

Run EDGAR Run: SEC dissemination in a high-frequency world (Detailed data steps)

By Jonathan L. Rogers, Douglas J. Skinner, Sarah L. C. Zechman

Given our project is based in large part on a proprietary/confidential dataset and we view some of our code itself to be proprietary, we are complying with the “Data and Code Sharing Policy for the Journal of Accounting Research” by providing “a detailed step-by-step description of the code or the relevant parts of the code.” In an ideal world, we would be able to provide sufficient detail so that future scholars could replicate our exact results. For replication to be possible, these scholars would need to obtain the data from a PDS subscriber. Given there is variation when individual PDS feeds are received (for the same document), exact replication would not be possible without getting the data from our specific subscriber (although aggregate differences would likely be immaterial). In addition to the step-by-step description, we have also included all of the SEC filenames, which uniquely identifies each of the 4,782 observations in our final sample (see Table 1 of the paper).

1. We start with a proprietary dataset that contains a maximum of three lines of data for each SEC filing. Each line contains the date of the filing, the time that the data provider received the filing (to the 1,000th of a second), the EDGAR file name, and whether the filing arrived from the primary PDS feed, the secondary PDS feed, or on the SEC website (from their scrape). Virtually all filings were received three times, once from each source, resulting in three lines of data for most SEC filings. When the filings were scraped from the SEC the dataset also includes a form type (e.g., form 4, 8-K).
2. We hired a computer science undergraduate student to write python code to 1) download all EDGAR indices, 2) use those indices to identify and download all form 4 and form 4/A filings between March 2012 and December 2013 and 3) extract EDGAR filename, filing date, acceptance time, company CIK and company name, and insider CIK and insider name from the filing.
3. Next, we download all insider stock transactions (Table 1 from the Thompson-Reuters [TR] database – downloaded on 06/09/2014). We treat the Thompson database as the

“master file” and merge all additional variables into this master file. We restrict this master file (using TR variables names) as follows:

```
where secdate ge '01MAR2012'd and secdate le '31DEC2013'd
and rolecode1 in
("CB","D","DO","H","OD","VC","AC","CC","EC","FC","MC","SC",
"AV","CEO","CFO","CI","CO","CT","EVP","O","OB","OP","OS","OT","O
X","P","S","SVP","VP")
and trancode in ("P","S") and cleanse ne "S" and cleanse ne "A"
and tprice ge 0 and shares ge 0
and formtype = '4';
```

4. In order to retain only one observation per filing, we keep only the largest purchase/sale per DCN. These steps create Line 1 of Table 1 (in the paper).
5. The TR insider database does not contain CIK numbers. As a result, we use the COMPUSTAT fundq database (downloaded 06/10/2014) to retrieve CIK numbers and gvkeys (using a 9 digit CUSIP match). We then merge the database created from our downloaded EDGAR files (by CIK number). We then run the following routine to match insider names per TR to insider names extracted from downloaded EDGAR filings and only retain observation with match_type = '1-good'. These steps generate Line 2 of Table 1 (in the paper).

/*A. The tfn secdate must equal the data from the edgar date_time stamp

B. The name of the insider (per EDGAR parse) must be more like the name insider (per TR) than the name of the company (per TR)

C. If the score comparing insider name per TR (called owner) to the extracted insider name from EDGAR .txt files (called INSIDER_NAME1) is greater than 500 we interpret this as a bad name match. Testing suggests that this cutoff certainly excludes some valid matches but results in fairly few invalid matches.

D. In a few cases, it appears that our name extraction code picks up the company name instead of the insider's name (potentially due to errors in the actual filings). Therefore, we test whether the downloaded insider name (again

INSIDER_NAME1) is a closer match to the TR company name (than the TR insider name). If so, we also consider this to be a bad match.

*/

```
compged(uppercase(a.owner),uppercase(b.INSIDER_NAME1)) as  
name_match_score,
```

```
if time_sec = "" then match_type = '4-no match      ';
```

```
else if
```

```
compged(uppercase(owner),uppercase(INSIDER_NAME1))>compged(uppercase(cname),uppercase(INSIDER_NAME1)) and  
compged(uppercase(cname),uppercase(INSIDER_NAME1)) < 500 then  
match_type='3-backwards      ';
```

```
else if compged(uppercase(owner),uppercase(INSIDER_NAME1))> 500 then  
match_type='2-big name diff';
```

```
else match_type = '1-good';
```

```
proc sort; by id_tfn_combined name_match_score match_type;  
proc sort NODUPKEY; by id_tfn_combined;
```

6. For each observation, we then check whether there were any other trades filed by the same CUSIP within 15 minutes (before or after) and exclude an observations with such trades. Applying this restriction results in Line 3 of Table 1 (in the paper).
7. We then exclude observations that were posted to the EDGAR website prior to 9:40am eastern time or posted after 3:30pm eastern time. Applying this restriction results in Line 4 of Table 1 (in the paper).
8. To get Line 5 of Table 1 (in the paper), we require sufficient TAQ data to compute all of our TAQ based variables. The following steps underlie this requirement.
 - a. First, we create a TAQ-CRSP link table using the TCLINK macro written by Rabih Moussawi at WRDS.¹
 - b. Second, we use the TAQ-CRSP link table to merge TAQ Symbols into the Line 4 of Table 1 dataset (based on CUSIP). When a single CUSIP matches to multiple TAQ

¹ Available at https://wrds-web.wharton.upenn.edu/wrds/research/macros/sas_macros/tclink.sas (last accessed on 01/05/2017).

Symbols, we retain that match with the greatest trading volume (per CRSP). If we are unable to match to a valid TAQ symbol, the observation is dropped.

- c. Third, we use the daily TAQ Quote Files (CQ files) from WRDS, to create a national best bid and offer (NBBO) dataset using the NBBO.sas file created by Rabih Moussawi at WRDS.² We define market price as the simple average of the best bid and best offer at each second (note: the best bid or offer may have been issued at that second or may have been carried forward from a prior second). Market spread is then defined as the best offer minus the best bid.
 - d. We merge the results from step b to step c in a one to many match (resulting in market prices for each of the 121 seconds centered on the relevant time).
 - e. We use the daily TAQ trading datasets (CT datasets) to calculate trading volume measures. We start by merging the observations in step d with the CT files for the same date and second. Since multiple trades can be executed in the same second, we collapse the resulting dataset by summing over all trades in a given second. Seconds without any trading volume are set to zero volume. We then calculate cumulative trading volume for a given second X by summing over all the volume for seconds t-60 to X (inclusive).
 - f. In order to control for “normal volume”, we repeat steps d and e for each of the 52 weeks prior to the filing. We then average the resulting volume measures. Note that the % *Abnormal Volume* variable in the paper, uses the cumulative control volume for the entire 121 second window as a deflator. Whenever the deflator volume is zero, the observation is set to missing. These steps result in the sample on Line 5 of Table 1 (in the paper).
9. The sample from step f above is merged with CRSP using the TAQ-CRSP Link table to get the permno. Using the permno, we match this sample to the CRSP daily stock file (dsf), which was downloaded on 03/18/2014. From the dsf, we use the bidlo and askhi variables to define the daily trading price range. We then compare this range to the insider

² Available at (https://wrds-web.wharton.upenn.edu/wrds/research/applications/sas_files/nbbo.sas) (last accessed on 01/05/2017).

transaction price (per TR) and exclude any observation where the insider's price is not within the trading price range, resulting in the sample on Line 6 of Table 1 (in the paper). This is the main sample used throughout the paper, although certain tests have additional data restrictions resulting in smaller sample sizes.

10. Several of our tests require that we merge data from additional sources (not detailed above). We rely on the COMPUSTAT funda to get total assets (merging based on gvkey and acceptance year). In addition, we capture media coverage by merging the Line 6 sample with the RavenPack Dow Jones (DJ) archive. Specifically, we take the RavenPack DJ story archive and merge in permno from the TAQ-CRSP link table (using a CUSIP match). DJ coverage is then defined as stories with a relevance score of 100, a category equal to either insider-buy or insider-sell, and a story time within 15 minutes after first_dissemination.