Excercise 4 - Centrality & Efficiency

Hugo Garcia (260791363)

March 4th, 2023

Load data

Load the following data: + applications from `app_data_sample.parquet` + edges from `edges_sample.csv`

```
# change to your own path!
data_path <- "C:/Users/hugog/Desktop/Exercise 4/"
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.
```

```
applications
```

```
## # A tibble: 2,018,477 x 16
##    applicat~1 filing_d~2 exami~3 exami~4 exami~5 exami~6 exami~7 uspc_~8 uspc_~9
##    <chr>      <date>     <chr>   <chr>   <chr>     <dbl>   <dbl> <chr>   <chr>
##  1 08284457   2000-01-26 HOWARD  JACQUE~ V         96082    1764 508     273000
##  2 08413193   2000-10-11 YILDIR~ BEKIR   L         87678    1764 208     179000
##  3 08531853   2000-05-17 HAMILT~ CYNTHIA <NA>      63213    1752 430     271100
##  4 08637752   2001-07-20 MOSHER  MARY    <NA>      73788    1648 530     388300
##  5 08682726   2000-04-10 BARR    MICHAEL E         77294    1762 427     430100
##  6 08687412   2000-04-28 GRAY    LINDA   LAMEY     68606    1734 156     204000
##  7 08716371   2004-01-26 MCMILL~ KARA    RENITA    89557    1627 424     401000
##  8 08765941   2000-06-23 FORD    VANESSA L         97543    1645 424     001210
##  9 08776818   2000-02-04 STRZEL~ TERESA  E         98714    1637 435     006000
## 10 08809677   2002-02-20 KIM     SUN     U         65530    1723 210     645000
## # ... with 2,018,467 more rows, 7 more variables: patent_number <chr>,
## #   patent_issue_date <date>, abandon_date <date>, disposal_type <chr>,
## #   appl_status_code <dbl>, appl_status_date <chr>, tc <dbl>, and abbreviated
## #   variable names 1: application_number, 2: filing_date,
## #   3: examiner_name_last, 4: examiner_name_first, 5: examiner_name_middle,
## #   6: examiner_id, 7: examiner_art_unit, 8: uspc_class, 9: uspc_subclass
```

```
edges
```

```
## # A tibble: 32,906 x 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
## # ... with 32,896 more rows
```

Get gender for examiners

```
library(gender)
```

```
## Warning: package 'gender' was built under R version 4.2.3
```

```
examiner_names <- applications %>%
  distinct(examiner_name_first)
examiner_names
```

```
## # A tibble: 2,595 x 1
##    examiner_name_first
##    <chr>
##  1 JACQUELINE
##  2 BEKIR
##  3 CYNTHIA
##  4 MARY
##  5 MICHAEL
##  6 LINDA
##  7 KARA
##  8 VANESSA
##  9 TERESA
## 10 SUN
## # ... with 2,585 more rows
```

Get a table of names and gender

```
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
  )
examiner_names_gender
```

```
## # A tibble: 1,822 x 3
##    examiner_name_first gender proportion_female
##    <chr>               <chr>              <dbl>
##  1 AARON               male              0.0082
##  2 ABDEL               male              0
##  3 ABDOU               male              0
##  4 ABDUL               male              0
##  5 ABDULHAKIM          male              0
##  6 ABDULLAH            male              0
##  7 ABDULLAHI           male              0
##  8 ABIGAIL             female            0.998
##  9 ABIMBOLA            female            0.944
## 10 ABRAHAM             male              0.0031
## # ... with 1,812 more rows
```

```
# remove extra colums from the gender table
examiner_names_gender <- examiner_names_gender %>%
  select(examiner_name_first, gender)
# joining gender back to the dataset
applications <- applications %>%
  left_join(examiner_names_gender, by = "examiner_name_first")
# cleaning up
rm(examiner_names)
rm(examiner_names_gender)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4614782 246.5    7923377 423.2  5033991 268.9
## Vcells 49730248 379.5   95687393 730.1 80045959 610.8
```

```
library(wru)
```

```
## Warning: package 'wru' was built under R version 4.2.3
```

```
examiner_surnames <- applications %>%
  select(surname = examiner_name_last) %>%
  distinct()
examiner_surnames
```

```
## # A tibble: 3,806 x 1
##    surname
##    <chr>
##  1 HOWARD
##  2 YILDIRIM
##  3 HAMILTON
##  4 MOSHER
##  5 BARR
##  6 GRAY
##  7 MCMILLIAN
##  8 FORD
##  9 STRZELECKA
## 10 KIM
## # ... with 3,796 more rows
```

We'll follow the instructions for the package outlined here https://github.com/kosukeimai/wru.

NOTE: I was getting errors running the original code block for examiner_race. I tried updating packages, and debugging for a long time but I believe it is my computer's software/environment setup preventing me. I asked for the csv output from a peer and am importing it instead. Original code bloack: examiner_race <- predict_race(voter.file = examiner_surnames, surname.only = T) %>% as_tibble() examiner_race

```
examiner_race <- read_csv(paste0(data_path,"examiner_race.csv"))
```

```
## Rows: 3806 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): surname
## dbl (5): pred.whi, pred.bla, pred.his, pred.asi, pred.oth
##
## 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.
```

```
examiner_race
```

```
## # A tibble: 3,806 x 6
##    surname    pred.whi pred.bla pred.his pred.asi pred.oth
##    <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
##  1 HOWARD       0.597   0.295    0.0275   0.00690   0.0741
##  2 YILDIRIM     0.807   0.0273   0.0694   0.0165    0.0798
##  3 HAMILTON     0.656   0.239    0.0286   0.00750   0.0692
##  4 MOSHER       0.915   0.00425  0.0291   0.00917   0.0427
##  5 BARR         0.784   0.120    0.0268   0.00830   0.0615
##  6 GRAY         0.640   0.252    0.0281   0.00748   0.0724
##  7 MCMILLIAN    0.322   0.554    0.0212   0.00340   0.0995
##  8 FORD         0.576   0.320    0.0275   0.00621   0.0697
##  9 STRZELECKA   0.472   0.171    0.220    0.0825    0.0543
## 10 KIM          0.0169  0.00282  0.00546  0.943     0.0319
## # ... with 3,796 more rows
```

```
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_
  ))
examiner_race
```

```
## # A tibble: 3,806 x 8
##    surname    pred.whi pred.bla pred.his pred.asi pred.oth max_race_p race
##    <chr>         <dbl>    <dbl>    <dbl>    <dbl>    <dbl>      <dbl> <chr>
##  1 HOWARD       0.597   0.295    0.0275   0.00690   0.0741      0.597 white
##  2 YILDIRIM     0.807   0.0273   0.0694   0.0165    0.0798      0.807 white
##  3 HAMILTON     0.656   0.239    0.0286   0.00750   0.0692      0.656 white
##  4 MOSHER       0.915   0.00425  0.0291   0.00917   0.0427      0.915 white
##  5 BARR         0.784   0.120    0.0268   0.00830   0.0615      0.784 white
##  6 GRAY         0.640   0.252    0.0281   0.00748   0.0724      0.640 white
##  7 MCMILLIAN    0.322   0.554    0.0212   0.00340   0.0995      0.554 black
##  8 FORD         0.576   0.320    0.0275   0.00621   0.0697      0.576 white
##  9 STRZELECKA   0.472   0.171    0.220    0.0825    0.0543      0.472 white
## 10 KIM          0.0169  0.00282  0.00546  0.943     0.0319      0.943 Asian
## # ... with 3,796 more rows
```

Join the data back to the applications table.

```
# removing extra columns
examiner_race <- examiner_race %>%
  select(surname,race)
applications <- applications %>%
  left_join(examiner_race, by = c("examiner_name_last" = "surname"))
rm(examiner_race)
rm(examiner_surnames)
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  4651609 248.5    7923377 423.2  6040103 322.6
## Vcells 51809613 395.3   95687393 730.1 94688015 722.5
```

Add Tenure

```
library(lubridate) # to work with dates
examiner_dates <- applications %>%
  select(examiner_id, filing_date, appl_status_date)
examiner_dates
```

```
## # A tibble: 2,018,477 x 3
##    examiner_id filing_date appl_status_date
##          <dbl> <date>      <chr>
##  1       96082 2000-01-26  30jan2003 00:00:00
##  2       87678 2000-10-11  27sep2010 00:00:00
##  3       63213 2000-05-17  30mar2009 00:00:00
##  4       73788 2001-07-20  07sep2009 00:00:00
##  5       77294 2000-04-10  19apr2001 00:00:00
##  6       68606 2000-04-28  16jul2001 00:00:00
##  7       89557 2004-01-26  15may2017 00:00:00
##  8       97543 2000-06-23  03apr2002 00:00:00
##  9       98714 2000-02-04  27nov2002 00:00:00
## 10       65530 2002-02-20  23mar2009 00:00:00
## # ... with 2,018,467 more rows
```

```
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)
examiner_dates
```

```
## # A tibble: 5,625 x 4
##    examiner_id earliest_date latest_date tenure_days
##          <dbl> <date>        <date>            <dbl>
##  1       59012 2004-07-28    2015-07-24         4013
##  2       59025 2009-10-26    2017-05-18         2761
##  3       59030 2005-12-12    2017-05-22         4179
##  4       59040 2007-09-11    2017-05-23         3542
##  5       59052 2001-08-21    2007-02-28         2017
##  6       59054 2000-11-10    2016-12-23         5887
##  7       59055 2004-11-02    2007-12-26         1149
##  8       59056 2000-03-24    2017-05-22         6268
##  9       59074 2000-01-31    2017-03-17         6255
## 10       59081 2011-04-21    2017-05-19         2220
## # ... with 5,615 more rows
```

```
applications <- applications %>%
  left_join(examiner_dates, by = "examiner_id")
rm(examiner_dates)
gc()
```

```
##            used  (Mb) gc trigger   (Mb)  max used   (Mb)
## Ncells  4665723 249.2   14310454  764.3  14310454  764.3
## Vcells 64188519 489.8  137965845 1052.6 137831562 1051.6
```

Pre-Processing

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

# Rename and remove unnecessary columns
```

```r
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_date'))
issued_apps$issued = 1
abandoned_apps$issued = 0

# Combine abandoned and issued dates
applications = rbind(abandoned_apps, issued_apps)
rm(abandoned_apps, issued_apps)
```

Calculate Application Processing Times

```r
app_proc_time = applications$end_date - applications$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
```

```r
# There were some errors in the dates which led to negative app_proc_time, remove these
applications$app_proc_time = app_proc_time
applications = applications[applications$app_proc_time >=0, ]
```

Calculate Centrality Scores for Each Examiner

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

# This is a directed  network, so Directed=TRUE
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"
```

```r
nodes = graph %>%
  activate(nodes) %>%
  mutate(degree = centrality_degree(),
         closeness = centrality_closeness(),
         betweenness = centrality_betweenness()) %>%
  rename(examiner_id = name) %>%  data.frame()
applications$examiner_id = as.character(applications$examiner_id)
applications = applications %>% left_join(nodes, by = 'examiner_id')
```

Drop NA Vaues and Set some Variables as factor

```
attach(applications)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     app_proc_time
```

```
applications = applications %>% drop_na(gender)
applications = applications %>% drop_na(race)
applications = applications %>% drop_na(degree)
applications = applications %>% drop_na(closeness)
applications = applications %>% drop_na(betweenness)
applications = applications %>% drop_na(tenure_days)

applications$gender = as.factor(applications$gender)
applications$race = as.factor(applications$race)
```

Use a Linear Regression Model to Estimate the Relationship between Centrality and Application Processing Times

Controlling for other characteristics of the examiner which might influence that relationship

```
applications_lm <- applications
lm = lm(applications_lm$app_proc_time ~
            applications_lm$degree +
            applications_lm$closeness +
            applications_lm$betweenness +
            applications_lm$gender +
            applications_lm$race +
            applications_lm$issued +
          applications_lm$tenure_days)
summary(lm)
```

```
##
## Call:
## lm(formula = applications_lm$app_proc_time ~ applications_lm$degree +
##     applications_lm$closeness + applications_lm$betweenness +
##     applications_lm$gender + applications_lm$race + applications_lm$issued +
##     applications_lm$tenure_days)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1296.3  -440.3  -118.0   305.0  4999.7
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.501e+03  8.122e+00 184.837  < 2e-16 ***
## applications_lm$degree       -2.024e-01  2.616e-02  -7.735 1.04e-14 ***
## applications_lm$closeness    -1.181e+02  2.422e+00 -48.747  < 2e-16 ***
## applications_lm$betweenness   9.741e-04  1.222e-04   7.972 1.57e-15 ***
## applications_lm$gendermale    2.483e+01  1.819e+00  13.649  < 2e-16 ***
## applications_lm$raceblack     2.065e+01  4.762e+00   4.336 1.45e-05 ***
## applications_lm$raceHispanic  1.799e+01  5.736e+00   3.136  0.00171 **
## applications_lm$raceother     4.832e+00  3.607e+01   0.134  0.89343
## applications_lm$racewhite    -5.895e+01  1.924e+00 -30.633  < 2e-16 ***
## applications_lm$issued        2.405e+01  1.751e+00  13.741  < 2e-16 ***
## applications_lm$tenure_days  -3.606e-02  1.341e-03 -26.893  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 644.9 on 594067 degrees of freedom
## Multiple R-squared:  0.01011,    Adjusted R-squared:  0.0101
## F-statistic:   607 on 10 and 594067 DF,  p-value: < 2.2e-16
```

Results 1

All independent variables are significant to predict the number of days to process a patent, except "Race - Other", and to a lesser degree of significance "Race- Hispanic", which is likely a result of there being relatively fewer Hispanic Examiners and thus fewer applications in total on their behalf.

The multiple R-squared value of 0.01011 indicates that the independent variables explain just 1% of the variation in the applications processing times. We also see that Closeness Centrality significantly influences application processing time whereby each additional degree of closeness decreases the application processing time by 118 days. Similarly for Degree of Centrality, but to a lesser extent, each additional Degree of Centrality decreases application processing time by about 20 days. Lastly, the Tenure of an examiner significantly influences their application processing times whereby each

In addition, based on the regression model's output we can also make comparisons between examiner characteristics and how they relate to their application processing times:

- Male Examiners take about 35 days longer, on average, to process applications than female examiners
- White Examiners take about 58 days less, on average, to process applications.

Linear Regression Including Interaction Variables for Gender and Centrality

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

```
##
## Call:
## lm(formula = applications_lm$app_proc_time ~ applications_lm$degree +
##     applications_lm$closeness + applications_lm$betweenness +
##     applications_lm$gender + applications_lm$race + applications_lm$tenure_days + 
##     applications_lm$issued + (applications_lm$gender * applications_lm$degree) + 
##     (applications_lm$gender * applications_lm$betweenness) + 
##     (applications_lm$gender * applications_lm$closeness))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1310.1  -440.5  -118.2   305.0  4991.4 
## 
## Coefficients:
##                                                           Estimate Std. Error
## (Intercept)                                             1.491e+03  8.369e+00
## applications_lm$degree                                  2.612e-01  5.401e-02
## applications_lm$closeness                              -1.023e+02  4.280e+00
## applications_lm$betweenness                            -1.302e-03  2.222e-04
## applications_lm$gendermale                              3.490e+01  2.735e+00
## applications_lm$raceblack                               1.880e+01  4.765e+00
## applications_lm$raceHispanic                            1.682e+01  5.756e+00
## applications_lm$raceother                               5.328e+00  3.606e+01
## applications_lm$racewhite                              -5.931e+01  1.925e+00
## applications_lm$tenure_days                            -3.576e-02  1.344e-03
## applications_lm$issued                                  2.450e+01  1.751e+00
## applications_lm$degree:applications_lm$gendermale      -6.074e-01  6.166e-02
## applications_lm$betweenness:applications_lm$gendermale  3.204e-03  2.638e-04
## applications_lm$closeness:applications_lm$gendermale   -2.077e+01  5.144e+00
##                                                        t value Pr(>|t|)    
## (Intercept)                                            178.190  < 2e-16 ***
## applications_lm$degree                                   4.836 1.32e-06 ***
## applications_lm$closeness                              -23.912  < 2e-16 ***
## applications_lm$betweenness                             -5.862 4.57e-09 ***
## applications_lm$gendermale                              12.761  < 2e-16 ***
## applications_lm$raceblack                                3.944 8.01e-05 ***
## applications_lm$raceHispanic                             2.922  0.00348 ** 
## applications_lm$raceother                                0.148  0.88253    
## applications_lm$racewhite                              -30.818  < 2e-16 ***
## applications_lm$tenure_days                            -26.604  < 2e-16 ***
## applications_lm$issued                                  13.991  < 2e-16 ***
## applications_lm$degree:applications_lm$gendermale       -9.852  < 2e-16 ***
## applications_lm$betweenness:applications_lm$gendermale  12.145  < 2e-16 ***
## applications_lm$closeness:applications_lm$gendermale    -4.038 5.38e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 644.8 on 594064 degrees of freedom
## Multiple R-squared:  0.0105, Adjusted R-squared:  0.01048 
## F-statistic: 484.9 on 13 and 594064 DF,  p-value: < 2.2e-16
```

Results 2

We see similar results than in the regression above for the significance of the independent variables. Regarding the interactions between gender and centrality measures, all are statistically significant.