Friday earnings announcements. Recently, more direct proxies of investor demand for information include studies of Google search volume for stock ticker symbols (e.g., Da, Engelberg, and Gao, 2011; Drake, Roulstone, and Thornock, 2012), and Yahoo Finance page views (Lawrence, Ryans, and Sun, 2016, 2017).

[2] `https://www.sec.gov/data/edgar-log-file-data-set`

EDGAR download logs allow for the study of demand for financial disclosures filed with the SEC. This paper compares the EDGAR download count methods used in Ryans (2017, hereinafter Ryans), Loughran and McDonald (2016, hereinafter LM), and Drake et al. (2015, hereinafter DRT). DRT describe factors associated with demand for EDGAR filings, and find that higher demand is associated with the speed at which earnings news is incorporated into returns. Dechow et al. (2016) show that SEC comment letters are downloaded at low rates compared to the 10-K reports on which they comment, and that there is a stronger negative price response when SEC comment letters are disclosed, if those comment letters are more heavily downloaded. Lee, Ma, and Wang (2015) create peer groups of firms by identifying those that are searched on EDGAR in close temporal proximity by individual users. LM note that financial disclosures are in general little used as a source of information. Masden (2017) finds that pre-earnings announcement returns are affected by attention to customers' downloads of EDGAR filings. Bozanic et al. (2016) use demand from IRS users to infer attention by tax authorities.

Researchers intending to study investor attention need to screen out any programmatic, or "robot", downloads so that the remaining observed page views correspond to human readers.[3] The EDGAR log files record the network (IP) address of each user downloading a document. The general robot-screening procedure is to calculate statistics about a user's download patterns over a day, then apply one or more tests to classify the user as a robot or a human. DRT developed the first method used in the literature to filter the EDGAR log files, though LM report that DRT's measure may provide misleading inferences due to misspecified robot-

[3]The terms *page views* and *downloads* are used interchangeably

classification thresholds. In particular, LM observe very few Form 4 downloads, in contrast to DRT's quite high rate of Form 4 downloads. Ryans (2017), hereinafter Ryans, studies investor attention to SEC comment letters, and uses a modified metric which is based on the contributions of DRT and LM, combined with a detailed analysis of actual user download patterns from the EDGAR log files. This paper compares the download methods used in DRT, LM, and Ryans, and provides examples of detailed log file analysis, illustrating that Ryans's method is effective at classifying readers as robots or humans in cases where the three measures disagree. Filings that are available in Interactive Data format, or that contain a number of exhibits (e.g., 10-Ks and 10-Qs), appear prone to upwardly biased download counts when using the procedures described in DRT and LM, but not when using Ryans's procedure, which only counts one download per user for cases where the EDGAR log records multiple page views for the same document in the same day.

In Section 2, I describe the methodologies used to calculate download counts according to LM, DRT, and Ryans. Section 3 provides a comparison of the three measures. Section 4 concludes.

2. EDGAR Log Data

The raw EDGAR log files contain a record for each filing download request,[4] and I begin by aggregating this file at the filing-day level. Since a filing is associated with a unique firm by its CIK number, this is equivalent to firm-filing-day aggregation. Each record in the aggregated file contains the following fields:

[4]The detailed log file record elements are described at `https://www.sec.gov/files/EDGAR_variables_FINAL.pdf`

1. *date*. YYYY-MM-DD format, Eastern Standard Time zone.
2. *cik*. The SEC's Central Index Key value associated with a filer (firm identifier).
3. *accession*. The SEC Accession Number, a document identifier (filing identifier).
4. *DRTpv*. Page views summarized according to the procedure in DRT.
5. *LMpv*. Page views summarized according to LM.
6. *Rpv*. Page views summarized according to Ryans.

Data files are provided for each calendar year beginning January 1, 2003, and currently ending March 31, 2016.[5] There are significant gaps in the data prior to March, 2003, and between September 24, 2005 and May 10, 2006, due to lost or corrupted log files. Additional data about each filing referenced in the log file can be identified by merging this data with the SEC EDGAR index file.[6]

2.1. Filtering Robot Downloads

In order to observe investors' use of financial information, downloads made by robots should be disregarded. Robots are either software programs that download large volumes of filings, or they may be smaller-volume software tools used to automate the download of targeted filings for later analysis. Because of the automated nature of the download, it is unlikely that a human is consuming the financial information immediately, and therefore is less likely to act upon the download. It is the consumption of financial information by an economic actor that is the construct of interest to most researchers.

[5] Available for academic use from the author.

[6] Further details about the EDGAR index files can be found at `https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm`

The initial task when summarizing the EDGAR download logs is therefore to screen out robot requests, and this is accomplished by first classifying IP addresses–the users' network address—as representing a robot if it meets one or more conditions. If it is not classified as a robot, the IP address is assumed to represent a human reader.

2.2. Classification Errors

There are two types of classification errors, based on the null hypothesis that a user is not a robot (i.e., a user is a human): Incorrectly classifying a human IP address as a robot is a Type I error, and incorrectly classifying a robot IP address as a human is a Type II error. Such errors can not always be precisely identified, for at least three reasons: First we cannot be sure, in some cases, of the type of user behind an IP addresses, based only on analysis of logs of the IP's activity. Is the user a sophisticated analyst who reads possibly hundreds of documents a day, or is it a small-scale robot that is populating a modest database? While an IP that downloads tens of thousands of filings a day is almost certainly a robot, an IP addresses that downloads only 10 filings may also be a robot, and analysis of the download patterns and type of content accessed will be the only clues available. Second, a single IP address could be used for both manual and automated downloads, either at different times during a day or at the same time if the robot program runs as a background process. In an ideal world, we would keep the manual download records and discard the robot activity, but in practice it may be impossible to separate the two activities from the log file evidence alone. Third, researchers may simply disagree on a common concept of a robot. If a user downloads 10 different companies' annual reports by viewing them in a web browser over a period of an hour, then the situation clearly represents human page views.

However, if a user runs a script to download the same set of annual reports in less than a minute, should we now treat those page views as being robot-generated? Intuitively, most researchers would classify the latter case as a robot, but some may disagree. A download count method that correctly classifies these borderline cases needs parameters that can make the appropriate distinctions.

Applying their respective screening parameters, all three of Ryans, DRT, and LM would keep the 10-view web browser user, but the scripting user would result in disagreement. Ryans would likely classify the user as a robot for violating a 3-CIK-per-minute limit, and DRT would eliminate the scripted downloads for violating a 5-download-per-minute limit. However, LM would keep the scripted downloads as human because the user made fewer than 50 total downloads during the day.

Depending on the research question, classification errors may be relatively unimportant. For example, the primary research question addressed by Lee et al. (2015) is to identify peer firms via clusters of companies that are searched together by individual users. If some humans are classified as robots and disregarded, it should not affect their inferences. Even if some low-volume robots are included as humans, their co-search patterns may still reflect peer groupings, because a human may have directed the robot on which filings to gather. On the other hand, if a researcher wants an accurate estimate of the number of people who viewed a particular filing on a particular day, the method chosen could take on significant importance.

Next, I discuss the methods used in Ryans, DRT, and LM. I first focus on LM because their method is the simplest, having a single classification threshold. I then discuss DRT, because their procedure involves a two-step test. Finally I

discuss Ryans, which uses a three-step procedure designed to achieve a reduction in both Type I and Type II errors relative to LM and DRT, by testing for patterns characteristic of both high-volume and low-volume robots, while aiming to retain download counts associated with sophisticated human users.

2.3. LMpv Robot Screening Procedure

LMpv is the sum of human downloads, calculated after screening out IP address with more than 50 downloads during a day as belonging to a robot. This is a more general version of the procedure used in Lee et al. (2015) who are only interested in searches to infer the identity of peer groups. As a result, Lee et al.'s (2015) research design includes a number of filtering steps that may be appropriate for identifying peer firms, but which are not appropriate in a general setting.[7] LM's assumptions of robot download behavior can be stated as a single rule:

1. Humans do not download more than 50 items during a day.

This is a simple and clear threshold that is likely to be broadly true. IP addresses that download significantly more than 50 items during a day are more likely to be robots, and IP addresses that download fewer than 50 items during a day are more likely to be human. For broad tests of investor attention to a variety of filing, download counts based on this criterion may be adequate. Two issues may be of concern in some settings: First, the procedure counts all downloads of a filing, even if they represent repeat downloads from the same IP address. Second,

[7] Lee et al. (2015) only consider S&P 1500 firms; IP addresses that search for two or more firms in a day; IP addresses that download fewer than 50 documents; and downloads of Forms 10-K, 10-Q, 8-K, S-1, and 14A. They readily admit (p. 415) that the 50-download cutoff for robots is "conservative", i.e. has a high rate of Type I error, but in their setting, a high rate of Type I error likely has no impact on inferences around human co-search patterns.

errors where humans are mistakenly classified as robots will be for those humans who download larger numbers of filings, and as a result $LMpv$ may bias against counting the page views of sophisticated financial statement readers. In Ryans, who studies investor attention to SEC comment letters, these relatively complex filings are more likely to be viewed by sophisticated analysts, making the $LMpv$ measure potentially unsuitable.

2.4. DRTpv Robot Screening Procedure

$DRTpv$ is calculated by classifying an IP address as belonging to a robot based on either the total number of requests during a day, or the maximum number of requests per minute during a day. Using their specific thresholds, their assumptions can be stated as:

1. Humans do not download more than 1,000 items during a day; *and*
2. Humans do not download more than 5 items per minute.

The high threshold on the first step allows sophisticated humans to be retained, at the expense of not rejecting some robots. The second step designed to catch those robots that pass step one, by identifying their more rapid frequency of downloads. This two step process should allow both for sophisticated human users to be retained while also eliminating low-volume robots. LM claim that the DRT procedure leads to classification errors, in particular those of Type II. If 1,001 requests in a day are likely to be the result of robot activity, would not a user making 999 requests during a day also be a robot mis-classified as a human? This criticism is tempered with the two-step procedure, as any robot making it past the first step is still subject to the 5 item per minute second step. In manual inspection of download patterns, I find that DRT's 5-item-per-minute threshold is probably

too low given EDGAR filing structures that have come into place in recent years. Therefore, DRT may inadvertently generate Type I errors via their second step while attempting to correct for Type II errors generated by their first step. The misclassification becomes more prevalent in recent years, as "Interactive Data" versions of forms such as 10-Ks and 10-Qs are presented in a way that human users can rapidly click through a series of footnotes, with each click registering as a download in the log files, increasing the chances that a human violates the five item per minute limit.

2.5. Rpv Robot Screening Procedure

Ryans augments and modifies these methods to try and reduce both Type I and II errors, based on detailed analysis of activity patterns by individual IP addresses in the detailed log files. Ryans adds a third step to DRT, and modifies the thresholds to levels characteristic of apparent human browsing patterns. Ryans's three step procedure examines downloads by each IP address, each day, calculating: (1) total number of downloads in a day, (2) maximum number of downloads in a one minute period, and (3) maximum number of separate company filings accessed in a single minute. As with DRT, the benefit of the multi-factor screen is that thresholds on each factor can be kept relatively high, reducing Type I errors, while testing for different characteristics of robot behavior to reduce Type II errors. *Rpv* is calculated based upon the following general assumptions and thresholds:

1. Humans do not download more than 25 items in a single minute; *and*
2. Humans do not download more than 3 different companies' items in a single minute; *and*

3. Humans do not download more than 500 items in a single day.

Individually, the thresholds may seem relatively high, for example 25 items per minute, or 500 items during a day. However, this threshold was set after observing user behavior in the log files, especially as it relates to viewing patterns of Interactive Data browsing of 10-K and 10-Q reports, which arose after the adoption of XBRL filings in 2009. When a user views the Interactive Data version of the filing, every click on a different footnote or financial statement generates another page view, and it is not uncommon for there to be 50 or even 100 items associated with an interactive 10-K.

Figure 1 illustrates an example index page for a 10-K, with a variety of exhibits or other items on which a user can click and generate separate download records in the EDGAR log. The Interactive Data button in the header section allows the user to access a presentation of the filing contents generated from the company's XBRL filing, and Figure 2 illustrates the Interactive Data presentation of an annual report. The income statement is visible in the main body of the page, and the menu to the left allows the user to quickly click on different financial statement and notes items, with each click registering in the EDGAR log file as a separate download. When faced with the pattern of downloads that result from viewing a single 10-K filing in Interactive Data format, DRT have a reasonable chance of classifying a human as a robot as it is easy to now access more than 5 footnotes in a minute. LM have a reasonable chance of classifying a human as a robot when the user accesses more than 50 footnotes during the day.

A key insight used in Ryans is that it is difficult for a human to access filings for different firms in quick succession, as the company search feature is somewhat slow to execute, requiring the user to navigate to the search page, enter an

identifier, and possibly select the relevant filer from a list. For a robot, the separate company filings are requested without needing to search, and so a robot downloading a set of documents is likely to access different CIKs in sub-minute intervals. Thus, if filings for more than 3 different companies are downloaded in a one-minute period, it is likely that the downloads are being conducted by a program, even if the automation is for a small total number of filings.

2.6. Other Issues: HTML-only Filings and Multiple Views by the Same IP Address

LM discusses results for measures using both total human downloads and HTML-only human downloads. For general research settings, limiting an analysis to HTML files is likely too restrictive. For example, HTML only page views ignores the human downloads related to Interactive Data viewing of 10-Ks and 10-Qs, because these items record XML-type views. Table 1 illustrates downloads for different forms according to each measure, and identifies certain forms for which LM's HTML-only measure ($LMpvH$) dramatically undercounts page views. In particular, Forms D, 13F-HR, and 4 show counts of zero or near-zero, while download counts of these forms by other measures consistently show a large number of page views (e.g., 8 downloads for Form 4s by LM's HTML measure vs. approximately 7,000 to 11,000 for the other measures in Panel A). These forms are available as human-readable XML types when viewed with a browser. Some Form 4s are surely of interest to humans: Facebook's third most downloaded filing, according to Rpv and $DRTpv$, is a Form 4 filed on May 22, 2012, reporting that founder and CEO Mark Zuckerberg sold $1.1 billion worth of shares.[8] Facebook has filed many Form 4s, yet only this one achieved such high levels

[8] `https://www.sec.gov/Archives/edgar/data/1326801/000120919112029812/xslF345X03/doc4.xml`

of demand, with 44,280 human page views on the day following its disclosure (untabulated). If robots are responsible for Form 4 downloads, then there would be little reason for this particular Form 4 to generate so much interest, as robots would presumably be gathering all Form 4s with relatively equal volume.

An HTML-only measure would also be poorly suited to the analysis of attention to comment letters in Ryans, as these filings (Form UPLOAD) are posted as PDF files. A clue as to the misspecification of an $LMpvH$ measure is the fact that CORRESP filings—which are HTML-formatted company responses to comment letters—are quite widely viewed, and it is intuitively incongruous that investors quite heavily consume CORRESP filings, but not the related UPLOADs. For example $LMpvH$ counts 968 views of Form CORRESP in Table 1 Panel A, but 0 views of form UPLOAD. In contrast Rpv counts 1,426 views of Form CORRESP and 1,265 views of form UPLOAD, which is more similar in magnitude.

Researchers should also consider if their download measure should count multiple views of a single filing by the same IP address during a day. Depending on the research design, it may not be appropriate to count these views as separate consumption events: one user making 50 clicks on an interactive 10-K's items is clearly not the same magnitude of consumption as are 50 views of the 10-K by different users. Both DRT and LM count all views of a filing by an IP address on a given day as separate downloads, whereas Ryans counts all such views as one.

Multiple views of the same document are recorded in three situations: (1) when a user views the Interactive Data version of a filing, (2) when a user views a number of exhibits attached to a filing, and (3) when a user returns to the document periodically throughout the day, e.g., by browsing back and forth between an 8-K and a 10-Q. Interactive Data is a presentation of a company's XBRL filing such

that the EDGAR server renders the financial statements and notes in the user's web browser from the underlying XBRL data. Each click on a financial statement or note by a user registers in the log file as an additional download item.

This discussion also illustrates that although it is relatively easy to generate multiple downloads for a single firm's filings in a short period of time, it is more difficult for a human to access another firm's filings, as they first have to return to the SEC home page, whence they may search for another firm's filings. This site structure allows Ryans's 3-CIK per minute threshold to be effective at identifying low volume robots because such robots can download different firms' filings without the search delay encountered by a human.

2.7. Detailed Measure Calculation Procedure

To summarize the daily EDGAR logs, the first step is to obtain the raw log files from the SEC web site. As of January, 2017, there were 4,839 daily files available (some containing no records), from January 1, 2003 through March 31, 2016. It is also useful to match the log file to the the EDGAR index file, to have ready access to the filing type and other information, such as the form's filing date. For each daily EDGAR log file, I perform the following procedure:[9]

1. Keep records with $code = 200$ — successful delivery of requested document by the EDGAR server.
2. Keep records with $idx = 0$ — remove index observations, as index pages provide the viewer a link to a filing, not the filing itself. Figure 1 is an example of an index page.

[9]The data files available from the author contain sample R code which implements the summarization procedure.

3. Keep records with $crawler = 0$ — remove identified web cawlers, e.g., Google's indexing robot.
4. Make a list of IP addresses to exclude from Rpv: IPs with any downloads per minute greater than 25, or with number of CIK's downloaded per minute greater than 3, or with more than 500 downloads during the day.
5. Calculate Rpv views per filing: number of unique IP addresses downloading the filing, where the IP address is not in the Rpv excluded list.
6. Make a list of IP addresses to exclude from $DRTpv$: IPs with any downloads per minute greater than 5, or IPs with more than 1,000 downloads per day.
7. Calculate $DRTpv$ per filing: number of log records for the filing, where the IP address is not in the $DRTpv$ excluded list.
8. Make a list of IP addresses excluded from $LMpv$: IPs with more than 50 downloads during the day.
9. Calculate $LMpv$ per filing: number of log records for the filing, where the IP address is not in the $LMpv$ excluded list.

I do not report robot page view statistics. While LM perform some analyses using robot page views, their robot page views are only recorded for filings also viewed by humans on that day. Researchers studying robot download patterns would likely not want to condition their observations on a filing also being download by humans. Furthermore, a data set that includes all filing-days accessed by robots is approximately ten times the size of the human-only data set.

3. Comparison of Rpv, $DRTpv$, and $LMpv$

To understand the sources of differences between the three methods, I conduct analyses of the detailed log files to compare the methods used in Ryans, LM,

and DRT. The first case considers broad patterns of filings viewed by users on the peak day for downloads of Facebook's S-1, February 2, 2012, which is currently the most downloaded filing in the EDGAR database. This day is primarily chosen because LM discuss their download counts for this filing, allowing me to verify that $LMpv$ is calculated accurately. More broadly, this case allows me to examine how the different methodologies handle counts for high-demand form types, showing general agreement but some surprising variation among the three measures in specific settings.

The second case study focuses on the disclosure date for Google's 2010 10-K filing, February 11, 2011. This case is chosen because Form 10-Ks are considered to be the most comprehensive financial disclosure and are widely studied in the accounting and finance literature. 10-Ks are also complex filings, as they are presented in different formats: in HTML, as an Interactive Data page, as machine readable XML documents, and are frequently accompanied by a variety of exhibit documents. In this setting, it is important to be aware of how the count methods reflect this complexity, whether they might tend to omit page views in some situations, and whether the count reflects bias that may occur by recording views of different components of the 10-K as separate downloads. I also examine the extent to which the methods in DRT, LM, and Ryans disagree on the robot-human categorization of specific IP addresses, and drill down to look at the traffic patterns of these IP addresses to see which method makes fewer apparent classification errors.

Table 1 provides some initial evidence that the threshold values in DRT and LM may be set to levels that screen out more sophisticated financial statement readers. Rpv gives relatively low page counts compared to $DRTpv$ and $LMpv$ for

10-Ks, but Rpv shows relatively high counts for UPLOAD and CORRESP filings, which may be consumed by more sophisticated users whose browsing patterns are more likely to breach the 50 download per day limit for $LMpv$ and the 5 download per minute limit for $DRTpv$.

3.1. Download Activity Surrounding Facebook's S-1 Filing

Table 2 Panel A provides descriptive statistics for page views according to the measures calculated in Ryans, DRT, and LM. These metrics are similar to those reported in Table 2 of LM. In their Table 4, LM report 115,558 page views for Facebook's S-1 for an HTML-only measure, which I refer to as $LMpvH$. My calculation of $LMpvH$ is 115,054, indicating immaterial differences, likely due to a subtle difference in our summarization procedures.[10] $DRTpv$ for this filing-day is 124,057 and for $LMpv$, 117,081, and Rpv counts only 78,848 page views. The DRT and LM methods, which include all downloads by a single IP address, may induce a significant upward bias in download counts for certain filing types, compared to Rpv. In untabulated analysis, I find that the 115,054 page views calculated for $LMpvH$ are generated by only 77,989 different IP addresses, $DRTpv$ counts 124,057 page views from 78,569 IP addresses, while Rpv counts 78,848 page views from an equal number of IP addresses. All three measures reflect downloads from similar numbers of IP addresses (Rpv counts 1 percent more IP addresses than $LMpv$), yet the download count magnitude varies by as much as 58 percent between the lowest and highest value when multiple item views are

[10] I only retain requests where the EDGAR server log records an HTTP code of 200: OK, meaning the server delivered the requested file. LM keep all records with 2xx codes, which include, e.g., 204: No Content, where the server fulfilled the request but did not return any document body. `https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html`

included. Inspection of the underlying log files indicates that the inflated values for $DRTpv$ and $LMpv$ are due to these measures counting both repeat views of S-1s by the same IP address and user views of the filing's sole exhibit, an HTML document reporting Ernst & Young's consent.[11]

Table 3 provides analysis of where Rpv, $DRTpv$, $LMpv$, and $LMpvH$ provide similar and different inferences by examining their respective download counts and correlations across a variety of popular filing types for all EDGAR filings during the day. The total download counts (DL) in Panel A indicate that counts using $DRTpv$ and $LMpv$ can be inflated due to the issues with counting multiple page views. $LMpv$ is 12 percent higher than Rpv for all downloads during the day, and $DRTpv$ is 36 percent higher than Rpv, as a result of multiple page views.

While Panel A shows very high correlations (1.00) for all four count methods when all filings are included, the correlations differ, potentially significantly, when studying specific forms. Panels B through E illustrate that Form S-1, 8-K, 10-K, and 10-Q filings have quite high correlations (thought $DRTpv$ has consistently the highest page counts amongst all the measures), with correlations ranging from 0.88 to 1.00. Panels F through H, however, show that the measures are less correlated when examining Forms 4, UPLOAD, and CORRESP. For the study of comment letters, the fact that $LMpvH$ ignores PDF filings makes it inappropriate for use, as these documents are exclusively filed in PDF form. For comment letters, Forms UPLOAD and CORRESP, correlations among Rpv, $DRTpv$, and $LMpv$ of 0.43 to 0.60 indicate the potential for the misclassification of human versus robot readers to affect inferences.

[11] `https://www.sec.gov/Archives/edgar/data/1326801/000119312512034517/0001193125-12-034517-index.htm`

3.2. Robot vs Human Classification for Readers of Google's 2010 10-K Filing

The Google 2010 10-K presents a reasonably high-demand filing, for which I study differences in classification of IP addresses by Rpv, $DRTpv$ and $LMpv$. I look at the detailed browsing patterns of a selection of IP addresses to identify how the three measures make Type I and Type II errors. Figure 3 shows intra-day page views of Google's 10-K by both humans in Panel A and robots in Panel A, on February 11, 2011. The visual comparison of downloads categorized by the methods in Ryans, LM, and DRT illustrate similar overall patterns, though Ryans appears to record a few more human downloads on this day, perhaps reflecting that Ryans retains some more sophisticated users, who may be those most likely to read a filing as soon as it is released.[12] Figure 4 shows download patterns for this filing over the long term. The download pattern is similar in overall inference to that shown in LM's Figure 3, for a General Electric 10-K, as it shows that downloads persist at a relatively high level for one year following disclosure, then drop significantly when the subsequent 10-K is filed. In contrast to Figure 3 illustrating Ryans's higher raw human download count on the release day, Figure 4 shows that Rpv provides relatively lower long term page view counts compared to $LMpv$ and $DRTpv$, as these measures include multiple page views due to users' viewing of Interactive Data items and exhibits.

Table 4 illustrates the variety of 10-K related files and the number of downloads by users who look at the 10-K, including both robots and humans. The `d10k.htm` file is the most popular, and it is the basic HTML filing that a human

[12] In untabulated results, Ryans identifies 261 page views on this day, but the value is reduced to Rpv of 91 after eliminating multiple views by the same IP address. By comparison, $DRTpv$ records 109 views, and $LMpv$ record 206 views, higher than Rpv, due to multiple view-counts. $LMpvH$ records the fewest, with 74 views, as the non-HTML views are eliminated.

would tend to read in their web browser. `R1.xml`, `R2.xml`, etc., are the human-readable Interactive Data items generated from XBRL filing data, and correspond to views of the various financial statements and footnotes. `dex2301.htm` and similar refer to views of exhibit documents attached to the 10-K filing. The remaining `.txt`, `-xbrl.zip`, `.xml` and `.xsd` filings relate to machine readable file formats for the filing text and XBRL data. These documents are likely requested by a robot, or may reflect an accidental click by a human.

After classifying IPs as robot or human, *Rpv* counts all views of a filing by an IP address as a single download. *DRTpv* and *LMpv* first screen out robots according to their respective procedures, then consider all of the remaining listed downloads as separate downloads. *LMpvH* only considers downloads of the `d10k.htm`, and `dex*.htm` files, but again records each as a separate download. These differences matter to calculating the number of page views to a factor of almost three for counting views of this 10-K on its release day: in untabulated analysis, I find that *LMpvH* calculates the fewest page views at 74, *Rpv* calculates 91, *DRTpv* calculates 109, and *LMpv* calculates the highest at 206.

To examine the underlying factors leading to differences in the download counts between these measures, I conduct a detailed analysis of the download patterns on a sample of IP addresses where classification as human or robot are likely to differ among the three measures. Overall, the number of IP classification disagreements is relatively small: on February 11, 2011 there are 33,996 human IP addresses that viewed EDGAR filings, as determined by any of the three methods, yet there are only 1,319 users (3.9 percent) for whom there was not a unanimous classification. However, focusing on the Google 10-K readers increases the fraction with disagreement: for the 144 unique IP addresses that viewed the Google

10-K according to all three measures, there is disagreement on the classification of 39 users (27.1 percent). This is a much more significant fraction, and I study five IP addresses from this group in detail to try and identify the sources of classification error.

I also select four IP addresses from among those that view the `.txt` version of the 10-K, which LM discuss is a filing format more likely to be accessed by robots. With these `.txt`-viewing users, we have a setting where the IP address accessing the file has a prior likelihood of being a robot, but where at least one of the three methods' classification rules have identified the IP as human. As a setting for likely classification errors, this is a good test case to identify the ability of the methods to make correct classifications. There are 24 users who view the `.txt` version of the Google 10-K, and the highest volume address records 33,696 downloads, clearly indicating a robot. However, the `.txt` file is also downloaded by an IP address that only records 2 total downloads for the day, and so is very likely a human. Where Rpv, $DRTpv$, and $LMpv$ unanimously agree on the classification of these 24 cases, manual inspection of the download logs shows they appear to be correctly classified. I choose another four representative text-downloading IP addresses to examine further: three with low total downloads over the day—with a prior likelihood of being human—as well as one with a mid-range of downloads, 3,048—and therefore more likely to be a robot.

Table 5 describes the summary browsing behavior of the sampled IP addresses and tabulates the robot or human classification made by each method. *IP Address* is the identifier of the user, representing their partially disguised network address.[13] *Day Total DL* is the total number of items downloaded during the day

[13]The SEC scrambles the last byte of the IP address to preserve confidentiality, and replaces the

by the IP address. LM classifies an IP as a robot if this value exceeds 50, DRT if it exceeds 1,000, and Ryans if it exceeds 500. $Max\ DL/min$ is the maximum number of downloads per minute during the day by the IP address. DRT classifies an IP as a robot if this value exceeds 5, while Ryans classifies as a robot if this value exceeds 25. $Max\ CIK/min$ is the maximum number of different CIK's files accessed per minute by the IP address, and Ryans classifies an IP as a robot if this value exceeds 3. *Rpv Robot* is TRUE if the method in Ryans classifies the IP address as a robot, but false otherwise, and the same holds for *DRT pv Robot* and *LM Robot* as classified by the DRT and LM methods, respectively.

I review the download records for each IP address, including all downloads made by those IP addresses during the day, and classify each as human or robot, based on the volume and frequency of filings viewed, as well as the predominance of human-readable or machine-readable files accessed. For example, IP address 69.25.75.*gii* has numerous periods throughout the day with greater than 40 downloads in a single minute, activity that would not be possible for a human to generate. IP address 64.129.127.*e ja* on the other hand visits a variety of filings, including 8-Ks and 10-Ks at a rate of a few minutes per filing from the hours of 9am until 8pm. Only filings related to the same firm are accessed with moderate frequency, for example a series of 8-Ks are accessed within a few minutes of each other, but different firms' filings are accessed with delay, indicative of a manual search. Because this user visits 269 filings during the day, LM classifies the IP as a robot, but since the IP accesses no more than 5 filings per minute, Ryans and DRT appear to accurately classify the user as a human. IP address 174.253.161.*jde* il-

---

digits with a consistent character string so that the IP address remains uniquely identifiable across time in the log file.

lustrates how a low volume of downloads can still be a robot. This IP address only downloads 11 filings, but all 11 are 10-K or 10-Q documents, the documents were all disclosed on this day, and they are all downloaded in XBRL-machine readable format, at a rate of four different companies per minute. In combination these factors indicate a robot, and the Ryans method categorizes it as such, while the LM and DRT methods classify it as a human because of the relatively low volume and frequency.

I report the manually verified classification in the *Manual Inspection* column. The *Undetermined* case, IP address 69.191.241.*aha*, appears to have several periods of robot-like downloads during the day, with access to several different machine-readable XML filings for different CIK's throughout the day, but the Google 10-K is viewed in HTML format seven minutes after viewing another just-filed 10-K, and there are no further downloads for the user on this day. This pattern might indicate that 69.191.241.*aha* represents both human and robot behavior during the same day, making it difficult to definitively determine the appropriate classification. For the 8 determinable cases, $LMpv$ accurately classifies 3 of 8, $DRTpv$ accurately classifies 5 of 8, and Rpv accurately classifies all 8.

4. Conclusions

The EDGAR log files provide an interesting data set for the study of attention to financial reporting disclosures. Drake et al. (2015), Loughran and McDonald (2016), and Ryans (2017) use three related screening procedures, with different criteria and thresholds, to classify and eliminate robot downloads from their count statistics. Overall, the three methods agree on classification as robot or human for more than 96 percent of IP addresses. However, when looking at a single popular

10-K filing, the agreement drops to 73 percent. Researchers should be aware that the DRT and LM measures count all downloads recorded for a file during a day, as opposed to Ryans whose measure counts the number of unique IP addresses downloading the file. Counting multiple downloads by the same user during a day has the potential to inflate download counts by up to 58 percent for specific filings and 36 percent for total downloads during a day. Further, measures that focus on HTML-only downloads may provide misleading inferences, especially for the subset of filings that are not made in HTML format, including Forms 4, 13F-HRs, and comment letters, as these filings are provided in human-readable XML and PDF format. Ryans's method screens out multiple downloads and a detailed analysis of the download logs indicates that it appears to accurately classify human users in cases where the three measures do not all agree.

References

Barber, B. M., Odean, T., 2008. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies 21 (2), 785–818.

Bozanic, Z., Hoopes, J. L., Thornock, J. R., Williams, B. M., 2016. IRS attention. Journal of Accounting Research Forthcoming.

Bushee, B. J., Core, J. E., Guay, W., Hamm, S. J., 2010. The role of the business press as an information intermediary. Journal of Accounting Research 48 (1), 1–19.

Chen, H., Noronha, G., Singal, V., 2004. The price response to S&P 500 index additions and deletions: Evidence of asymmetry and a new explanation. Journal of Finance 59 (4), 1901–1930.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. Journal of Finance 66 (5), 1461–1499.

Dechow, P. M., Lawrence, A., Ryans, J., 2016. SEC comment letters and insider sales. The Accounting Review 91 (2), 401–439.

DellaVigna, S., Pollet, J. M., 2009. Investor inattention and Friday earnings announcements. Journal of Finance 64 (2), 709–749.

Drake, M. S., Roulstone, D. T., Thornock, J. R., 2012. Investor information demand: Evidence from Google searches around earnings announcements. Journal of Accounting Research 50 (4), 1001–1040.

Drake, M. S., Roulstone, D. T., Thornock, J. R., 2015. The determinants and consequences of information acquisition via EDGAR. Contemporary Accounting Research 32 (3), 1128–1161.

Engelberg, J. E., Parsons, C. A., 2011. The causal impact of media in financial markets. Journal of Finance 66 (1), 67–97.

Gervais, S., Kaniel, R., Mingelgrin, D. H., 2001. The high-volume return premium. Journal of Finance 56 (3), 877–919.

Irvine, P. J., 2003. The incremental impact of analyst initiation of coverage. Journal of Corporate Finance 9 (4), 431–451.

Lawrence, A., Ryans, J., Sun, Y., 2016. Yahoo Finance search and earnings announcements. Working Paper, University of California at Berkeley.

Lawrence, A., Ryans, J., Sun, Y., 2017. Investors' demand for sell-side research: SEC filings, media coverage, and market factors. The Accounting Review In-Press.

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

Lehavy, R., Sloan, R. G., 2008. Investor recognition and stock returns. Review of Accounting Studies 13 (2-3), 327–361.

Loughran, T., McDonald, B., 2016. The use of EDGAR filings by investors. Journal of Behavioral Finance Forthcoming.

Masden, J., 2017. Anticipated earnings announcements and the customer-supplier anomaly. Journal of Accounting Research Forthcoming.

Monga, V., Chasan, E., 2015. The 109,894-word annual report. The Wall Street Journal June 1.
URL http://www.wsj.com/articles/the-109-894-word-annual-report-1433203762

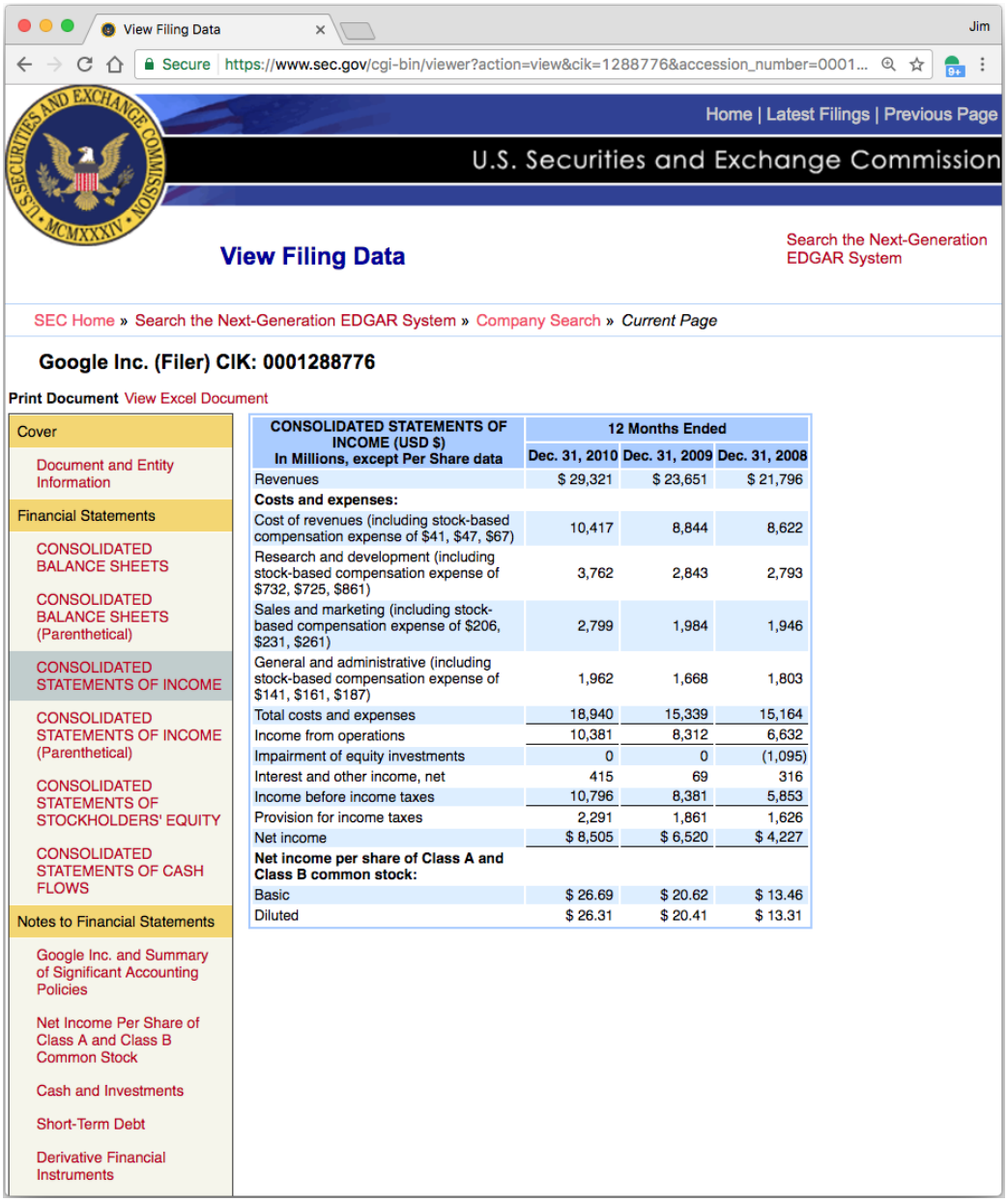
Ryans, J. P., 2017. Textual classification of SEC comment letters. Working Paper, London Business School.

Shleifer, A., 1986. Do demand curves for stocks slope down? Journal of Finance 41 (3), 579–590.

Figure 1: Google's 2010 10-K on the EDGAR Web Site

This figure illustrates the EDGAR index page for Google's 2010 10-K (`https://www.sec.gov/Archives/edgar/data/1288776/000119312511032930/0001193125-11-032930-index.htm`). In the header section there is an *Interactive Data* button for browsable access to the underlying XBRL data. Note the primary filing is `d10k.htm`, the HTML version of the 10-K viewable in a browser. There are also seven exhibits, and the complete filing in a machine-readible `.txt` format. Below the main document files are a set of data files with `.xml` and `.xsd` extensions. *LMpv* and *DRTpv* record for every file viewed by a user, including each Interactive Data footnote or item, whereas *Rpv* records only one download for the user if they access any of the items during a day.

Figure 2: Interactive Data Version of Google's 2010 10-K

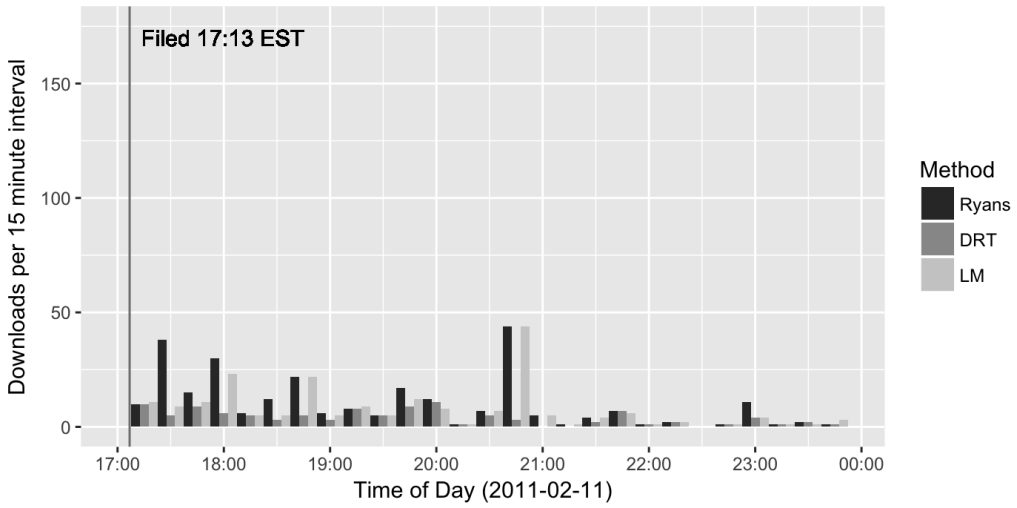


CONSOLIDATED STATEMENTS OF INCOME (USD $) In Millions, except Per Share data	12 Months Ended		
	Dec. 31, 2010	Dec. 31, 2009	Dec. 31, 2008
Revenues	$ 29,321	$ 23,651	$ 21,796
Costs and expenses:			
Cost of revenues (including stock-based compensation expense of $41, $47, $67)	10,417	8,844	8,622
Research and development (including stock-based compensation expense of $732, $725, $861)	3,762	2,843	2,793
Sales and marketing (including stock-based compensation expense of $206, $231, $261)	2,799	1,984	1,946
General and administrative (including stock-based compensation expense of $141, $161, $187)	1,962	1,668	1,803
Total costs and expenses	18,940	15,339	15,164
Income from operations	10,381	8,312	6,632
Impairment of equity investments	0	0	(1,095)
Interest and other income, net	415	69	316
Income before income taxes	10,796	8,381	5,853
Provision for income taxes	2,291	1,861	1,626
Net income	$ 8,505	$ 6,520	$ 4,227
Net income per share of Class A and Class B common stock:			
Basic	$ 26.69	$ 20.62	$ 13.46
Diluted	$ 26.31	$ 20.41	$ 13.31

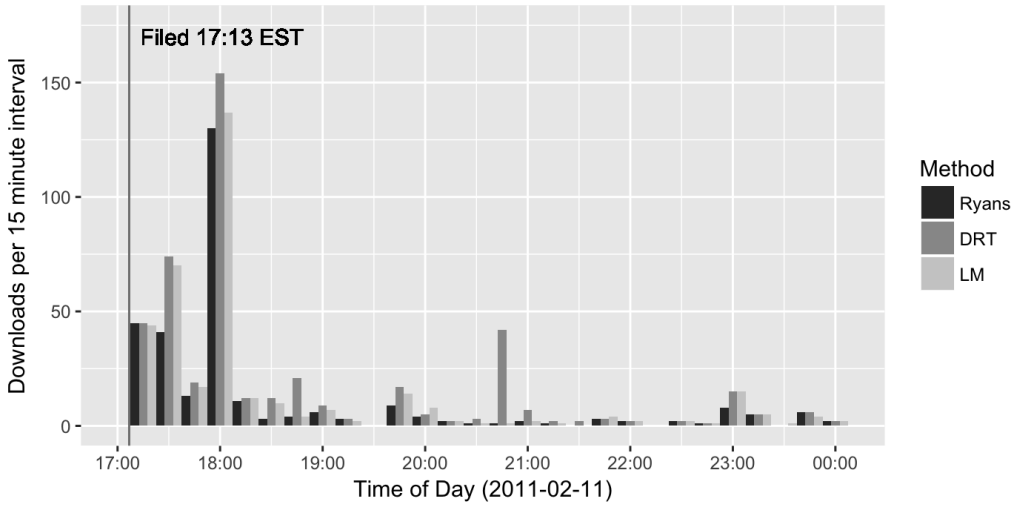
This figure illustrates the EDGAR Interactive Data page for Google's 2010 10-K. The income statement is illustrated in the body of the page, and the menu bar to the left allows the user to click on each of the financial statements and notes. Every item clicked is recorded as a download in the EDGAR log file.

Figure 3: Human vs. Robot Downloads of Google's 2010 10-K on its Disclosure Day

Panel A: Human Downloads

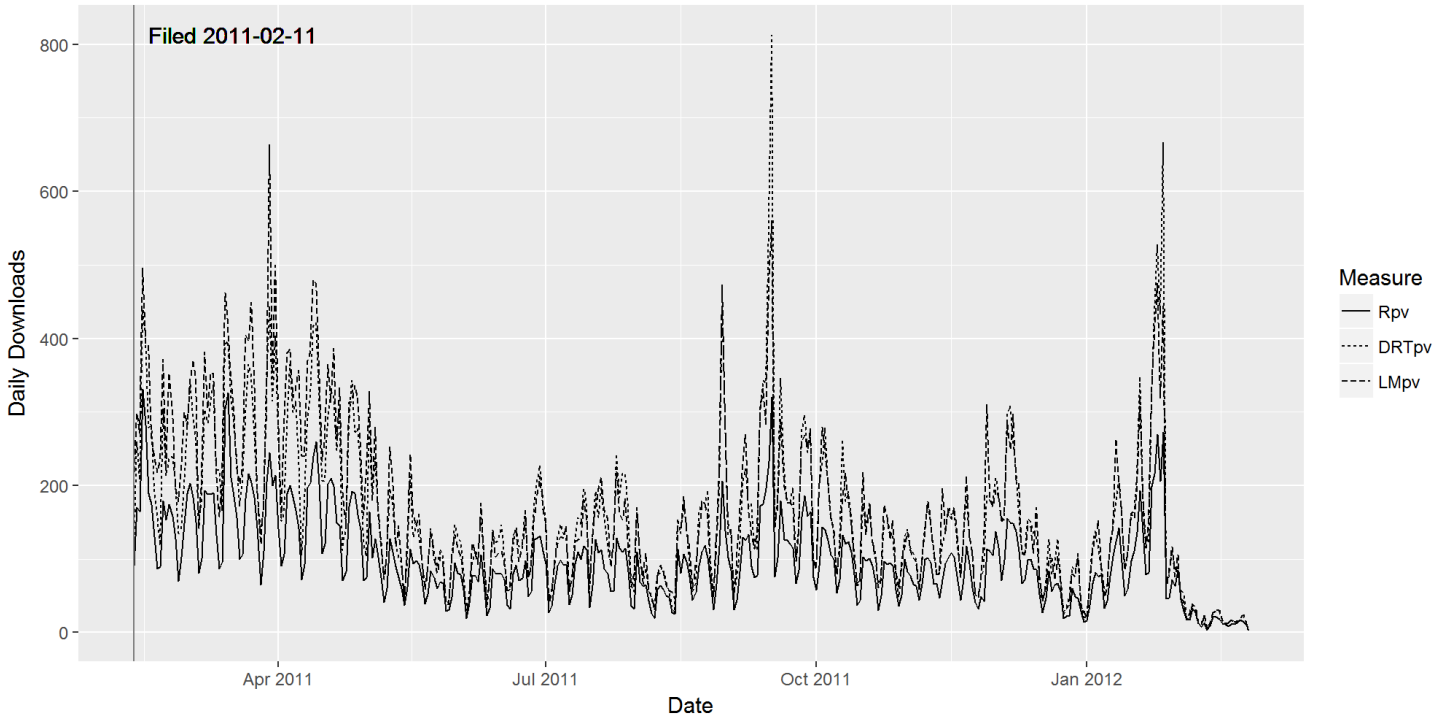


Panel B: Robot Downloads



This figure illustrates EDGAR downloads of Google's 2010 10-K on February 11, 2011, the date it was filed. Panel A illustrates the count of human downloads by 15 minute period, according to the Rpv methodology (inferences are similar using $DRTpv$ or $LMpv$), and Panel A illustrates the count of robot downloads by 15 minute period.

Figure 4: Long-Term Download Patterns for Google's 2010 10-K



This figure illustrates EDGAR downloads of Google's 2010 10-K from February 11, 2011, the date it was filed, until one month following the next year's 10-K, February 26, 2012. All three measures are plotted, and while Rpv is consistently a lower level than the other two measures, reflecting multiple same-day downloads by individual IP addresses with the $LMpv$ and $DRTpv$ measures, the download changes are very similar, illustrating the relatively high correlation among Rpv, $LMpv$, and $DRTpv$ (0.97 - 0.98) for this filing.

Table 1: Download Count Comparison by Filing Type

Panel A: Facebook Peak S-1 Download Day: February 2, 2012

Form	Rpv	DRTpv	LMpv	LMpvH
S-1	81,633	127,835	120,117	117,905
8-K	31,512	39,592	29,334	27,671
10-K	30,835	41,420	34,353	29,219
10-Q	24,440	32,399	27,052	24,708
4	8,677	10,705	7,375	8
DEF 14A	8,664	9,293	5,798	5,199
S-1/A	4,182	5,647	4,268	4,084
6-K	3,997	4,832	3,802	3,599
424B5	2,616	2,639	1,735	1,421
424B2	2,074	1,927	1,033	946
424B3	2,031	2,228	1,531	1,376
SC 13G/A	1,937	2,210	1,685	947
20-F	1,858	2,548	2,124	1,992
D	1,856	2,772	2,182	0
13F-HR	1,626	1,980	1,543	0
SC 13D/A	1,602	2,043	1,740	1,410
10-K/A	1,462	1,824	1,458	1,232
CORRESP	1,426	1,466	1,016	968
DEFA14A	1,380	1,500	1,054	965
SC 13G	1,371	1,499	1,127	685
485BPOS	1,358	1,846	929	755
FWP	1,314	1,183	725	603
424B4	1,309	1,542	983	881
UPLOAD	1,265	1,289	828	0
497	1,057	1,223	576	475
Total - All Forms	246,418	334,344	277,062	242,101

Panel B: Google 2010 10-K Disclosure Day: February 11, 2011

Form	Rpv	DRTpv	LMpv	LMpvH
8-K	29,964	33,356	26,471	24,494
10-K	29,555	33,259	28,462	22,541
10-Q	22,655	23,599	20,441	16,830
4	9,889	10,310	8,647	27
DEF 14A	9,066	8,565	5,071	4,648
SC 13G/A	6,854	7,522	6,417	3,496
S-1	4,972	6,709	6,008	5,747
13F-HR	4,471	4,962	3,729	0
SC 13G	3,896	4,424	3,621	1,866
S-1/A	3,339	3,886	3,024	2,896
6-K	3,306	3,700	2,966	2,695
424B5	2,194	1,999	1,315	1,110
424B3	1,872	1,785	1,318	1,163
SC 13D/A	1,803	1,929	1,381	1,104
10-K/A	1,770	2,015	1,575	1,263
424B2	1,655	1,504	978	913
20-F	1,636	1,892	1,656	1,547
D	1,606	2,236	1,959	0
DEFA14A	1,416	1,326	1,076	999
424B4	1,334	1,353	902	824
UPLOAD	1,168	1,185	991	0
CORRESP	1,032	983	827	755
485BPOS	1,021	1,299	908	505
SC 13D	1,013	1,090	874	612
FWP	976	814	581	494
Total - All Forms	173,574	191,054	153,279	110,422

These panels presents comparative download counts by form type for two sample days. Panel A corresponds to February 2, 2012, the day of Facebook's S-1 peak downloads. Panel B corresponds to February 1, 2011, the day Google's 2010 10-K was disclosed. . Rpv are page views calculated according to Ryans (2017), $DRTpv$ are page views calculated according to Drake et al. (2015), $LMpv$ are page views calculated according to the NR_total measure in Loughran and McDonald (2016), and $LMpvH$ is their NR_HTML measure, or page views of HTML documents only. By including only HTML files, $LMpvH$ shows an unusually small number of Form 4s downloads, and 0 downloads of Forms D, 13F-HR, and UPLOAD (comment letters), as these are filings that, while human readable, are not delivered as HTML documents.

Table 2: Sample Day Descriptive Statistics

Panel A: Facebook Peak S-1 Download Day: February 2, 2012

	N	mean	sd	min	q1	q10	q25	median	q75	q90	q99	max
Rpv	104,391	2.361	244.067	0	0	1	1	1	1	3	11	78,848
DRTpv	104,391	3.203	384.008	0	0	0	1	1	2	4	17	124,057
LMpv	104,391	2.654	362.415	0	0	0	0	1	1	3	15	117,081
LMpvH	104,391	2.319	356.136	0	0	0	0	0	1	3	13	115,054

Panel B: Google 2010 10-K Disclosure Day: February 11, 2011

	N	mean	sd	min	q1	q10	q25	median	q75	q90	q99	max
Rpv	105,341	1.648	7.265	0	0	1	1	1	1	3	10	2,047
DRTpv	105,341	1.814	10.922	0	0	0	1	1	2	3	13	3,148
LMpv	105,341	1.455	10.201	0	0	0	0	1	1	3	12	2,947
LMpvH	105,341	1.048	9.683	0	0	0	0	0	1	2	10	2,924

These tables presents descriptive statistics for two sample days. Panel A corresponds to February 2, 2012, the day of Facebook's S-1 peak downloads. Panel B corresponds to February 1, 2011, the day Google's 2010 10-K was disclosed. Rpv are page views calculated according to Ryans (2017), $DRTpv$ are page views calculated according to Drake et al. (2015), $LMpv$ are page views calculated according to the NR_total measure in Loughran and McDonald (2016), and $LMpvH$ is their NR_HTML measure, or page views of HTML documents only.

Table 3: Correlations Among *Rpv*, *DRTpv*, *LMpv*, and *LMpvH* for Selected Filing Types

Panel A: All Filing Types

	DL	Rpv	DRTpv	LMpv
Rpv	246,418			
DRTpv	334,344	1.00***		
LMpv	277,062	1.00***	1.00***	
LMpvH	242,101	1.00***	1.00***	1.00***

Panel B: S-1

	DL	Rpv	DRTpv	LMpv
Rpv	81,633			
DRTpv	127,835	1.00***		
LMpv	120,117	1.00***	1.00***	
LMpvH	117,905	1.00***	1.00***	1.00***

Panel C: 8-K

	DL	Rpv	DRTpv	LMpv
Rpv	31,512			
DRTpv	39,592	0.88***		
LMpv	29,334	0.90***	0.91***	
LMpvH	27,671	0.90***	0.91***	0.99***

Panel D: 10-K

	DL	Rpv	DRTpv	LMpv
Rpv	30,835			
DRTpv	41,420	0.97***		
LMpv	34,353	0.96***	0.97***	
LMpvH	29,219	0.97***	0.97***	0.98***

Panel E: 10-Q

	DL	Rpv	DRTpv	LMpv
Rpv	24,440			
DRTpv	32,399	0.90***		
LMpv	27,052	0.83***	0.84***	
LMpvH	24,708	0.81***	0.82***	0.99***

Panel F: 4

	DL	Rpv	DRTpv	LMpv
Rpv	8,677			
DRTpv	10,705	0.69***		
LMpv	7,375	0.73***	0.67***	
LMpvH	8	0.00	0.03**	0.04***

Panel G: UPLOAD

	DL	Rpv	DRTpv	LMpv
Rpv	1,265			
DRTpv	1,289	0.50***		
LMpv	828	0.43***	0.44***	
LMpvH	0	NaNNA	NaNNA	NaNNA

Panel H: CORRESP

	DL	Rpv	DRTpv	LMpv
Rpv	1,426			
DRTpv	1,466	0.60***		
LMpv	1,016	0.58***	0.57***	
LMpvH	968	0.58***	0.55***	0.98***

This table presents the download count (*DL*) for, and Pearson correlations among, *Rpv*, *DRTpv*, *LMpv*, and *LMpvH* for selected filing types on a sample day, February 2, 2012 (Facebook's S-1 peak download day). These panels illustrate that for many filing types, correlations between the different measures of EDGAR downloads are similar. Correlations for all filings are 1.00 in Panel A and 0.96 to 0.98 in Panel D for 10-K filings. However correlations for other filings are much lower, for example 0.43 to 0.50 for form UPLOAD, and *LMpvH* cannot measure demand for Form UPLOAD, because this form is provided in PDF, and therefore *LMpvH* registers no downloads, even though *LMpvH* measures 968 downloads for CORRESP filings, which are the company responses to comment letters.

Table 4: Google 10-K Downloads by Extension Type

Extension	Downloads
d10k.htm	99
.txt	36
goog-20101231.xml	36
-xbrl.zip	34
R1.xml	28
goog-20101231.xsd	21
goog-20101231_cal.xml	20
goog-20101231_lab.xml	20
goog-20101231_def.xml	20
goog-20101231_pre.xml	20
R2.xml	15
R4.xml	12
R3.xml	10
dex2301.htm	8
R6.xml	8
R7.xml	8
dex2101.htm	7
dex301.htm	7
dex302.htm	7
R8.xml	6
Total - All Extensions	574

This table presents the number of downloads recorded for the 20 most-downloaded file extensions for Google's 2010 10-K filed on February 11, 2011. The page providing access to these files is illustrated in Figure 1. *Downloads* is the number of raw downloads, including robots. Extension is the file extension in the EDGAR web log associated with the downloads, i.e. which file associated with the 10-K package was actually requested. `d10k.htm` is the basic HTML filing that a human reader would tend to read in their web browser. `R1.xml`, `R2.xml`, `R3.xml`, etc., are human-readable "Interactive Data" paragraphs generated from the XBRL filing data, and correspond to views of the various financial statements and footnotes. `dex2301.htm` and similar refer to views of exhibit documents attached to the 10-K filing. The remaining `.txt`, `-xbrl.zip`, and `goog*.xml/.xsd` filings relate to machine readable file formats for the filing text and XBRL data. These documents are either requested by a robot, or may reflect an accidental click by a human.

Table 5: Comparison of Robot Classification for Case Study IP Addresses

IP Address	Day Total DL	Max DL/min	Max CIK/min	DRTpv Robot	LMpv Robot	Rpv Robot	Manual Inspection
69.25.75.gii	3,048	89	47	TRUE	TRUE	TRUE	Robot (TRUE)
64.129.127.eja	269	5	3	FALSE	TRUE	FALSE	Human (FALSE)
12.34.6.gjj	52	3	3	FALSE	TRUE	FALSE	Human (FALSE)
69.191.241.aha	46	4	4	FALSE	FALSE	TRUE	Undetermined
208.103.34.dfg	41	20	1	TRUE	FALSE	FALSE	Human (FALSE)
75.150.52.cih	27	8	4	TRUE	FALSE	TRUE	Robot (TRUE)
212.162.50.dif	18	4	4	FALSE	FALSE	TRUE	Robot (TRUE)
174.253.161.jde	11	4	4	FALSE	FALSE	TRUE	Robot (TRUE)
184.88.26.jdd	2	1	1	FALSE	FALSE	FALSE	Human (FALSE)

This table presents descriptive statistics the categorization of particular IP addresses by Ryans, DRT, and LM. The IP addresses are chosen because they each downloaded Google's 2010 10-K on the filing date, five were classified differently by the three methods (12.34.6.gjj, 174.253.161.jde, 212.162.50.dif, 64.129.127.eja, and 69.191.241.aha), and four are addresses that downloaded the `.txt` document format, believed to be only of interest to robots (184.88.26.jdd, 208.103.34.dfg, 69.25.75.gii, and 75.150.52.cih). The SEC scrambles the last byte of the IP address to preserve confidentiality, and replaces the digits with consistently applied characters so that the IP address remains uniquely identifiable across time in the log file. $MaxDL/min$ is the maximum number of downloads per minute during the day by the IP address. $DayTotalDL$ is the total number of items downloaded during the day by the IP address. $MaxCIK/min$ is the maximum number of different CIK's files accessed per minute by the IP address. *RpvRobot* is TRUE if the method in Ryans classifies the IP as a robot, false otherwise, and the same holds for $DRTpvRobot$ and *LMRobot* as classified by DRT and LM respectively. Manual Inspection indicates that manual review of the IP's comprehensive download records for the day appears to be be consistent with a robot or human behavior. The *Undetermined* case appears to have several robot-like downloads during the day, but the Google 10-K is viewed independently in HTML format, allowing for the possibility that the IP address has the behavior of both a human and a robot during the same day.