e3

2024-04-08

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(gender)
library(wru)
```

```
##
## Please cite as:
##
## Khanna K, Bertelsen B, Olivella S, Rosenman E, Rossell Hayes A, Imai K
## (2024). _wru: Who are You? Bayesian Prediction of Racial Category Using
## Surname, First Name, Middle Name, and Geolocation_. R package version
## 3.0.1, <https://CRAN.R-project.org/package=wru>.
##
## Note that wru 2.0.0 uses 2020 census data by default.
## Use the argument ‘year = "2010"‘, to replicate analyses produced with earlier package versions.
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
##
```

```
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
## The following object is masked from 'package:tibble':
##
##     as_data_frame
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```
library(ggraph)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(arrow)
```

```
##
## Attaching package: 'arrow'
##
## The following object is masked from 'package:lubridate':
##
##     duration
##
## The following object is masked from 'package:utils':
##
##     timestamp
```

```
library(readr)
library(dplyr)
library(tidyr)
```

Data loading and preprocessing

```
data_path <- "C:/Users/chens/Desktop/orgb/"
applications <- read_parquet(paste0(data_path,"app_data_sample.parquet"))
edges <- read_csv(paste0(data_path,"edges_sample.csv"))
```

```
## Rows: 32906 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): application_number
## dbl  (2): ego_examiner_id, alter_examiner_id
## date (1): advice_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

demographics

```
# Gender

examiner_names=applications %>% distinct(examiner_name_first)
examiner_names_gender=examiner_names %>%
  do(results = gender(.$examiner_name_first, method = "ssa")) %>%
  unnest(cols = c(results), keep_empty = TRUE) %>%
  select(examiner_name_first = name, gender, proportion_female)

# Join gender data back
applications=applications %>%
  left_join(examiner_names_gender, by = "examiner_name_first")
```

```
# Race

examiner_surnames=applications %>% select(surname = examiner_name_last) %>% distinct()
examiner_race=predict_race(voter.file = examiner_surnames, surname.only = T) %>% as_tibble()
```

```
## Predicting race for 2020

## Warning: Unknown or uninitialised column: `state`.

## Proceeding with last name predictions...

## i All local files already up-to-date!

## 701 (18.4%) individuals' last names were not matched.
```

```
examiner_race=examiner_race %>%
  mutate(max_race_p = pmax(pred.asi, pred.bla, pred.his, pred.oth, pred.whi)) %>%
  mutate(race = case_when(
    max_race_p == pred.asi ~ "Asian",
    max_race_p == pred.bla ~ "black",
    max_race_p == pred.his ~ "Hispanic",
    max_race_p == pred.oth ~ "other",
    max_race_p == pred.whi ~ "white",
    TRUE ~ NA_character_
  ))

# Join race data back
applications=applications %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
```

```r
# Tenure
examiner_dates=applications %>%
  select(examiner_id, filing_date, appl_status_date)
examiner_dates=examiner_dates %>%
  mutate(start_date = ymd(filing_date), end_date = as_date(dmy_hms(appl_status_date)))
examiner_dates=examiner_dates %>%
  group_by(examiner_id) %>%
  summarise(
    earliest_date = min(start_date, na.rm = TRUE),
    latest_date = max(end_date, na.rm = TRUE),
    tenure_days = interval(earliest_date, latest_date) %/% days(1)
  ) %>%
  filter(year(latest_date)<2018)

# Join tenure data back
applications=applications %>% left_join(examiner_dates, by = "examiner_id")
```

processing time

```r
applications=applications %>%
  mutate(
    final_decision_date = coalesce(patent_issue_date, abandon_date),
    app_proc_time = as.numeric(difftime(final_decision_date, filing_date, units = "days"))
  )
```

Centrality measures

```r
unique_examiner_ids=unique(c(edges$ego_examiner_id, edges$alter_examiner_id))

g=graph_from_data_frame(edges[, c("ego_examiner_id", "alter_examiner_id")], directed = TRUE, vertices =
```

```
## Warning in graph_from_data_frame(edges[, c("ego_examiner_id",
## "alter_examiner_id")], : In ‘d’ ‘NA’ elements were replaced with string "NA"
```

```
## Warning in graph_from_data_frame(edges[, c("ego_examiner_id",
## "alter_examiner_id")], : In ‘vertices[,1]’ ‘NA’ elements were replaced with
## string "NA"
```

```r
centrality_entire=data.frame(
  examiner_id = V(g)$name,
  degree_centrality = degree(g, mode = "out"),
  betweenness_centrality = betweenness(g, directed = TRUE),
  closeness_centrality = closeness(g, mode = "out")
)

centrality_entire$examiner_id=as.numeric(centrality_entire$examiner_id)
```

```
## Warning: NAs introduced by coercion
```

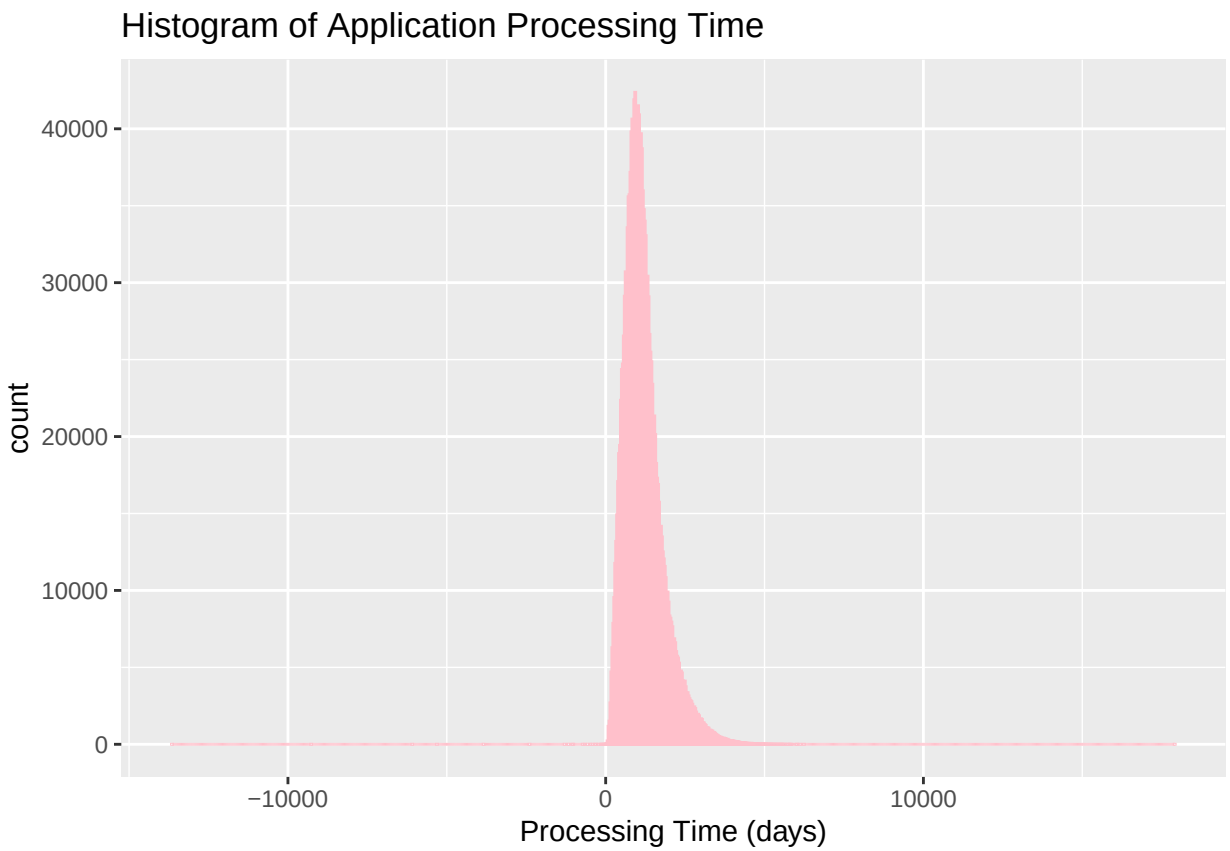
```r
applications=applications %>%
  left_join(centrality_entire, by = "examiner_id")

centrality_entire=data.frame(
  examiner_id = V(g)$name,
  degree_centrality = degree(g, mode = "out"),
  betweenness_centrality = betweenness(g, directed = TRUE),
  closeness_centrality = closeness(g, mode = "out")
)

centrality_entire$examiner_id=as.numeric(centrality_entire$examiner_id)
```

```
## Warning: NAs introduced by coercion
```

```r
applications=applications %>%
  left_join(centrality_entire, by = "examiner_id")
```

```r
# visualization
ggplot(applications, aes(x = app_proc_time)) +
  geom_histogram(binwidth = 30, fill = "blue", color = "pink") +
  labs(title = "Histogram of Application Processing Time", x = "Processing Time (days)")
```

```
## Warning: Removed 329761 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Regression Analysis

linear regression model

```
# Estimate the linear regression model with degree_centrality as the independent variable
degree_model=lm(
  app_proc_time ~ degree_centrality.x + gender + race + tenure_days,
  data = applications_clean
)

summary(degree_model)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree_centrality.x + gender + race +
##     tenure_days, data = applications_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2537.6  -442.5  -119.0   305.7  4933.2
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.543e+03  8.002e+00 192.856  < 2e-16 ***
## degree_centrality.x  1.509e-01  2.523e-02   5.980 2.24e-09 ***
## gendermale           2.716e+01  1.818e+00  14.937  < 2e-16 ***
## raceblack            4.762e+00  4.762e+00   1.000  0.31739
## raceHispanic         1.599e+01  5.749e+00   2.781  0.00542 **
## raceother            9.462e+00  3.615e+01   0.262  0.79349
## racewhite           -6.491e+01  1.925e+00 -33.726  < 2e-16 ***
## tenure_days         -4.627e-02  1.294e-03 -35.768  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.5 on 594077 degrees of freedom
##   (236232 observations deleted due to missingness)
## Multiple R-squared:  0.005387,   Adjusted R-squared:  0.005376
## F-statistic: 459.7 on 7 and 594077 DF,  p-value: < 2.2e-16
```

The degree_model has been constructed to explore the dynamics between an examiner's degree centrality in the USPTO advice network, demographic characteristics, and the time it takes to process patent applications. The model treats 'degree_centrality', 'gender', 'race', and 'tenure_days' as predictors of the 'app_proc_time', which is the outcome variable.

An analysis of the model's performance reveals an adjusted R-squared value of 0.0053769. This statistic suggests that the model accounts for approximately 0.33% of the variability in the application processing time. In practical terms, while the model captures a positive association between the independent variables and the application processing time, the low adjusted R-squared indicates a relatively weak explanatory power. Such a small percentage points to the possibility that other unexamined factors might play a significant role in influencing 'app_proc_time'.

The low explanatory power of this model raises questions about the potential complexities underlying the patent examination process that are not captured by the included variables. For instance, factors such as the complexity of the patent application, the field of invention, the workload of the examiners, and their interaction with the advice network may have significant impacts that are not encapsulated by the centrality

measure alone. Moreover, individual differences between examiners, such as their decision-making style and efficiency, could also contribute to the variation in processing times, beyond what demographic factors can explain.

relationship differ by examiner gender

```
# Degree centrality model with interaction
degree_gender_interaction=lm(
  app_proc_time ~ degree_centrality.x * gender + race + tenure_days,
  data = applications_clean
)
summary(degree_gender_interaction)
```

```
##
## Call:
## lm(formula = app_proc_time ~ degree_centrality.x * gender + race +
##     tenure_days, data = applications_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2538.4  -442.7  -118.7   305.7  4939.9
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.533e+03  8.059e+00 190.228  < 2e-16 ***
## degree_centrality.x             6.129e-01  5.092e-02  12.037  < 2e-16 ***
## gendermale                      3.675e+01  2.037e+00  18.043  < 2e-16 ***
## raceblack                       5.074e+00  4.762e+00   1.065  0.28670
## raceHispanic                    1.807e+01  5.752e+00   3.142  0.00168 **
## raceother                       9.810e+00  3.614e+01   0.271  0.78607
## racewhite                      -6.512e+01  1.925e+00 -33.837  < 2e-16 ***
## tenure_days                    -4.578e-02  1.295e-03 -35.363  < 2e-16 ***
## degree_centrality.x:gendermale -6.103e-01  5.842e-02 -10.447  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.4 on 594076 degrees of freedom
##   (236232 observations deleted due to missingness)
## Multiple R-squared:  0.00557,    Adjusted R-squared:  0.005557
## F-statistic:   416 on 8 and 594076 DF,  p-value: < 2.2e-16
```

The degree-gender interaction model presents a nuanced view of how gender may modify the relationship between an examiner's centrality in the USPTO network and their patent application processing times(Adjusted R-squared: 0.005557). The significant interaction term between degree centrality and gender points to a differential effect: while a higher degree centrality tends to be associated with longer processing times, this effect is not uniform across genders.

Specifically, for male examiners, the influence of degree centrality on prolonging the application processing time is mitigated, as indicated by the negative coefficient of the interaction term. This suggests that male examiners with higher centrality—potentially indicating a more significant advisory role or greater involvement in complex cases—might not experience as much of an increase in processing time as their female counterparts with similar centrality levels.

However, it's important to note that the model does not robustly explain the variance in processing times, given the low overall R-squared value. This implies that while the interaction between centrality and gender

is statistically discernible, there are still many aspects of the processing time that remain unaccounted for by this model. Other factors, perhaps related to the institutional environment, the nature of the applications themselves, or the support systems in place for examiners, could be influential and warrant further investigation.

The interaction effect observed prompts a deeper consideration of how structural and social factors within the USPTO might differentially affect the workflows of male and female examiners. It raises questions about the presence of potential gender-based differences in task allocation, access to resources, or the burden of informal roles that might not be immediately apparent from quantitative measures alone.

Findings: The low adjusted R-squared value from the degree_model suggests that the factors included in the analysis—degree centrality, gender, race, and tenure days—offer limited predictive power for application processing times. This weak fit signals that the patent examination process is influenced by a complex interplay of factors beyond an examiner's network position and basic demographic characteristics. The USPTO may need to consider a broader range of variables to more accurately forecast processing times and identify areas for efficiency improvements.

Although degree centrality has a quantifiable impact on processing times, its overall effect is small. This could imply that while network centrality captures some aspects of an examiner's role within the advice network, it does not necessarily translate into a large impact on their workflow efficiency. The USPTO could investigate the nature of these networks further to understand how to better leverage them for improved processing times.

The significant interaction between degree centrality and gender highlights a differential impact on processing times, suggesting that gender may play a role in how network centrality affects examiners. For male examiners, higher centrality is associated with less of an increase in processing time than for female examiners. This could point to underlying gender dynamics within the workplace that affect job performance. Understanding these dynamics could be crucial for the USPTO in creating a more balanced work environment.