

```
start_time <- Sys.time()
cat("===== START TIME =====")

## ===== START TIME =====

print(.start_time)

## [1] "2026-02-02 21:22:38 PST"

library(RPostgres)
library(dplyr)

##

## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##   filter, lag

## The following objects are masked from 'package:base':
##   intersect, setdiff, setequal, union

library(data.table)

##

## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##   between, first, last

library(tibble)
library(stringr)
library(farr)

##

## Attaching package: 'farr'

## The following object is masked from 'package:base':
##   truncate

library(lubridate)

##

## Attaching package: 'lubridate'

## The following objects are masked from 'package:data.table':
##   hour, isoweek, nday, minute, month, quarter, second, wday, week,
##   yday, year

## The following objects are masked from 'package:base':
##   date, intersect, setdiff, union

library(ggplot2)
library(purrr)

##

## Attaching package: 'purrr'

## The following object is masked from 'package:data.table':
##   transpose

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
##   forcats 1.0.0     ✓ tidyr  1.3.0
##   readr  2.1.5

## -- Conflicts ----- tidyverse_conflicts() -----
## * data.table::between() masks dplyr::between()
## * dplyr::filter() masks stats::filter()
## * data.table::first() masks dplyr::first()
## * lubridate::hour() masks data.table::hour()
## * lubridate::isoweek() masks data.table::isoweek()
## * dplyr::last() masks stats::lag()
## * data.table::last() masks dplyr::last()
## * lubridate::nday() masks data.table::nday()
## * lubridate::minute() masks data.table::minute()
## * lubridate::month() masks data.table::month()
## * lubridate::quarter() masks data.table::quarter()
## * lubridate::second() masks data.table::second()
## * purrr::compose() masks data.table::tcompose()
## * lubridate::wday() masks data.table::wday()
## * lubridate::week() masks data.table::week()
## * lubridate::yday() masks data.table::yday()
## * lubridate::year() masks data.table::year()
## * Use the conflicted package <http://conflicted.r-lib.org/> to force all conflicts to become errors

library(hmisc)
library(haven)
library(fixest)
library(roadkill)
theme_set(theme_bw())

path = "~/Users/gliangianli/Dropbox/RDB_code/"

sdc = readRDS(paste0(path, "1_data_processing/data/sample_selection/sdc/SDC_IPD_2008_2025.RDS")) %>%
  filter(is.na('Dates: Filing Date') | 'Dates: Filing Date' >= "2008-01-01")

# ipo_files = sdc %>% filter('SDC File Form (Number)' != 0) %>% dplyr::select('SDC File Form (Number)') %>% distinct()
# write_csv(ipo_files, paste0(path, "1_data_processing/data/sample_selection/sdc/SDC_files.csv"))

sdc_cik = read_csv(paste0(path, "1_data_processing/data/sample_selection/sdc/SDC_files_EDGAR.csv"))

## Row names:
## Rows: 29257 Columns: 12
## --- Column specification
##   (3): cik, sl_filing_name, fw_filing_name dbl (6): ..., SEC.File.Form.Number.,
##   Number of filings, sl, fw, EFFECT date (3): sl_filing_date, fw_filing_date,
##   EFFECT_filing_date
## 1 Use `spec()` to retrieve the full column specification for this data. 1
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## " " -> "...1"
```

```
sdc = sdc %>% left_join(sdc_cik %>% mutate(SEC.File.Form.Number. = as.character(SEC.File.Form.Number.)), by = c("SEC File Form (Number)" = "SEC.File.Form..Number."))

# nrow(sdc) # 975,833

sdc_ipo_yes = sdc %>% filter('IPO Flag' == "True") # 56,731
sdc_ipo_no = sdc %>% filter(is.na('IPO Flag') | 'IPO Flag' != "True") # 919,102

# index from WRDS SEC Analytics Suite
sl_since_2010 = readRDS(paste0(path, "1_data_processing/data/filing_index/20250615/reference_table_sl_20250615.RDS")) %>%
  filter(date == "2010-01-01") %>%
  group_by(cik) %>% mutate(first_filing_date_since_2010 = min(fdate)) %>%
  mutate(not_first_since_2010 = ifelse(fdate == first_filing_date_since_2010, 0, 1)) %>% ungroup() # multiple S-is in one day - make them all first

# nrow(sl_since_2010) # 16,690

# keep those can be mapped to sdc_ipo_yes (or filter out those with filing_name map to sdc_ipo_no - this would give us larger sample)
sl = sl_since_2010 %>% mutate(filing_name = sub("^/.*\\.txt$", "\\1", fname)) %>%
  filter(filing_name != sdc_ipo_yes$sl_filing_name)

# nrow(sl) # 6,013

# link from WRDS
link = readRDS(paste0(path, "1_data_processing/data/sample_selection/wrds/RDS_gvkey_cik_link_20250617.RDS"))

# msc from WRDS
msc = readRDS(paste0(path, "1_data_processing/data/sample_selection/wrds/msc_msc_20250617.RDS"))

# ccm from WRDS
ccm = readRDS(paste0(path, "1_data_processing/data/sample_selection/wrds/RDS_ccm_20250617.RDS"))
```

1. map to SDC & CRSP/Compustat

```
cik_gvkey_permno = sl %>% inner_join(link, by = "cik") %>% filter(fdate >= sec_start_date, fdate <= sec_end_date) %>%
  inner_join(ccm %>% inner_join(msc %>% group_by(permno) %>%
    mutate(min_namedt = min(namedt), max_namedt = max(namedt)) %>% slice(1) %>% ungroup()) %>%
    dplyr::select(permno, min_namedt, max_namedt, shred),
    by = c("permno" = "permno")),
    by = "gvkey") %>%
  filter(fdate < min_namedt) %>% # sl must be before min_namedt
  group_by(fname) %>% mutate(n = n(), m = length(unique(flag))) %>%
  filter(n == 1 | (n > 1 & m == 1 & flag == 3) | (n > 1 & m == 1)) %>% ungroup() %>% keep higher rank linking
  group_by(cik) %>% filter(fdate == max(fdate)) %>% ungroup() # keep the latest filing for each cik

## Warning in inner_join(., link, by = "cik"): Detected an unexpected many-to-many relationship between 'x' and 'y'.
## 1 Row 3 of 'x' matches multiple rows in 'y'.
## 1 Row 4363 of 'y' matches multiple rows in 'x'.
## 1 If a many-to-many relationship is expected, set 'relationship =
## "many-to-many" to silence this warning.

## Warning in inner_join(., ccm %>% inner_join(msc %>% group_by(permno) %>% Detected an unexpected many-to-many relationship between 'x' and 'y'.
## 1 Row 71 of 'x' matches multiple rows in 'y'.
## 1 Row 4539 of 'y' matches multiple rows in 'x'.
## 1 If a many-to-many relationship is expected, set 'relationship =
## "many-to-many" to silence this warning.

cik_gvkey_permno_unique_linked = cik_gvkey_permno %>% group_by(gvkey) %>%
  mutate(o = length(unique(cik)), p = length(unique(lpermno))) %>% filter(o == 1, p == 1) %>% slice(1) %>% ungroup() %>%
  # one permno, multiple gvkey
  group_by(lpermno) %>% mutate(q = length(unique(gvkey))) %>% filter(q == 1) | (linkdt == min(linkdt)) %>% ungroup() %>%
  dplyr::select(fname, fdate, cik, gvkey, lpermno, first_filing_date_since_2010, not_first_since_2010, filing_name) %>%
  mutate(link = 1)

cik_gvkey_permno_non_unique = cik_gvkey_permno %>% group_by(gvkey) %>%
  mutate(o = length(unique(cik)), p = length(unique(lpermno))) %>% filter(o != 1 | p != 1) %>% ungroup()

cik_gvkey_permno_non_unique_linked_1 = cik_gvkey_permno_non_unique %>% filter(link_end_date >= min_namedt, link_end_date <= max_namedt) %>%
  # link_end_date must be between the date in crsp
  group_by(gvkey) %>% mutate(o = length(unique(cik)), p = length(unique(lpermno))) %>% filter(o == 1, p == 1) %>% slice(1) %>% ungroup() %>%
  dplyr::select(fname, fdate, cik, gvkey, lpermno, first_filing_date_since_2010, not_first_since_2010, filing_name) %>%
  mutate(link = 2)

cik_gvkey_permno_non_unique_linked_2 = cik_gvkey_permno_non_unique %>% filter(lgvkey != cik_gvkey_permno_non_unique_linked_1$gvkey) %>%
  group_by(gvkey) %>% filter(fdate == max(fdate)) %>% # keep 1 share
  filter(min_namedt == min(min_namedt)) %>% mutate(o = length(unique(cik)), p = length(unique(lpermno))) %>%
  filter(o == 1, p == 1) %>% slice(1) %>% ungroup() %>%
  dplyr::select(fname, fdate, cik, gvkey, lpermno, first_filing_date_since_2010, not_first_since_2010, filing_name) %>%
  mutate(link = 3)

cik_gvkey_permno_non_unique_linked_3 = cik_gvkey_permno_non_unique %>% filter(lgvkey != cik_gvkey_permno_non_unique_linked_1$gvkey) %>%
  filter(lpermno != cik_gvkey_permno_non_unique_linked_2$gvkey) %>%
  mutate(o = length(unique(cik)), p = length(unique(lpermno))) %>% filter(o == 1, p == 1) %>% slice(1) %>% ungroup() %>%
  dplyr::select(fname, fdate, cik, gvkey, lpermno, first_filing_date_since_2010, not_first_since_2010, filing_name) %>%
  mutate(link = 4)

cik_gvkey_permno_linked = rbind(cik_gvkey_permno_unique_linked, cik_gvkey_permno_non_unique_linked_1,
                                cik_gvkey_permno_non_unique_linked_2, cik_gvkey_permno_non_unique_linked_3)

# three fname matches with two gvkey - just drop for now
cik_gvkey_permno_linked = cik_gvkey_permno_linked %>% group_by(fname) %>% mutate(n = n()) %>% filter(n == 1) %>% dplyr::select(-n) %>% ungroup()

cik_gvkey_permno_linked = cik_gvkey_permno_linked %>%
  filter(fname != c("data/095126/0001193125-11-184110.txt", # choose permno = 13077, not 20583
    "edgar/data/1326428/0001047469-12-000036.txt", # choose permno = 13665, not 16488
    "edgar/data/1326428/0001193125-12-281472.txt")) # choose permno = 13647, not 16246

# drop those cross the treatment window (n = 13)
cik_gvkey_permno_linked = cik_gvkey_permno_linked %>% filter((first_filing_date_since_2010 < "2022-11-30" & fdate >= "2022-11-30"))

saveRDS(cik_gvkey_permno_linked %>% filter(fdate <= "2024-12-31"), paste0(path, "1_data_processing/data/sample_selection/final_sample/1_sl_map_to_sdc_crsp_compustat.RDS"))

# nrow(cik_gvkey_permno_linked %>% filter(fdate <= "2024-12-31")) # 3,257
```

2. map to SDC but not CRSP/Compustat

```
sl_other_in_sdc = sl %>% filter(ifname != cik_gvkey_permno_linked$fname) %>%
  filter(filing_name != cik_gvkey_permno_linked$filing_name) %>%
  filter((first_filing_date_since_2010 < "2022-11-30" & fdate >= "2022-11-30"))

# nrow(sl_other_in_sdc) # 2,689

unique_sl_other_in_sdc = sl_other_in_sdc %>% dplyr::select(filing_name) %>% distinct()

# nrow(unique_sl_other_in_sdc) # 2,689

saveRDS(sl_other_in_sdc %>% filter(fdate <= "2024-12-31"), paste0(path, "1_data_processing/data/sample_selection/final_sample/2_sl_map_to_sdc_but_not_crsp_compustat.RDS"))

# nrow(sl_other_in_sdc %>% filter(fdate <= "2024-12-31")) # 2,627
```

3. other

```
# get cik network
list_of_linked_cik_filing_name_pair = rbind(cik_gvkey_permno_linked %>% dplyr::select(cik, filing_name) %>% distinct(),
      sl_other_in_sdc %>% dplyr::select(cik, filing_name) %>% distinct()) %>% distinct()

forms_sl_since_1994 = readRDS(paste0(path, "1_data_processing/data/filing_index/20250615/reference_table_sl_since_1994_20250615.RDS"))

# in the full sample since 1994
list_of_cik_filing_name_pair_in_full = forms_sl_since_1994 %>% mutate(filing_name = sub("^/.*\\.txt$", "\\1", fname)) %>% dplyr::select(cik, filing_name) %>% distinct()

cik_related = list_of_cik_filing_name_pair_in_full %>% left_join(list_of_cik_filing_name_pair_in_full, by = "filing_name") %>% filter(cik.x != cik.y) %>%
  dplyr::select(cik.x, cik.y) %>%
  rename("cik" = "cik.x",
         "cik_related" = "cik.y") %>%
  filter(cik != cik_related) %>%
  filter(cik != list_of_linked_cik_filing_name_pair$cik) # 762 related CIK

## Warning in left_join(., list_of_cik_filing_name_pair_in_full, by = "filing_name"): Detected an unexpected many-to-many relationship between 'x' and 'y'.
## 1 Row 2 of 'x' matches multiple rows in 'y'.
## 1 Row 26 of 'y' matches multiple rows in 'x'.
## 1 If a many-to-many relationship is expected, set 'relationship =
## "many-to-many" to silence this warning.

# 1. focus in the sample of very initial S-1 filing for each CIK
sl_other = sl_since_2010 %>% filter(not(first_filing_date_since_2010 == 0)) %>% # 9,727
  mutate(filing_name = sub("^/.*\\.txt$", "\\1", fname)) %>%
  filter(ifname != cik_gvkey_permno_linked$fname, sl_other_in_sdc$fname) %>% # 4,073
  filter(filing_name != cik_gvkey_permno_linked$filing_name, sl_other_in_sdc$filing_name) %>% # 4,063
  filter(cik != cik_gvkey_permno_linked$cik, sl_other_in_sdc$cik) %>% # 2,438
  filter(cik != cik_related$cik_related) %>% # 2,116
  group_by(cik) %>% filter(recession == min(recession)) %>% ungroup() # 2,070

# 2. exclude those dropped during sample selection because they are dupe
cik_gvkey_pair = link %>% dplyr::select(cik, gvkey) %>% distinct()
gvkey_permno_pair = ccm %>% dplyr::select(gvkey, lpermno) %>% distinct()
permno_min_namedt = msc %>% group_by(permno) %>% mutate(min_namedt = min(namedt)) %>% slice(1) %>% ungroup() %>% dplyr::select(permno, min_namedt)

sl_other_dup = sl_other %>% left_join(cik_gvkey_pair, by = "cik") %>%
  left_join(gvkey_permno_pair, by = "gvkey") %>%
  left_join(permno_min_namedt, by = c("lpermno" = "permno")) %>%
  filter(gvkey != cik_gvkey_permno_linked$gvkey) | (lpermno != cik_gvkey_permno_linked$lpermno) | (fdate > min_namedt) %>%
  dplyr::select(fdate,filing_name) %>% distinct() # 326

## Warning in left_join(., gvkey_permno_pair, by = "gvkey"): Detected an unexpected many-to-many relationship between 'x' and 'y'.
## 1 Row 58 of 'x' matches multiple rows in 'y'.
## 1 Row 811 of 'y' matches multiple rows in 'x'.
## 1 If a many-to-many relationship is expected, set 'relationship =
## "many-to-many" to silence this warning.

sl_other = sl_other %>% anti_join(sl_other_dup, by = c("fname")) # 1,744

# 3. add an indicator to those filings not the first S-1 filings (i.e., first is before 2010) drop?
sl_first_cik_since_1994 = forms_sl_since_1994 %>% group_by(cik) %>%
  mutate(min_fdate_since_1994 = min(fdate)) %>% slice(1) %>% dplyr::select(cik, min_fdate_since_1994)

sl_other = sl_other %>% inner_join(sl_first_cik_since_1994, by = c("cik")) %>%
  mutate(first_filing_date_since_1994 = ifelse(fdate == min_fdate_since_1994, 1, 0))

sl_other = sl_other %>% filter(first_filing_date_since_1994 == 1) # 1,456

unique_sl_other = sl_other %>% dplyr::select(filing_name) %>% distinct()

# nrow(unique_sl_other) # 1,103

# doesn't matter which one to drop, we only care about the underlying filings
sl_other = sl_other %>% filter(fdate <= "2024-12-31") %>%
  # use the first filing date
  group_by(filing_name) %>% arrange(fdate, by_group = TRUE) %>% slice(1) %>% ungroup()

saveRDS(sl_other, paste0(path, "1_data_processing/data/sample_selection/final_sample/3_sl_not_sdc_crsp_compustat.RDS"))

# nrow(sl_other) # 1,011
```

merge together

```
sl_sections = readRDS(paste0(path, "1_data_processing/data/sample_selection/other/sl_with_sections_identified.RDS"))

group_1 = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/1_sl_map_to_sdc_crsp_compustat.RDS"))
group_2 = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/2_sl_map_to_sdc_but_not_crsp_compustat.RDS"))
group_3 = readRDS(paste0(path, "1_data_processing/data/sample_selection/final_sample/3_sl_not_sdc_crsp_compustat.RDS"))

df = rbind(group_1 %>% filter(filing_name != sl_sections$filing_name) %>% mutate(group = 1) %>% dplyr::select(fname, filing_name, cik, fdate, group),
            group_2 %>% filter(filing_name != sl_sections$filing_name) %>% mutate(group = 2) %>% dplyr::select(fname, filing_name, cik, fdate, group),
            group_3 %>% filter(filing_name != sl_sections$filing_name) %>% mutate(group = 3) %>% dplyr::select(fname, filing_name, cik, fdate, group))

# table(df$group)
#   1     2     3
# 3135 1599  686

saveRDS(df, paste0(path, "1_data_processing/data/sample_selection/final_sample/sample_sl.RDS"))

# nrow(df) # 5,420

.end_time <- Sys.time()
cat("===== END TIME =====")

## ===== END TIME =====

print(.end_time)

## [1] "2026-02-02 21:23:11 PST"

cat("Elapsed:", difftime(.end_time, .start_time, units = "secs"), "seconds")

## Elapsed: 33.459 seconds
```