2023 Organizational network analysis group project

Project overview

Context

The U.S. Patent and Trademark Office (USPTO) is one of the most important organizations for supporting innovation and economic growth. The USPTO employs over 10,000 patent examiners whose primary task is to assess inventions for patentability and issue patents when appropriate.

Inventors submit patent applications, pay the associated fees and receive a decision from a patent examiner. If all of the inventor's claims made in a patent application are allowed by the examiner—i.e., determined to be novel and non-obvious in light of prior inventions—the examiner issues a patent.

If some or all of the claims are rejected as not meeting the patentability criteria, the examiner issues a formal rejection of the application, which presents the applicant with a choice: abandon the application and stop the process, or continue the process by revising the application and resubmitting it to the agency for another round of consideration. The latter option is associated with additional fees, but is virtually always open: the applicant can continue to reapply with no limit on the number of attempts. (This process is described in more detail here.)

The challenge

Since innovation is dynamic and because patent rights can be highly valuable for inventors, the timing of the USPTO's decisions on patent applications is a sensitive issue for inventors. The USPTO has faced some pressure from policymakers and regulators to address the problem of long review times. The agency is interested in understanding what affects the examination time—the time from the date the application is filed until a definitive decision is made. Understanding the causes and the variation in the examination time can help reduce the backlog.

Inequities among examiners

The agency suspects—and has some indirect evidence—that examiners' work, mobility across units, promotion across pay grades and attrition differ systematically by demographic characteristics. Of special concern are patterns related to examiner's gender and race/ethnicity, not only because such differences may pose obvious ethical challenges, but also because they increase the risk of legal liability for the agency. The USPTO would like to understand what, if any, patterns in examiners' work, mobility, promotion and attrition vary systematically by race and/or gender.

Team project

Your team's goal is to perform the analysis of the agency's data as a consulting team would. Base on the analysis, you will write a short report (no longer than 10 pages including all figures table and references, if any). Your analysis will focus on one or more of the following questions:

- What are the organizational and social factors associated with the length of patent application prosecution?
- What is the role of network structure here?
- What is the role of race and ethnicity in the processes described in the questions above?

Your report should offer convincing analysis, as well as projections and related recommendations.

You can and should start working on the project as soon as your group is formed. Our weekly exercises will help you build the analysis strategy and move you toward the project's goal.

Data

The data on examiners and patent applications come from various public sources. I have modified some of the source files to make them smaller and more manageable, but feel free to use the original sources if you feel compelled.

Small data files are in the *CSV* format, larger ones are in *parquet* format to reduce size. To read *CSV* files, use `readr` package that is part of the `tidyverse` package. To read *parquet* files, use `arrow` package.

Other packages I recommend for working with the data:

- `lubridate` for working with dates
- `stringr` for working with character strings
- `skimr` for quick summaries of data columns
- `gender` for inferring gender from name (there are other options)
- `wru` for predicting race/ethnicity based on name

Patent applications

The largest data file is on patent applications. It comes from the USPTO 2016 PatEx dataset. The full data file takes between 0.5Gb and 1Gb of memory, depending on how you load it, so I have created a smaller subsample of the data for your convenience, using the following restrictions:

- The year of `filing_date` is 2000 or later
- Only 4 of the 9 Technology Centers comprising the agency are included: 1600, 1700, 2100 and 2400
- Only 15 relevant variables are included (see PatEx data dictionary)

```
app_data_sample %>%
  tbl_vars()
```

```
## <dplyr:::vars>
##  [1] "application_number"   "filing_date"          "examiner_name_last"
##  [4] "examiner_name_first"  "examiner_name_middle" "examiner_id"
##  [7] "examiner_art_unit"    "uspc_class"           "uspc_subclass"
## [10] "patent_number"        "patent_issue_date"    "abandon_date"
## [13] "disposal_type"        "appl_status_code"     "appl_status_date"
## [16] "tc"
```