Crawling EDGAR

Diego García*
Kenan-Flagler Business School, UNC at Chapel Hill

Øyvind Norli
Norwegian School of Management

March 21st, 2012

Abstract

While the title may lead you to think that this paper is about spiders, it is about firms in the United States reporting relevant business information to the Securities and Exchange Commission (SEC). The paper is meant to serve as a primer for economists in the computing details of searching for information on the Internet. One important goal of the paper is to show how simple open-source computer scripts can be generated to access financial data on firms that interact with regulators in the United States.

*Corresponding author, 4409 McColl, CB#3490, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599–3490, tel: 1–919–962–8404, fax: 1–919–962–2068, email: diego_garcia@unc.edu

1 Introduction

Business relevant information is more easily available today than ever before. Information about corporations, investors, and security markets get disseminated through the Internet almost instantaneously. For the most part, the available information is unstructured in the form of a text. It is easy to see that a strategy of trading on information acquired from free form text would become more profitable the faster you are able to read the text. Hence it is not surprising that text analytics is becoming increasingly important on Wall Street.[1] Hoping to capture the current mood of investors, some traders are using computer programs to monitor and decode the words, opinions, rants and even keyboard-generated smiley faces posted on social networking sites.[2] Academia has followed suit. Computerized decoding of "textual information" into quantitative metrics has become an important area of research in financial economics.

This paper is meant to be a teaser to researchers in financial economics that lowers the costs of entry into the field of text analytics. The paper develops, presents and explains a set of simple Perl programs that will allow access to the electronic filing system (EDGAR) used by the U.S. Securities and Exchange Commission (SEC) to disseminate business relevant information. To illustrate how to download and extract information from EDGAR, we use Form 8-K to analyze executive turnover (new hires and departures of corporate executives). We investigate if there is a calendar effect in executive turnover (there is). But the findings on this particular question are not the main point of this paper. Our key contribution is to show how easy it is to access and analyze the various forms that companies and investors file electronically with the SEC.

The empirical literature that uses textual data as their main data source is growing. Tetlock, Saar-Tsechansky, and Macskassy (2008), García and Norli (2012), Phillips and Hoberg (2010), and Kogan, Routledge, Sagi, and Smith (2009) analyze the annual report filed by firms on Form 10-K. Another strand of the literature has focused on textual analysis of newspaper articles: Tetlock (2007) picks up investor sentiment by analyzing reports on the state if the stock market, while Dougal, Engelberg, García, and Parsons (2012) use exogenous scheduling of Wall Street Journal columnists to identify a causal relation between financial reporting and stock market performance. Engelberg (2008) analyze earnings announcements, Hoberg and Hanley (2012) study IPO prospectuses.[3]

[1] Text analytics covers tagging and annotations, word counting, pattern recognition, etc. The purpose of text analytics is to turn an unstructured text into data that can be analyzed.

[2] USA Today, May 4th, 2011, "Wall Street traders mine tweets to gain a trading edge."

[3] Other related papers include Chan (2003), Barber and Odean (2008), Engelberg and Parsons (2011), DellaVigna and Pollet (2009), Loughran and McDonald (2011), García (2012), Solomon (2009), Davis, Piger,

The rest of the paper proceeds as follows. In section 2 we present some details on the EDGAR filing system. The next section presents a few simple algorithms to extract basic information from 8-K statements filed with the SEC through EDGAR. Section 4 presents an analysis of the calendar effects around aggregate filings of 8-K statements that discuss executive turnover. The last section concludes.

2 EDGAR

Companies and others are required by law to file a number of different forms with the U.S. Securities and Exchange Commission (SEC). The main purpose of filing these forms is to make certain types of information available to investors and corporations—and by that improve the efficiency of security markets. Historically these forms have been filed with the SEC on paper. In the early 1990s the SEC developed the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system to handle electronic form filing.[4] As of May 6, 1996 all public U.S. companies were required to make all their filings, with a few exceptions, on EDGAR. More importantly, any person with access to a computer linked to the Internet can obtain and read these filings within seconds after they are filed.

As researchers looking for relevant information on companies with operations in the United States, we have traditionally relied on databases such as Compustat, ExecuComp, SDC Platinum, etc. These databases are attractive because their owners have collected data from companies' filings and organized the information in a structured way. Most of the information that is found in Compustat comes from Form 10-K. Most of the information in ExecuComp comes from Proxy filings (Form DEF 14A) and Forms 3, 4, and 5. The merger information in SDC Platinum relies heavily on the forms filed during the period leading up to a merger. Since the introduction of EDGAR, researchers have had easy access to this "standard" information in addition to an enormous amount of information not found in any other database.

To get an idea of what type of information that is available through EDGAR we move on to looking at the most common forms filed with SEC through EDGAR. Table 1 reports the filing frequency of the 20 most commonly filed forms over the period 1994 through 2011.

and Sedor (2007), Loughran and McDonald (2008).

[4]The SEC describes EDGAR as follows: "EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (SEC). Its primary purpose is to increase the efficiency and fairness of the securities market for the benefit of investors, corporations, and the economy by accelerating the receipt, acceptance, dissemination, and analysis of time-sensitive corporate information filed with the agency." For more information on EDGAR, visit: `http://www.sec.gov/edgar.shtml`.

The first column in the Table contains a short description of the form. The second column contains the form code used on EDGAR. The third column contains the total number of times a forms is filed during the whole sample period.

The most common EDGAR filing is Form 4. For the sample period 1994–2011 this form is filed more than four million times. Form 4 is used to report purchases or sales of securites by persons who are the beneficial owner of more than 10 percent of any class of any equity security, or who are directors or officers of the issuer of the security. This form would, for example, allow you to study the granting of options to officers or directors. Table 1 also shows that Form 4/A is a commonly used form. When "/A" is appended to a form code it means that the filing is an amendment to an existing filing. Thus, a specific corporate event could be linked to an initial filing and a subsequent string of amendments to this initial filing. Form 3 and Form 5, also prevalent in EDGAR, deal with similar ownership issues.

The second most common EDGAR filing is Form 8-K, with more than one million filings. Companies have to use this form to file information on issues that are of "material importance" for the firm. The 8-K statements include information on changes in management, new significant contracts, merger negotiations, lawsuits, etc. In the next sections of the paper we will use the 8-K Forms to illustrate how one can use simple computerized parsing to extract information from the EDGAR filings.

Another important subset of EDGAR is comprised of Form SC 13D (commonly referred to as Schedule 13D) and Form SC 13G. Filing of these forms are triggered when someone crosses the the 5% ownership threshold in a firm. The 13Ds are "active" investors, say those seeking control of the firm, whereas the 13Gs are from "passive" investors. There are on the order of 1 million such filings (including ammendments).

Annually and quarterly statements also figure prominently in the EDGAR system. There are well over 400,000 10-Q forms, and over 100,000 10-K statements. Other forms that come up in the "top-twenty" list in Table 1 are: foreign firms' current reports, Form 6-K; forms having to do with issuance of securities, from prospectuses, such as Form 424B3, to exemptions from regulation D; forms specific to institutional managers, such as quarterly holdings reported on Form 13F.

Filers in the EDGAR system are uniquely identified using the Central Index Key (CIK). For the sample perod 1994 through 2011, there are 452,830 unique CIKs in the EDGAR database. Only a fraction of these CIKs are publicly traded firms. There are many filers that are private firms. These private firms include manufacturing firms, but also hedge funds and mutual funds. You will also receive a CIK if you are filing on behalf of yourself as an individual.

Table 2 reports the number of filers (unique CIKs) that file a particular type of form. We see there are more than 171,000 filers that have filed a form 4 (or an associated ammendment) at some point during the sample period. There are about 40,000 filers that have filed 13-Ds, with a very similar number of 13-G filers. The total number of firms that file some type of 10-K report adds up to over 36,000. This is similar in magnitude to the number of firms that file 8-K statements.

The three last columns of Table 2 report descriptive statistics on the number of filings of a particular form (for the subset of firms that actually file that particular form.) We see that while the number of firms that file 10-K and 8-K statements is about the same, a given firm typically files more 8-K statements than 10-K statements. On average, companies that file 8-K have typically filed over 30 such statements, while they, on average, only have filed six 10-K statements. This is most noticable when looking at the last column in Table 2, which lists the maximum number of a given form by a given firm. Chase, the financial conglomerate, has filed a total of 1,347 8-K statements. On the other hand, the firm with the most 10-K statements file (Old Republic International, an insurance firm, which filed many ammendments), only has a total of 67 10-K filings. This contrast is also found for other types of filings. Fidelity has filed over 27,000 13-G statements with the SEC, and GAMCO over 5,000 13-D statements.

3 Downloading Filings from EDGAR

In this section we explain how to download filings from EDGAR. The section will walk you through the following routine tasks with the EDGAR database: (a) Downloading and reading quarterly master index files. These files contains, among other things, a link to the file for every form filed during the quarter, (b) downloading and reading all 8-Ks filed during the period 1994 through 2011, (c) extracting information from the 8-Ks. We provide a set of Perl routines that should be easily adapted to particular research projects.[5]

Perl is an open source interpreter language best known for its powerful text processing facilities.[6] This makes it a natural choice for the problem at hand. The Perl code that we provide in this paper is written for a Unix operating system. But it can be easily adapted

[5]Engelberg and Sankaraguruswamy (2007) also provide a set of SAS routines to crawl EDGAR documents. Our paper is more comprehensive, in terms of laying out a particular question and tackling it directly, as well as providing some new facts on the EDGAR database itself. Diego is thankful to Joey for prompting him to learn how to crawl.

[6]Most Unix systems come with Perl installed. For Windows there are several implementations available. We have tested Strawbery Perl with success for this project. The reader may want to search online for a Perl primer, or read the "Camel book" (Wall, Christiansen, and Orwant, 2000).

for other operating systems.

Downloading Quarterly EDGAR Index Files. To be able to effectively download company filings from EDGAR, you will have to know where EDGAR stores the files associated with each filing made through the system. For this purpose EDGAR provide a set of quarterly index files. The index file for a specific quarter contains information about every filing made during the quarter. The first entry for the last quarterly file of 2011 reads:

```
1000032|BINCH JAMES G|4|2011-12-02|edgar/data/1000032/0001181431-11-058482.txt
```

The CIK of this filer is `1000032`. Since this entry refers to a Form 4 filing, the filer is a person and not a company. The filing date is December 2, 2011. The textfile associated with the file can be downloaded from:

```
ftp.sec.gov/edgar/data/1000032/0001181431-11-058482.txt
```

You will find all documents submitted for a given filing in the folder (the last folder name is the accession number without dashes):

```
ftp.sec.gov/edgar/data/1000032/000118143111058482/
```

This folder can also be accessed using the http protocoll and your favorite browser:

```
http://edgar.sec.gov/Archives/edgar/data/1000032/0001181431-11-058482-index.html
```

Program 1 presents a Perl program that downloads the master files of the EDGAR database. The routine starts by calling a package, `LWP::UserAgent`, which is one of the many packages provided in `www.cpan.org` in order to interact with the Internet from a perl script. We include others, i.e. `WWW::Mechanize`, in what follows. The program then opens a browser, by creating the object `$ua`, and then grabs files from the EDGAR ftp site, going one quarter/year at a time. The output is saved into 80 different files with the following structure:

```
...
2001QTR1master
2001QTR2master
2001QTR3master
2001QTR4master
2002QTR1master
...
```

Each master file can take as much as 30 Mb of hard disk space, especially for files during the last 10 years. The master files will be the input for the next program.

Organize Index information by CIK. Although it is possible to access and analyze information in EDGAR filings without downloading and saving any files to you local computer, it is more efficient to have the index files and associated filings saved locally. The reason why this saves time is that you will have to analyze documents several times during the course of a research project. Having everything locally implies that you are not vulnerable to the speed of the EDGAR ftp server. With a fast Internet connection that never goes down one can leave the files on the EDGAR server and read from there whenever necessary. But, most of us will not have this luxury.

There are many ways to organize the files and the information you download from EDGAR. We first create a directory structure with individual files that contain the entries of each individual CIK. This is clearly something of interest, as typically we are interested on the behavior of particular economic entities. Furthermore, such a directory structure is a must-do when working with the amounts of data that one may need when working with EDGAR. Program 2 breaks the EDGAR master files into smaller master files for each CIK. The program also creates a directory structure with 906 different directories (labeled 000–905). Each directory has 500 different files, where each file stores the content of the EDGAR masterfiles for that particular CIK as a `.dat` file (the 906 directories work out to take care of the over 450,000 distinct CIKs).

Summarizing information contained in the master files. Before jumping into the analysis of master files, we need to give a small aside on hardware requirements for this project. The first one is obvious, sufficient disk space. The raw text files from the EDGAR database can take from over 200 Gigabytes for all 8-K statements, to 10-20 Gigabytes for 13D/13F/13G statements (10-K statements take up on the ballpark of 100 Gigabytes). Another technical issue is the number of different files one must deal with. On a unix system, our experience is that creating directory structures with under 1,000 "objects" per tree node works well (Program 2 uses 500).

Program 3 reads the `*.dat` files created by Program 2. To be able to find the `*.dat` files we have created a master list of these files using the Unix command: `ls -Rl * > AllEDGARfiles.dat`. If you are on a different system than Unix and do not have access to this command you may have to let Program 2 create the file `AllEDGARfiles.dat`.

Program 3 creates two files. The first, `statsEDGAR.dat`, will be used as input to other programs. This file contains: (1) the directory where the `cik.dat` file resides, (2) the number of EDGAR filings for that CIK, (3) the cik number, (4) the date of the first entry in the first quarter file for that cik, (5) the date of the last entry in the last quarter file for that cik,

followed by the number of filings for particular forms (namely 10-K, 10-Q, 6-K, 8-K, 4, 13G, 13D, 13F, 424Bs, 424B3). Note how the code identifies files of different types by different means. For example, it counts forms 4 only if the string in the EDGAR file is either `"4"` or `"4/A"`. A 10Q form is defined by whether the form type contains either the strings `"10Q"` or `"10-Q"`. Table 2, discussed in the previous section, is constructed from the information in `statsEDGAR.dat`.

The second file created by Program 3, `formtypesEDGAR.dat`, is a list of all form types. It seemed to be a good idea to have a comprehensive list of all the strings that populate the form type field in EDGAR. Table 1 is constructed with information from this file. We note that we count different text strings as different types of forms, so ammendments are counted as different forms in Table 1.

Downloading all 8-K filings for 1994-2011. We end this section with a crawling algorithm. Program 4 downloads all Form 8-K form EDGAR. The program opens `statsEDGAR.dat` and first looks for CIKs that have filed at least one Form 8-K. For each CIK that has filed an 8-K, it opens the local master file for that CIK, looks for the strings 8-K or 8K in the form name, and saves those appropriately to the local disk. When all 8-Ks for all CIKs are identified, lines 37 through 56 in Program 4 download the 8-K text files from EDGAR and save the file to the local hard drive. Note how the actual crawling code in Program 4 is minimal. Most of the code is simply formatting the relevant filenames correctly. In the next section we analyze the content of the downloaded 8-K filings.

4 Analyzing Form 8-K

Say that you are working on a project on executives, and you want to know when "material evidence" is reported by firms with respect to executive turnover. The "current report" Form 8-K discussed previously, would be a natural place to look for this type of information.[7] Figure 1 plots the daily number of 8-K filings in the EDGAR database for the full sample period. It is interesting to note how there seem to be two regimes shifts. One in 1997, just after filing electronically via EDGAR became mandatory. The second shift might be related to the enaction of the Sarbanes-Oxley Act in July of 2002. While these shifts are interesting on their own, we will only use the full sample of 8-K statements as a control group in what follows.

[7] Clearly 10-Q and 10-K statements also have information on executive turnover. The choice of 8-K statements is mostly due to the fact that they have been mostly neglected in the financial economics literature.

In order to gather data for a project on executives' turnover, we need to come up with a set of text strings that we can use as keywords in a search using our 8-K database. For example, one could create a file `wordlist.dat` which contains the strings: departure of directors, election of directors, appointment of officers, resignation of directors. Clearly a current report (8-K) that contains such text strings will be discussing executives and corporate governance topics. The choice of strings in this example is motivated by the discussion in the EDGAR database on 8-K statements, which details what type of information should be disclosed on corporate governance and management (see in particular Item 5.02). At this point, the reader can imagine any given set of text strings that she may be interested in.

Program 5 reads through all 8-K statements in our database and counts the number of occurences of the set of text strings saved in `wordlist.dat`. Since this is a particularly important type of routine, we will comment on the `grep` command that takes care of matching words. The command looks for occurences of a given string, ignoring case and looking for all matches (the `ig` flags). The flags `\b` are "word boundaries." This is important in some cases, as strings of text can be part of a word (rather than a whole word). The program starts by loading a list of files, from `matched.8k`, which is an output file from a "check" program that verifies what files was actually downloaded by Program 4.

Figure 2 plots the time-series of the number of 8-K statements that mention a given string on a given day (the string being the title of each plot), for the time period 2005–2011.[8] Two of our four strings pick up a substantial number of 8-K statements: There are 123,177 8-K statements that contain "departure of directors," and 160,258 that have "election of directors." The other two strings are less common, with only 1,478 and 1,991 distinct 8-K statements mentioning "appointment of officers" and "resignation of directors." Furthermore, the correlation between the number of 8-K statements mentioning the first two strings is over 0.9, so we collapse the two metrics into a single count of the number of 8-K statements that mention either of the two strings.[9] Figure 2 and the fact that more than 15% of all 8-K statements contain the strings we tried, suggest that we have meaningful metric of executive turnover.

Since the focus of the paper is on presenting algorithms that readers can recycle for their own questions, we focus on a simple test that looks at calendar effects in executive turnover. Does turnover occur randomly throughout the year, or is there clustering around

[8] The number of hits of our text strings is significantly smaller prior to 2005.

[9] There are 123,004 current reports that mention both "departure of directors," and "election of directors." There are 37,254 that mention "election of directors," but not "departure of directors." There are only 173 that include "departure of directors" but not "election of directors." We choose to neglect the other two strings simply due to the lack of sufficient reports that include them.

year's end? Do firms announce executive turnover questions on random days throughout the week, or do they wait to release information until Friday? The former question can rule out some economic theories in which agents do not use "anchors" such as year's end when making decisions. The latter could shed some light on the common folklore of "bad earnings announcements on Fridays." Finally, given we also have the time-series of all 8-K statements, we can study the same questions for the whole universe of "current reports," and see if there are any differences.

In an attempt to answer these questions, we let Y_t denote the number of firms filing an 8-K statement with either of our two strings "departure of directors" and "election of directors." Also let X_t denote the total number of 8-K statements filed on a given day (irrespective of its content). We estimate the following models:

$$X_t = \alpha_x + \boldsymbol{\beta}_x \mathbf{D}_t + \boldsymbol{\gamma}_x \mathbf{M}_t + \boldsymbol{\eta}_x \mathbf{T}_t + \upsilon_t; \quad (1)$$

$$Y_t = \alpha_y + \boldsymbol{\beta}_y \mathbf{D}_t + \boldsymbol{\gamma}_y \mathbf{M}_t + \boldsymbol{\eta}_y \mathbf{T}_t + \epsilon_t; \quad (2)$$

where $\mathbf{D}_t$ denotes a vector of day-of-the-week dummies (Monday being the omitted one), and $\mathbf{M}_t$ denotes month-of-the-year dummies. The vector $\mathbf{T}_t$ denotes time-trends controls, namely a linear term $t - 2004$, and a quadratic term $(t - 2004)^2$. We estimate the model for the period 2005–2011.

Table 3 presents the point estimates. The first two numeric columns give the results for all current report filings, whereas the last two contain those pertaining to the 8-K statements dealing with executive turnover (as defined previously). Turning to the weekly effects, we see that the dummies for Tuesday-Friday are all significantly positive. The evidence suggests that Mondays have 52.9 less 8-K statements than Tuesdays. The day of the week with the most filings is Thursday. As it turns out, Friday is the day with the second least filings, only 22.7 more than Mondays. The differences between the Tuesday-Thurday dummies, and those corresponding to Mondays and Fridays are economically large—anywhere from 20 to 90 more filings.

Turning to the evidence on filings having to do with executive turnover, we see a very different picture. There is virtually no statistical difference between Mondays, Tuesdays, Wednesdays, or Thursdays. On the other hand, Fridays have an average of 10.3 more 8-K statements dealing with executive turnover than a Monday. Thus, we conclude that firms do seem to be particularly fond of Fridays as a day when to file material information having to do with executive turnover. Whether this preference for the end of the week is strategic behavior by the firm, or whether it is related to the actual content of the 8-K statements,

seems like an interesting avenue for future research.

We next study the seasonality along the different months of the year. The estimates are presented in Panel B of Table 3. Looking at the evidence on the full sample of 8-K statements, we find that three months are associated with particularly high levels of current reports, February, May, and November. Another three months exhibit particularly low levels, June, September, and December. This seasonality is most likely the result of other quarterly and annual filings. Recall the 8-K statements are giving regulators "material evidence" between the other required filings, most notably annual and quarterly statements. The latter acts as a substitute to 8-K statements on the months when they are filed (typically March, June, September, December).

The results for the 8-K associated with executive turnover present a significantly different pattern. The largest point estimates correspond to the months of February, May and December, whereas the six months starting in June and ending in November are associated with a significantly lower incidence of material evidence on executives.

One natural next step would be to do further text analysis of this subset of 8-K statements. We could try to extract the names of the executives involved via regular expressions, and more details as to the reasons for the departures or the background of new directors. Another possibility would be to cross our 8-K metrics, at the individual firm level, with other databases in Finance. Clearly price reactions to such announcements seem an interesting route to pursue, and using the CIK/GVKEY link file one can access both Compustat and CRSP.

5 Conclusion

This short paper has presented a simple set of Perl routines to crawl and read the EDGAR database. It has presented a self-contained sequence of programs that allowed us to see when firms file material evidence outside the quarterly and annual statements (via 8-K statements), focusing on material evidence having to do with executive turnover.

The exercise itself was meant to be a simple illustration of how to access information filed through EDGAR. But we uncovered a few notable findings: Fidelity filing thousands of 13-G statements, a structural break on 8-K filings around shortly after the passing of the Sarbanes-Oxley Act, and some striking difference in the timing of 8-K statements having to do with executive turnover.

While the study of 10-K statements can probably be considered mainstream finance by now, it strikes us as somewhat surprising that EDGAR filings are typically analyzed in isolation. There is clearly much to be learned about how firms sequentially release information

to the market and to regulators, and the interactions between the different types of filings.

The main goal of this project was to create a set of blueprints that other researchers can use to expand our understanding of formal communications between economic entities and the SEC. The electronic materials that accompany this paper provide further scripts to study the EDGAR database. While we do not plan to work on EDGAR for the rest of our careers, we are willing to put in the time to create a set of algorithms that can be reused by others. We hope that this short article will help lower the costs of the computing part of crawling EDGAR, so that financial economists can focus on finding interesting questions that may be answered with some simple Perl scripts.

References

Barber, B., and T. Odean, 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.

Chan, W. S., 2003, "Stock price reaction to news and no-news: drift and reversal after headlines," *JFE*, 70, 223–260.

Davis, A. K., J. Piger, and L. M. Sedor, 2007, "Beyond the Numbers: Managers' Use of Optimistic and Pessimistic Tone in Earnings Press Releases," working paper, University of Oregon.

DellaVigna, S., and J. M. Pollet, 2009, "Investor inattention and Friday earnings announcements," *Journal of Finance*, pp. 709–749.

Dougal, C., J. Engelberg, D. García, and C. Parsons, 2012, "Journalists and the stock market," *Review of Financial Studies*, forthcoming.

Engelberg, J., 2008, "Costly information processing: evidence from earnings announcements," working paper, UNC at Chapel Hill.

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

Engelberg, J., and S. Sankaraguruswamy, 2007, "How to gather data using a web crawler: An application using SAS to search EDGAR," working paper, UCSD.

García, D., 2012, "Sentiment during recessions," working paper, UNC at Chapel Hill.

García, D., and Ø. Norli, 2012, "Geographic dispersion and stock returns," *Journal of Financial Economics*, forthcoming.

Hoberg, J., and K. Hanley, 2012, "Litigation Risk, Strategic Disclosure and the Underpricing of Initial Public Offerings," *Journal of Financial Economics*, forthcoming.

Kogan, S., B. R. Routledge, J. S. Sagi, and N. Smith, 2009, "Information Content of Public Firm Disclosures and the Sarbanes-Oxley Act," working paper, University of Texas at Austin.

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

——— , 2011, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *Journal of Finance*, 66, 35–65.

Phillips, G., and J. Hoberg, 2010, "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies*, 23, 3773–3811.

Solomon, D., 2009, "Selective publicity and stock prices," working paper, USC.

Tetlock, P. C., 2007, "Giving content to investor sentiment: the role of media in the stock market," *Journal of Finance*, 62(3), 1139–1168.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, 2008, "More than words: quantifying language to measure firms' fundamentals," *Journal of Finance*, 63(3), 1437–1467.

Wall, L., T. Christiansen, and J. Orwant, 2000, *Programming Perl*. O'Reilly Media, Cambridge, Mass.

Program 1: Crawling the EDGAR masterfiles

```
# Program to download master files from EDGAR
# Copied almost verbatim from the LWP::UserAgent documentation

use LWP::UserAgent;
my $ua = LWP::UserAgent->new;
$ua->timeout(10);
$ua->env_proxy;

for($year=1993; $year<2012; $year=$year+1){
    for($i=1; $i<5; $i=$i+1){
  $quarter = "QTR" . $i;
  $filegrag = "ftp://ftp.sec.gov/edgar/full-index/" . $year . "/" . $quarter . "/master.gz"
      ;

  # This commands gets the file from EDGAR
  my $response = $ua->get($filegrag);

  # Now just pipe the output to a file named appropriately
  $filename = $year . $quarter . "master";
  open(MYOUTFILE, ">" . $filename);
  if ($response->is_success) {
      print MYOUTFILE $response->decoded_content;  # or whatever
  }
  else {
      die $response->status_line;
  }
  close(MYOUTFILE);
    }
}
```

Program 2: Creating a data and directory structure

```
##########################################################################
# This program grabs all the master files downloaded
# and pipes the content corresponding for each CIK into
# a new file.
##########################################################################

for($year=1993; $year<2012; $year=$year+1){
    for($i=1; $i<5; $i=$i+1){

        # First load each quarterly file
  $filea = $year . "QTR" . $i . "master";
  open(MYINFILE, $filea); #assumes master files are in root
  @linesgn = <MYINFILE>;
  close(MYINFILE);
  $sizegn = @linesgn;

  $numerox = sprintf("%03d", $numeroj); # format directory name
        $commandunix = "mkdir " . $numerox . "\n";
        system($commandunix); #creates directory 000

  # For each entry in the quarterly file
  for($j=0; $j<$sizegn; $j++)
  {
      # Pick only lines that correspond to EDGAR filings (.txt)
      if($linesgn[$j] =~ m/txt/){
    @arraydata = split(/\|/, $linesgn[$j]); # gets the data from each line
    $numero = sprintf("%07d", $arraydata[0]); # formats CIK number
    $filec = ">>" . $numerox . "/" . $numero . ".dat"; # The ">>" appends

    # Now just write into file
    open(MYOUTFILE, $filec); # opens file
    # To save file as csv
    $datagn=$linesgn[$j];
    $datagn =~ s/,/;/g;
    $datagn =~ s/\|/,/g;
    print MYOUTFILE $datagn;
    close(MYOUTFILE);
      }
      if($j> 500*($numeroj+1)){ # Each directory has 500 files
    $numeroj = $numeroj+1;
    $numerox = sprintf("%03d", $numeroj); # format directory name
                $commandunix = "mkdir " . $numerox . "\n";
                system($commandunix); #creates new directory
      }
  }
    }
}
```

Program 3: Reading the master files

```
################################################################################
# This program reads through the individual files from the QTRmaster EDGAR indexes.
# It counts the number of forms, as well as the number of particular types of
# forms (10-K, 8-K, etc). It also pipes the form fields (as strings) into another file
################################################################################

open(MYINFILE, "AllEDGARfiles.dat"); # Load file with list of .dat files
@linesgn = <MYINFILE>;
close(MYINFILE);
$sizegn = @linesgn; #Number of files

open(MYOUTFILE, ">statsEDGAR.dat");
open(MYOUTFILEB, ">formtypesEDGAR.dat");

for($i=0; $i<$sizegn; $i=$i+1){
    #For each dat file
    if($linesgn[$i] =~ m/dat/){
  $filename = $linesgn[$i];
  chomp($filename);

    #Initialize counters
  $tenk = $eightk = $fourk = $thirteeng = $thirteend = $fourteen = 0;
  $sixk = $tenq = $thirteenf = $code424b = 0;

    #Open the file with filing information (from master files) for each CIK
  open(MYINFILE, $filename);
  @datafile = <MYINFILE>;
  close(MYINFILE);

    #CIK code
  $cikcode = $filename;
  $cikcode =~ s/.dat//g;

    #Count file lines and different text strings
  $countlines = @datafile; #number of entries in each .dat file
  for($j=0; $j<$countlines; $j=$j+1){
      @arraydata = split(/,/, $datafile[$j]);
      print MYOUTFILEB $arraydata[2] . "\n";
      if($arraydata[2] == "4" || $arraydata[2] == "4\/A"){$fourk = $fourk+1;}
      if($arraydata[2] =~ /10K/ || $arraydata[2] =~ /10\-K/){$tenk = $tenk+1;}
      if($arraydata[2] =~ /10Q/ || $arraydata[2] =~ /10\-Q/){$tenq = $tenq+1;}
      if($arraydata[2] =~ /8K/ || $arraydata[2] =~ /8\-K/){$eightk = $eightk+1;}
      if($arraydata[2] =~ /6K/ || $arraydata[2] =~ /6\-K/){$sixk = $sixk+1;}
      if($arraydata[2] =~ /13F/){$thirteenf = $thirteenf+1;}
      if($arraydata[2] =~ /13G/){$thirteeng = $thirteeng+1;}
      if($arraydata[2] =~ /13D/){$thirteend = $thirteend+1;}
      if($arraydata[2] =~ /424B/){$code424b = $code424b+1;}
  }

  print MYOUTFILE "$filex,$countlines,$cikcode,$tenk,$tenq,$eightk,$sixk,$fourk,";
  print MYOUTFILE "$thirteeng,$thirteend,$thirteenf,$code424b\n";
    }
}
close(MYOUTFILE);
close(MYOUTFILEB);
```

Program 4: Crawling 8-K statements

```
use WWW::Mechanize;

######################################################################
# This program starts with the list of all CIKs with certain counts
# and crawls the 8-K statements
######################################################################

open(MYINFILE, "statsEDGAR.dat");
#print MYOUTFILE "$filex,$countlines,$cikcode,$tenk,$tenq,$eightk,$sixk,$fourk,";
#print MYOUTFILE "$thirteeng,$thirteend,$thirteenf,$code424b\n";
@linesgn = <MYINFILE>;
close(MYINFILE);
$sizegn = @linesgn;

for($i=0; $i<$sizegn; $i=$i+1){
    @arraydata = split(/,/, $linesgn[$i]);

    if($arraydata[5]>0){ #require CIK to have some 8K filed
  @arraydatab = split(/,/, $linesgn[$i]);

  # This points to where the file for a given CIK is in the file structure
  $filename = "../" . $arraydata[0];

  # Get CIK and create directory
  @arraydatac = split(/\//, $arraydatab[0]);
  $cik = $arraydatac[2];
  $cik =~ s/.dat//;
  $dirname = $arraydatac[1];
  $makemed = "mkdir " . $dirname . "/" . $cik;
  system($makemed); # creates directory

        # Reads in the cik.dat file
  open(MYINFILE, $filename);
  @datagn = <MYINFILE>;
  close(MYINFILE);
  $lenx = @datagn;

  # Crawl each occurence of a 13G filing
  for($j=0; $j<$lenx; $j=$j+1){
      @arraydata = split(/\,/, $datagn[$j]);
      if($arraydata[2] =~ m/8K/ || $arraydata[2] =~ /8\-K/){
                # Starts crawler, not checking for errors
    my $mech = WWW::Mechanize->new( autocheck => 0 );
    # Grabs address
    @arraydatad = split(/\//, $arraydata[4]);
    # Formats output file name
    $filenamea =  $dirname . "/" . $cik . "/" . $arraydatad[3];
    chomp($filenamea);
    # This is the file from the EDGAR archives
    $filecrawl = "http://www.sec.gov/Archives/" . $arraydata[4];
    # This crawls the file and saves it to the hard drive
    $mech->get($filecrawl, ':content_file' => $filenamea);
      }
  }
    }
}
```

Program 5: Reading 8-K statements

```
################################################################################
# This program reads through 8K statements and counts the number of occurrences
# of a given set of words
################################################################################

$start_run = time();

open(MYINFILE, "matched.8k");
@linesgn = <MYINFILE>;
close(MYINFILE);
$sizegn = @linesgn;

open(MYINFILE, "wordlist.dat");
@wordlisting = <MYINFILE>;
close(MYINFILE);
$lengthfile = @wordlisting;

open(MYOUTFILE, ">EightKwordlist.dat");

for($i=0; $i<$lengthfile; $i=$i+1){
    $litiword[$i] = $wordlisting[$i];
    chomp($litiword[$i]);
}

for($i=0; $i<$sizegn; $i=$i+1){
    chomp($linesgn[$i]);
    @arraydata = split(/,/, $linesgn[$i]);
    @arraydatab = split(/\//, $arraydata[0]);
    $dirname = $arraydatab[1];
    $cik = $arraydatab[2];
    $cik =~ s/.dat//;
    $filename1 = $dirname . "/" . $cik . "/" . $arraydata[1];

    open(MYINFILE, $filename1);
    @datafiling = <MYINFILE>;
    close(MYINFILE);
    $lendatafile = @datafiling;

    print MYOUTFILE "$dirname,$cik,$filename1";

    for($j=0; $j<$lengthfile; $j=$j+1){
  $countwords[$j] = 0;
  my $count = grep(/\b$litiword[$j]\b/ig, @datafiling);
  $countwords[$j] = $count;
  print MYOUTFILE ",$countwords[$j]";
    }

    print MYOUTFILE "\n";

}

my $end_run = time();
my $run_time = $end_run - $start_run;
print "\n\nJob took $run_time seconds\n\n";
```

Table 1: Most frequent EDGAR filing codes

The table presents the frequencies of appareances of different types of filings in the EDGAR database. The time period is 1993–2011. A filing is considered if and only if the same text string (i.e. "4") appears in the form field of the EDGAR master files.

Form	Form code	Frequency
Changes in ownership	4	4,028,202
Current report filing	8-K	1,030,605
5% passive ownership triggers ammendments	SC 13G/A	433,902
Quarterly report	10-Q	422,366
Initial ownership report	3	391,533
Definite materials	497	286,299
5% passive ownership triggers	SC 13G	278,735
Current report of foreign issuer	6-K	230,052
Change on a prospectus	424B3	188,880
5% active ownership triggers ammendments	SC 13D/A	165,010
Changes in ownership ammendments	4/A	150,442
Annual report on ownership changes	5	149,358
Annual report	10-K	128,566
Regulation D exemption, issuance	REGDEX	126,754
Quarterly holdings, institutional managers	13F-HR	126,016
Proxy statements	DEF 14A	125,634
Quarterly report, small business	10QSB	120,120
Registration of securities, investment companies	24F-2NT	116,126
Registration management investment companies	485BPOS	112,269
5% active ownership triggers	SC 13D	86,722

Table 2: EDGAR filers

The table presents the frequencies of appareances of different types of filings in the EDGAR database. The time period is 1993–2011. The second column presents the number of distint firms (CIKs) that file any version of the form type listed (i.e. both forms with "4" or "4/A" are counted). Columns 3–5 present the mean/median/maximum number of filings across all CIKs that file a given form.

Form type	Number of firms	Mean # of filings	Median # filings	Maximum # of filings
4	171,922	24.3	6	9,236
13-D	41,872	6.0	2	5,300
13-G	39,092	18.2	4	27,089
10-K	36,707	6.5	4	67
8-K	35,673	30.5	10	1,347
424B3	35,448	5.3	1	5,781
10-Q	27,708	23.4	18	157
14	22,027	11.8	8	290
497	20,127	19.6	4	2,618
485	13,512	11.0	5	670
X17A	7,935	6.9	7	20
13-F	7,788	23.6	19	278
424B2	7,720	8.8	1	5,239
NSAR	6,411	20.8	16	301
11-K	3,125	10.2	7	190
6-K	2,420	96.3	43	2,888

Table 3: Calendar effects on current report filings

The table presents the estimates of the models

$$\begin{aligned} X_t &= \alpha_x + \boldsymbol{\beta}_x \mathbf{D}_t + \boldsymbol{\gamma}_x \mathbf{M}_t + \boldsymbol{\eta}_x \mathbf{T}_t + \upsilon_t; \\ Y_t &= \alpha_y + \boldsymbol{\beta}_y \mathbf{D}_t + \boldsymbol{\gamma}_y \mathbf{M}_t + \boldsymbol{\eta}_y \mathbf{T}_t + \epsilon_t; \end{aligned}$$

where X_t denotes the number of 8-K statements filed with the SEC on a given date t, Y_t denotes the number of 8-K statements filed with the SEC that contain one of the two strings "departure of directors" and "election of directors." The variable $\mathbf{D}_t$ denotes a vector of day-of-the-week dummies (Monday being the omitted one), and $\mathbf{M}_t$ denotes month-of-the-year dummies. The vector $\mathbf{T}_t$ denotes time-trends controls, namely a linear term $t - 2004$, and a quadratic term $(t - 2004)^2$.

	Current reports (8-K)		**Current reports (8-K) on executive turnover**	
	Point Estimate	t-value	Point Estimate	t-value
A. Weekly effects				
Tuesday	52.9	6.9	2.9	1.8
Wednesday	41.8	5.5	−1.4	−0.9
Thursday	92.2	12.0	1.6	1.0
Friday	22.7	3.0	10.3	6.6
B. Monthly effects				
February	63.8	5.3	12.6	5.1
March	−4.3	−0.4	1.4	0.6
April	23.6	2.0	−6.4	−2.7
May	85.2	7.2	7.7	3.2
June	−55.5	−4.8	−7.6	−3.2
July	1.9	0.2	−16.2	−6.7
August	12.8	1.1	−15.8	−6.7
September	−59.2	−5.0	−17.3	−7.1
October	28.9	2.4	−13.7	−5.7
November	82.4	6.9	−5.3	−2.2
December	−24.4	−2.1	6.6	2.7
C. Time−effects				
Linear time term	−53.5	−8.8	21.3	17.1
Quadratic time term	2.8	4.4	−2.2	−16.8
Intercept	509.2	32.7	41.6	13.1
Adjusted R−squared	0.388		0.300	

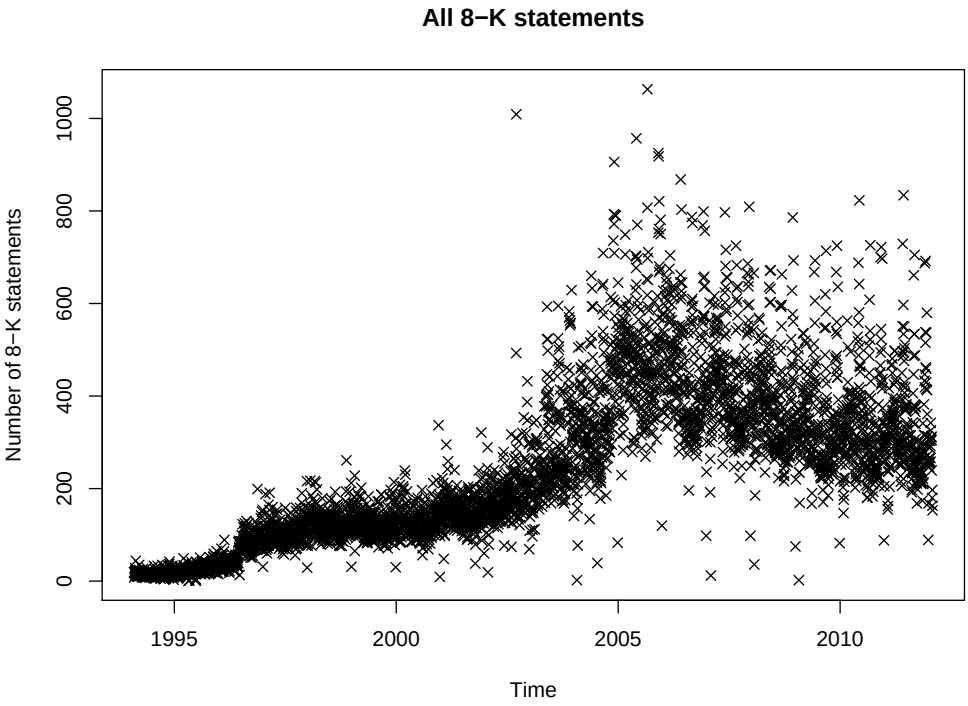

Figure 1: Plots of the number of unique CIKs that file a 8−K statement on a given day.

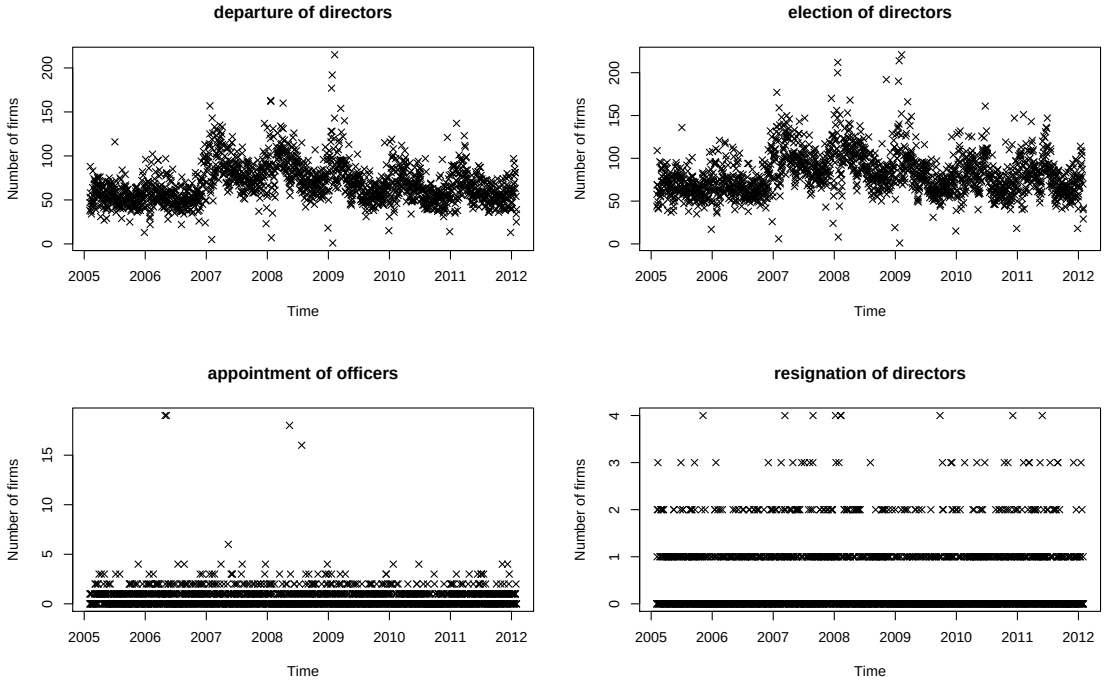

Figure 2: Plots of the number of unique CIKs that file a 8−K statement on a given day containing one the strings in the title (not case sensitive).