Centrality and Efficiency

Ximena_Rodriguez

2023-04-03

This analysis presents the relationship between the efficiency of patent application processing and measures of centrality in an organizational network. Specifically, we will examine a dataset of patent applications from the United States Patent and Trademark Office and explore how the centrality of individual examiners in the organizational network relates to the processing time of patent applications.

Loading data

First we are going to load the data and see the attributes of each file.

```
load(file='applications.Rda')
load(file='edges.Rda')
print("Here are the attributes of applications:")
```

```
## [1] "Here are the attributes of applications:"
```

```
names(applications)
```

```
##  [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"                   "gender"               "race"
## [19] "earliest_date"        "latest_date"          "tenure_days"
```

```
print("Here are the attributes of edges:")
```

```
## [1] "Here are the attributes of edges:"
```

```
names(edges)
```

```
## [1] "application_number" "advice_date"        "ego_examiner_id"
## [4] "alter_examiner_id"
```

Calculating number of days between Filing Date and End Date

In this code chunk, we are preparing a dataset of patent applications for analysis by selecting only those that have been either abandoned or issued. We do this by checking if the “abandon_date” or “patent_issue_date” column contains any missing values.

We then combine the selected applications into a single data frame and remove unnecessary columns, having a single column of end date which gives us if the patent was either abandoned or issued.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.2.0
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the  ]8;;http://conflicted.r-lib.org/ conflicted package ]8;;  to force all conflicts t
o become errors
```

```
# Select patent applications that have either been abandoned or issued
abandoned_apps = applications[!is.na(applications$abandon_date),]
issued_apps = applications[!is.na(applications$patent_issue_date),]

# Rename columns and remove unnecessary columns
abandoned_apps = abandoned_apps %>% rename(end_date = abandon_date) %>% select(-c('patent_issue_
date'))
issued_apps = issued_apps %>% rename(end_date = patent_issue_date) %>% select(-c('abandon_dat
e'))
issued_apps$issued = 1
abandoned_apps$issued = 0

# Combine abandoned and issued patent applications into a single data frame
apps = rbind(abandoned_apps, issued_apps)
rm(abandoned_apps, issued_apps)
```

In the code below we are calculating the processing time, and we see that we have some data quality issues as the Min is -13636.

```
app_proc_time = apps$end_date - apps$filing_date
app_proc_time = as.numeric(app_proc_time)
summary(app_proc_time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  -13636     765    1079    1190    1481   17898
```

Here we will filter the processing times that are greater than zero.

```
apps$app_proc_time = app_proc_time
apps = apps[apps$app_proc_time >=0, ]
rm(applications, app_proc_time)

edges = edges %>% inner_join(apps[c('application_number')], by = 'application_number')
head(edges, 10)
```

```
## # A tibble: 10 × 4
##    application_number advice_date ego_examiner_id alter_examiner_id
##    <chr>              <date>                <dbl>             <dbl>
##  1 09402488           2008-11-17            84356             66266
##  2 09402488           2008-11-17            84356             63519
##  3 09402488           2008-11-17            84356             98531
##  4 09445135           2008-08-21            92953             71313
##  5 09445135           2008-08-21            92953             93865
##  6 09445135           2008-08-21            92953             91818
##  7 09479304           2008-12-15            61767             69277
##  8 09479304           2008-12-15            61767             92446
##  9 09479304           2008-12-15            61767             66805
## 10 09479304           2008-12-15            61767             70919
```

Modeling

Linear regression is a statistical technique used to model the relationship between a dependent variable (in this case, days of patent processing) and one or more independent variables (such as measures of centrality or attributes of examiner).

Linear regression can help us identify the strength and direction of the relationship between these variables, as well as provide estimates of the magnitude of the effect of the independent variables on the dependent variable.

```
library(tidygraph)
```

```
##
## Attaching package: 'tidygraph'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggraph)
edges = edges %>% rename(to = alter_examiner_id,
                     from = ego_examiner_id)

graph = as_tbl_graph(x = edges[c('to','from')], directed = TRUE , mode = 'out')
```

```
## Warning in graph_from_data_frame(x, directed = directed): In `d' `NA' elements
## were replaced with string "NA"
```

```
nodes = graph %>%
  activate(nodes) %>%
  mutate(degree = centrality_degree(),
         closeness = centrality_closeness(),
         betweenness = centrality_betweenness()) %>%
  rename(examiner_id = name) %>%  data.frame()

apps$examiner_id = as.character(apps$examiner_id)



apps = apps %>% left_join(nodes, by = 'examiner_id')
```

Linear Regression: Model 1

First,we will drop any rows that have missing values in the variables “app_proc_time”, “degree”, “closeness”, “betweenness”, “gender”, “race”, or “tenure_days”. This is done using the “drop_na” function from the “tidyverse” package.

Then, we are converting the “gender” and “race” variables to factor variables, which will allow us to include them in the regression model.

```
attach(apps)
names(apps)
```

```
##  [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"        "end_date"             "disposal_type"       
## [13] "appl_status_code"     "appl_status_date"     "tc"                  
## [16] "gender"               "race"                 "earliest_date"       
## [19] "latest_date"          "tenure_days"          "issued"              
## [22] "app_proc_time"        "degree"               "closeness"           
## [25] "betweenness"
```

```
apps = apps %>%  drop_na(app_proc_time, degree,closeness, betweenness, gender, race,tenure_days)

apps$gender = as.factor(apps$gender)
apps$race = as.factor(apps$race)

lm_1 = lm(app_proc_time ~ degree + closeness + betweenness + gender + race + tenure_days + issue
d)

summary(lm_1)
```

```
## 
## Call:
## lm(formula = app_proc_time ~ degree + closeness + betweenness + 
##     gender + race + tenure_days + issued)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1296.2  -440.4  -118.0   305.1  4999.6 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   1.500e+03  8.126e+00 184.615  < 2e-16 ***
## degree       -2.074e-01  2.619e-02  -7.919  2.4e-15 ***
## closeness    -1.179e+02  2.423e+00 -48.650  < 2e-16 ***
## betweenness   9.781e-04  1.223e-04   7.995  1.3e-15 ***
## gendermale    2.504e+01  1.821e+00  13.749  < 2e-16 ***
## raceblack     2.103e+01  4.763e+00   4.417  1.0e-05 ***
## raceHispanic  1.826e+01  5.736e+00   3.183  0.00146 ** 
## raceother     5.083e+00  3.607e+01   0.141  0.88792    
## racewhite    -5.857e+01  1.926e+00 -30.412  < 2e-16 ***
## tenure_days  -3.592e-02  1.341e-03 -26.782  < 2e-16 ***
## issued        2.369e+01  1.753e+00  13.514  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 644.9 on 592693 degrees of freedom
##   (1095980 observations deleted due to missingness)
## Multiple R-squared:  0.01004,	Adjusted R-squared:  0.01002 
## F-statistic: 601.2 on 10 and 592693 DF,  p-value: < 2.2e-16
```

Results Model 1

- All independent variables but “Race: Other” are significant to predict the number of days to process a patent.
- Having high degree of centrality and closeness on average gives more efficiency in processing the patents. To be specific, on average the processing time of an application is reduced by 117 days for each additional degree of closeness centrality that an examiner has.
- White examiners take 59 days less on average processing applications holding all other variables equal.
- Male examiners take 25 days more on average processing applications holding all other variables equal.
- Issued patents take 23 days more on average than the abandoned patents holding all other variables equal.
- The multiple R-squared is 0.01004, which means that the independent variables explain only about 1% of the variation in the dependent variable.

Linear Regression: Model 2

```
lm_2 = lm(app_proc_time ~ degree + closeness + betweenness + gender + race + tenure_days+issued
+ gender*degree + gender*betweenness + gender*closeness)

summary(lm_2)
```

```
## 
## Call:
## lm(formula = app_proc_time ~ degree + closeness + betweenness + 
##     gender + race + tenure_days + issued + gender * degree + 
##     gender * betweenness + gender * closeness)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1310.2  -440.5  -118.1   305.1  4991.8 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             1.491e+03  8.372e+00 178.040  < 2e-16 ***
## degree                  2.575e-01  5.406e-02   4.763 1.90e-06 ***
## closeness              -1.028e+02  4.281e+00 -24.003  < 2e-16 ***
## betweenness            -1.301e-03  2.224e-04  -5.850 4.92e-09 ***
## gendermale              3.488e+01  2.736e+00  12.747  < 2e-16 ***
## raceblack               1.921e+01  4.766e+00   4.030 5.58e-05 ***
## raceHispanic            1.717e+01  5.756e+00   2.982 0.002864 ** 
## raceother               5.645e+00  3.606e+01   0.157 0.875607    
## racewhite              -5.894e+01  1.926e+00 -30.599  < 2e-16 ***
## tenure_days            -3.563e-02  1.344e-03 -26.503  < 2e-16 ***
## issued                  2.413e+01  1.754e+00  13.760  < 2e-16 ***
## degree:gendermale      -6.094e-01  6.173e-02  -9.872  < 2e-16 ***
## betweenness:gendermale  3.208e-03  2.641e-04  12.147  < 2e-16 ***
## closeness:gendermale   -1.991e+01  5.146e+00  -3.869 0.000109 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 644.7 on 592690 degrees of freedom
##   (1095980 observations deleted due to missingness)
## Multiple R-squared:  0.01042,    Adjusted R-squared:  0.0104 
## F-statistic: 480.3 on 13 and 592690 DF,  p-value: < 2.2e-16
```

Results Model 2

- We see similar results in regards of the independent variables that were included in Model 1
- Differences rely on the interaction terms which are also significant to predict the number of processing days for the patents.
- For every extra closeness degree that the male has, he will reduce the processing days by 19 days than for females.