Preliminary Trademark Rebrand Report

Aniket Kesari

May 30, 2017

Setup

Description of Data Source

I pulled data from the United States Patent and Trademark Office’s Open Data website (https://www.uspto.gov/learning-and-resources/electronic-data-products/trademark-case-files-dataset-0). The database contains all trademark applications from 1884-2017. I included the data sources I use here, as well as the relevant documentation in the Dropbox folder. The main source has several different data sources, and I pull from three: “Case Files,” “Prior Mark Files,” and “Owner Files.” I triangulate between these three files by matching them on the “serial_no” (serial number) variable that identifies each unique trademark.

Case File

The “case_file.csv” file contains the full data for all trademark applications, including the dates they were filed and registered. There is also information about how UPSTO handled the case (can recover this through the ’event" markers). This file is important because it contains the filing dates that motivate the project.

Prior Marks File

The prior marks file contains information regarding whether a trademark registration had prior registrations. According to the documentation, USPTO recommends that an entity filing for a trademark notes whether they are claiming prior ownership over the same or similar trademark, which provides a good basis for understanding whether the new trademark is a rebranding of an old one. However, this step is not explicitly required, so there may be missing data. The dataset contains a list of prior registration numbers for each serial number.

Owner File

The Owwner File contains information about the owners of trademarks, including their addresses, names, etc. For this analysis, I used the “legal entity” variable to subset into corporate owners of trademarks.

Load in the Data

The major issue with these data sources is that they are *very* large. The Case File contains approximately 8 million observations and 120 variables, meaning there are close to 1 billion cells in the dataset. I wasn’t able to load the full dataset, so I subsetted both the Case File and Owner File to the first 2 million observations, which takes us up to the year 1989 (indicating that nearly 6 million registrations came in the ~30 years since then). Here I load in all of the data:

```r
# Load necessary packages
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
# Set working directory
setwd("C:/Users/Owner/Dropbox/Trademark Rebrand Project")
# The main file is way too big to read in, subsetting to 2 million observations
# Load each data set, subset case and owner files to first 2 million observations
case_file <- read.csv("C:/Users/Owner/Dropbox/Trademark Rebrand Project/case_file.csv/case_file.csv", nr
owner_file <- read.csv("C:/Users/Owner/Dropbox/Trademark Rebrand Project/owner.csv/owner.csv", nrow=2000
prior_mark <- read.csv("C:/Users/Owner/Dropbox/Trademark Rebrand Project/prior_mark.csv/prior_mark.csv")
```

Number of Prior Registrations

Here I simply count the number of unique serial numbers in the Prior Mark data. We have approximately 1 million trademarks that have at some point, had a prior mark.

```
# Count number of unique ID's in the Prior Mark data
prior_regs <- prior_mark %>%
  summarise(n_distinct(serial_no))
print(prior_regs)
```

```
##   n_distinct(serial_no)
## 1               1005198
```

Clean Data

Used this serial_no ID to test cleaning throughout: 73609715

Here, I clean the data. The code below has markup to explain what I did at each step:

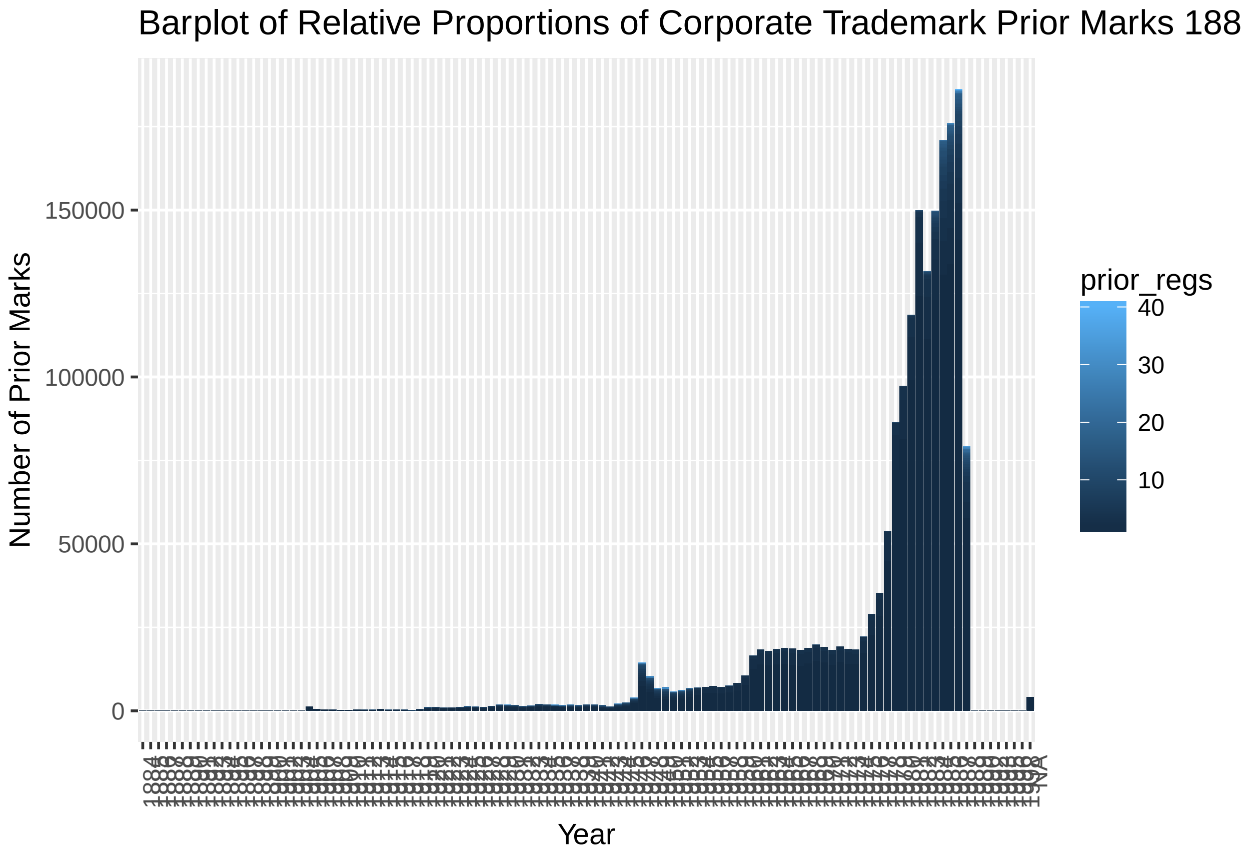
```
# Create a new dataset
new_file <- case_file %>% # Start with the Case File
              # Select variables of interest
              select(serial_no, ir_auto_reg_dt, filing_dt, registration_dt, related_other_in) %>%
              # Merge Prior Mark data in, matching by "serial number" and retain all values
              full_join(prior_mark, by="serial_no") %>%
              # Merge Owner data in, matching by "serial number" and retain all values
              full_join(owner_file, by="serial_no") %>%
              # Subset to corporations only (legal entity ID = 3)
              filter(own_entity_cd == 3) %>%
              group_by(serial_no) %>% # Group observations by serial number
              # Add a new column with the number of prior marks for each serial number
              mutate(prior_regs = n_distinct(prior_no))
new_file$filing_dt <- as.Date(new_file$filing_dt, format="%Y-%m-%d") # Convert filing date variable to
new_file$registration_dt <- as.Date(new_file$registration_dt, format="%Y-%m-%d") # Convert registration
```

```
new_file$Year <- format(new_file$filing_dt, "%Y") # Extract Year from filing date

# Write a csv with this clean data
write.csv(new_file, "Trademark Rebrands 1884-1989.csv")
```

Plotting

This preliminary plot is a barplot that shows the proportion of prior marks in each year

```
PriorRegBarPlot <- new_file %>%
                              group_by(Year, prior_regs) %>%
                              summarise(n=n()) %>%
                              ggplot() + aes(x = Year, y = n, fill = prior_regs) + geom_bar(stat="identit

PriorRegBarPlot
```



Interpretation: The number of prior marks dramatically increased over the course of the 100-year timespan here. While there isn't much evidence that the frequency of new marks increased dramatically (this is indicated by the light blue), we need more data from the next 30 years. Further, I suspect if we subset further down to owners with more than 1 trademark, we'll see better results.

Steps from here

1. *Figure out how to load in the full dataset*: The current dataset crashes my computer if I try to load it - I may have to offload the initial load to a cluster and then subset it.
2. *Clean the Data*: There are lots of duplicate observations because of inconsistent record keeping, I'll need to clean more to get better results
3. *More Plots*: The barplot lends support for the idea that the rate of trademark filings has increased rapidly in recent years. Out of 130 years on record, the first 100 years only make up about 25% of *all* recorded trademarks, and I suspect we'll see more rebranding in the 1990s onward. A scatterplot showing the linear relationship may also be better at that stage.
4. *More Relationships*: We also have information about owners' geography, names, etc., so we may be able to do some more interesting analysis with less noisy data. I also need to check to see how to match "prior_no" from Prior Marks to "serial_no" to figure out the frequency of rebranding for each individual trademark (right now the "prior_no" doesn't seem to correspond to a modern "serial_no").
5. *Modeling*: It may be worth checking if there were changes in legal regimes or other reasons that caused any spikes in trademark filings.