Exercise 4

Niki Mahmoodzadeh

2024-04-08

Introduction

This task aims to explore how a patent examiner's network centrality influences the duration of patent application processes.

Data loading and preprocessing

Initially, I import the dataset and prepare it for subsequent analysis.

```
applications = read_csv("/Users/nikimahmoodzadeh/Downloads/672_project_data-2/app_data_sample.csv", show_col_types = FALSE)
edges = read_csv("/Users/nikimahmoodzadeh/Downloads/672_project_data-2/edges_sample.csv",show_col_types = FALSE)
```

Data Preprocessing

Estimating examiner demographics

```
## 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.
```

Creating processing time variable

Centrality measures

Next, I generate a distinct list of examiner IDs as a preparatory step before diving into the main analysis.

```
## 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"
```

Following this, I begin to calculate various centrality metrics for further examination.

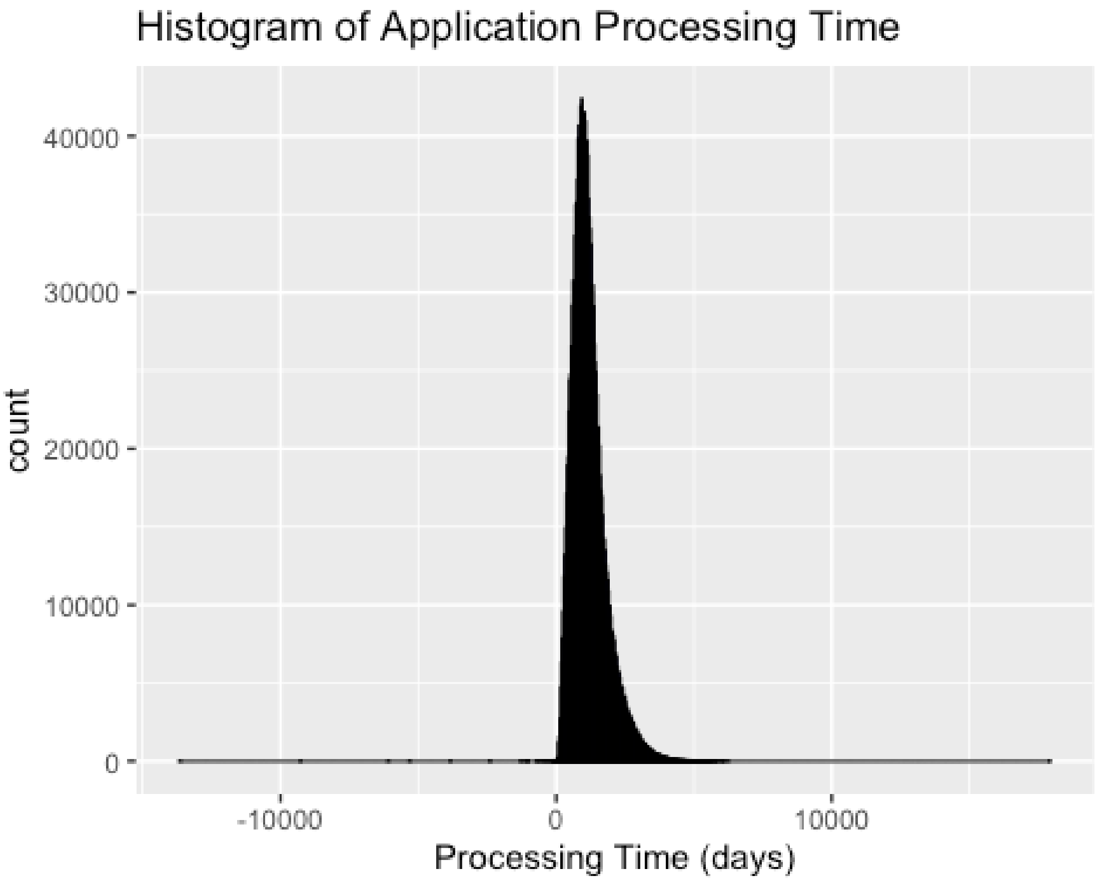
```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

Exploratory Data Analysis

```
##  [1] "application_number"         "filing_date"
##  [3] "examiner_name_last"         "examiner_name_first"
##  [5] "examiner_name_middle"       "examiner_id"
##  [7] "examiner_art_unit"          "uspc_class"
##  [9] "uspc_subclass"              "patent_number"
## [11] "patent_issue_date"          "abandon_date"
## [13] "disposal_type"              "appl_status_code"
## [15] "appl_status_date"           "tc"
## [17] "gender.x"                   "race.x"
## [19] "earliest_date.x"            "latest_date.x"
## [21] "tenure_days.x"              "gender.y"
## [23] "proportion_female"          "pred.whi"
## [25] "pred.bla"                   "pred.his"
## [27] "pred.asi"                   "pred.oth"
## [29] "max_race_p"                 "race.y"
## [31] "earliest_date.y"            "latest_date.y"
## [33] "tenure_days.y"              "final_decision_date"
## [35] "app_proc_time"              "degree_centrality.x"
## [37] "betweenness_centrality.x"   "closeness_centrality.x"
## [39] "degree_centrality.y"        "betweenness_centrality.y"
## [41] "closeness_centrality.y"

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



Regression Analysis

First, I will remove the missing values in degree, betweenness, and closeness centrality.

Degree centrality linear regression model

I conduct an analysis to construct a linear regression model, using degree centrality as the predictor variable.

```
##
## Call:
## lm(formula = app_proc_time ~ degree_centrality.x + gender.x +
##       race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
##  -2518.1  -444.2  -118.6   306.9  4921.0
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.268e+03  2.065e+00 613.789  < 2e-16 ***
## degree_centrality.x 2.111e-01  2.502e-02   8.437  < 2e-16 ***
## gender.xmale        3.371e+01  1.800e+00  18.727  < 2e-16 ***
## race.xblack                1.162e+00   4.772e+00     0.244   0.807546
## race.xHispanic             2.139e+01   5.760e+00     3.714   0.000204  ***
## race.xother                5.505e+01   3.621e+01     1.520   0.128413
## race.xwhite               -6.752e+01   1.924e+00   -35.089     < 2e-16  ***
## tenure_days.x              1.084e-04   9.098e-06    11.919     < 2e-16  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 647.9 on 598624 degrees of freedom
##   (231685 observations deleted due to missingness)
## Multiple R-squared:  0.00335,           Adjusted R-squared:  0.003339
## F-statistic: 287.5 on 7 and 598624 DF,  p-value: < 2.2e-16
```

Explanation on degree centrality linear regression model

The model constructed includes variables such as degree centrality, gender, race, and tenure days, predicting application processing time. The model's adjusted R-squared value is 0.003339, indicating a mere 0.33% variance in processing time can be accounted for by these variables, suggesting a poor model fit.

Betweenness centrality linear regression model

I proceed to estimate a linear regression model, this time with betweenness centrality as the predictor.

```
##
## Call:
## lm(formula = app_proc_time ~ betweenness_centrality.x + gender.x +
##      race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##      Min       1Q  Median         3Q      Max
## -2517.1  -444.2  -118.4        306.6  4920.1
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.268e+03  2.029e+00 625.112  < 2e-16 ***
## betweenness_centrality.x  1.473e-03  1.184e-04  12.445  < 2e-16 ***
## gender.xmale                   3.326e+01  1.801e+00  18.472  < 2e-16 ***
## race.xblack                     1.453e+00   4.770e+00     0.305   0.760671
## race.xHispanic                  2.213e+01   5.760e+00     3.842   0.000122  ***
## race.xother                     5.788e+01   3.620e+01     1.599   0.109880
## race.xwhite                    -6.724e+01   1.924e+00   -34.948     < 2e-16  ***
## tenure_days.x                   1.078e-04   9.097e-06    11.855     < 2e-16  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 647.9 on 598624 degrees of freedom
##   (231685 observations deleted due to missingness)
## Multiple R-squared:  0.00349,           Adjusted R-squared:  0.003478
## F-statistic: 299.5 on 7 and 598624 DF,  p-value: < 2.2e-16
```

Explanation on betweenness centrality linear regression model

This model, incorporating betweenness centrality, gender, race, and tenure days to predict processing time, achieves an adjusted R-squared of 0.003478. This further implies that betweenness centrality is an ineffective predictor of processing time.

Closeness centrality linear regression model

Next, I estimate a linear regression model with closeness centrality as the predictor.

```
# Closeness centrality linear regression model
closeness_model=lm(
  app_proc_time ~ closeness_centrality.x + gender.x + race.x + tenure_days.x,
  data = applications_clean
)
summary(closeness_model)
```

```
##
## Call:
## lm(formula = app_proc_time ~ closeness_centrality.x + gender.x +
##     race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2552.9  -442.0  -118.4   306.5  5008.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.304e+03  2.086e+00 624.891  < 2e-16 ***
## closeness_centrality.x -1.290e+02  2.261e+00 -57.082  < 2e-16 ***
## gender.xmale            3.109e+01  1.796e+00  17.311  < 2e-16 ***
## race.xblack             2.027e+01  4.769e+00   4.251 2.13e-05 ***
## race.xHispanic          2.111e+01  5.743e+00   3.676 0.000237 ***
## race.xother             2.501e+01  3.611e+01   0.693 0.488593
## race.xwhite            -6.175e+01  1.921e+00 -32.149  < 2e-16 ***
## tenure_days.x           9.540e-05  9.076e-06  10.512  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.2 on 598624 degrees of freedom
##   (231685 observations deleted due to missingness)
## Multiple R-squared:  0.008628,  Adjusted R-squared:  0.008616
## F-statistic: 744.3 on 7 and 598624 DF,  p-value: < 2.2e-16
```

Explanation on closeness centrality linear regression model

The closeness centrality model, including the same set of variables, yields an adjusted R-squared of 0.008616, showing a slight improvement over the previous models but still indicating weak predictive capability.

Combined model of linear regression

Finally, I estimate a combined linear regression model that includes all centrality measures.

```
##
## Call:
## lm(formula = app_proc_time ~ degree_centrality.x + betweenness_centrality.x +
##          closeness_centrality.x + gender.x + race.x + tenure_days.x,
##          data = applications_clean)
##
## Residuals:
##       Min       1Q  Median       3Q       Max
##  -2554.1  -441.8  -118.6   306.2  5008.8
##
## Coefficients:
##                                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              1.307e+03   2.194e+00 595.466   < 2e-16 ***
## degree_centrality.x        -1.845e-01   2.601e-02   -7.094 1.31e-12 ***
## betweenness_centrality.x   4.676e-04   1.204e-04    3.882 0.000103 ***
## closeness_centrality.x   -1.319e+02   2.363e+00 -55.843   < 2e-16 ***
## gender.xmale                             3.090e+01   1.796e+00   17.204   < 2e-16 ***
## race.xblack                              1.989e+01   4.771e+00    4.168 3.07e-05 ***
##   race.xHispanic                          2.085e+01      5.745e+00        3.629      0.000285     ***
##   race.xother                             2.486e+01      3.611e+01        0.688      0.491212
##   race.xwhite                            -6.191e+01      1.922e+00    -32.207         < 2e-16     ***
##   tenure_days.x                           9.461e-05      9.076e-06       10.424         < 2e-16     ***
##   ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.2 on 598622 degrees of freedom
##       (231685 observations deleted due to missingness)
## Multiple R-squared:   0.008727,                     Adjusted R-squared:   0.008712
## F-statistic: 585.6 on 9 and 598622 DF,   p-value: < 2.2e-16
```

Explanation on combined model

The combined model exhibits an adjusted R-squared of 0.008712, a minor improvement over the closeness model's 0.008616, underscoring the marginal enhancement achieved by combining these centrality measures.

Analysis to see if this relationship differ by examiner gender

Degree-Gender interaction

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

```
##
## Call:
## lm(formula = app_proc_time ~ degree_centrality.x * gender.x +
##        race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##        Min         1Q  Median          3Q       Max
##   -2519.3  -444.4  -118.3           307.0  4928.8
##
## Coefficients:                                                  Pr(>|t|)
##                                     Estimate Std. Error t value
## (Intercept)                        1.259e+03  2.176e+00 578.838        < 2e-16 ***
## degree_centrality.x                7.435e-01  5.085e-02  14.620        < 2e-16 ***
## gender.xmale                       4.465e+01  2.017e+00  22.139       <2e-16   ***
## race.xblack                        1.581e+00  4.771e+00       0.331    0.740
## race.xHispanic                       2.373e+01   5.762e+00     4.119   3.81e-05 ***
## race.xother                          5.490e+01   3.620e+01     1.516      0.129
## race.xwhite                         -6.766e+01   1.924e+00   -35.168    < 2e-16 ***
## tenure_days.x                        1.080e-04   9.097e-06    11.877    < 2e-16 ***
## degree_centrality.x:gender.xmale    -7.021e-01   5.838e-02   -12.025    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 647.8 on 598623 degrees of freedom
##    (231685 observations deleted due to missingness)
## Multiple R-squared:  0.003591,            Adjusted R-squared:  0.003578
## F-statistic: 269.7 on 8 and 598623 DF,  p-value: < 2.2e-16
```

Explanation on Degree-Gender interaction

In the model analyzing the interaction between degree centrality and gender, a significant interaction indicates varying effects of degree centrality on processing time by gender, with a mitigated effect observed for male examiners.

Betweenness-Gender interaction

```
# Betweenness centrality model with interaction
betweenness_gender_interaction=lm(
  app_proc_time ~  betweenness_centrality.x *  gender.x +  race.x +  tenure_days.x,
  data = applications_clean
)
summary(betweenness_gender_interaction)

##
## Call:
## lm(formula = app_proc_time ~ betweenness_centrality.x * gender.x +
##        race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##      Min      1Q  Median      3Q     Max
## -2515.5  -444.0  -118.9   306.6  4916.9
##
## Coefficients:
```

```
##                                              Estimate   Std. Error   t value  Pr(>|t|)
## (Intercept)                                 1.272e+03    2.050e+00   620.440   < 2e-16
## betweenness_centrality.x                   -5.457e-04    2.184e-04    -2.498  0.012479
## gender.xmale                                2.855e+01    1.851e+00    15.426   < 2e-16
## race.xblack                                 4.172e-01    4.770e+00     0.087  0.930313
## race.xHispanic                              2.116e+01    5.760e+00     3.674  0.000239
## race.xother                                 5.940e+01    3.620e+01     1.641  0.100814
## race.xwhite                                -6.723e+01    1.924e+00   -34.947   < 2e-16
## tenure_days.x                               1.079e-04    9.096e-06    11.863   < 2e-16
## betweenness_centrality.x:gender.xmale     2.856e-03      2.597e-04    10.998   < 2e-16
##
## (Intercept)                             ***
## betweenness_centrality.x                *
## gender.xmale                            ***
## race.xblack
## race.xHispanic                          ***
## race.xother
## race.xwhite                             ***
## tenure_days.x                           ***
## betweenness_centrality.x:gender.xmale   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 647.8 on 598623 degrees of freedom
##   (231685 observations deleted due to missingness)
## Multiple R-squared:  0.003691,    Adjusted R-squared:  0.003678
## F-statistic: 277.2 on 8 and 598623 DF,  p-value: < 2.2e-16
```

Explanation on Betweenness-Gender interaction

The model examining betweenness centrality and gender interaction demonstrates that higher betweenness centrality may lengthen processing times, especially for male examiners, though its overall explanatory power is minimal.

Closeness-Gender interaction:

```
# Closeness centrality model with interaction
closeness_gender_interaction=lm(
  app_proc_time ~ closeness_centrality.x * gender.x + race.x + tenure_days.x,
  data = applications_clean
)
summary(closeness_gender_interaction)

##
## Call:
## lm(formula = app_proc_time ~ closeness_centrality.x * gender.x +
##     race.x + tenure_days.x, data = applications_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2554.3  -441.9  -118.8   306.2  5000.1
##
## Coefficients:
```

```
##                                          Estimate   Std. Error    t value    Pr(>|t|)
## (Intercept)                              1.300e+03   2.279e+00   570.570    < 2e-16  ***
## closeness_centrality.x                  -1.172e+02   4.013e+00   -29.199    < 2e-16  ***
## gender.xmale                             3.598e+01   2.256e+00    15.950    < 2e-16  ***
## race.xblack                              1.963e+01   4.772e+00     4.113   3.91e-05  ***
## race.xHispanic                           1.932e+01   5.765e+00     3.351   0.000806  ***
## race.xother                              2.341e+01   3.612e+01     0.648   0.516840
## race.xwhite                             -6.185e+01   1.921e+00   -32.198    < 2e-16  ***
## tenure_days.x                            9.487e-05   9.077e-06    10.451    < 2e-16  ***
## closeness_centrality.x:gender.xmale     -1.740e+01   4.856e+00    -3.582   0.000341  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646.2 on 598623 degrees of freedom
##     (231685 observations deleted due to missingness)
## Multiple R-squared:  0.008649,          Adjusted R-squared:  0.008636
## F-statistic: 652.8 on 8 and 598623 DF,  p-value: < 2.2e-16
```

Explanation on Closeness-Gender interaction

The closeness centrality and gender interaction model show that closeness centrality typically reduces processing times, but this effect is less pronounced for male examiners.

Combined-Gender interaction:

```
# Combined model with interaction
combined_gender_interaction=lm(
   app_proc_time ~ (degree_centrality.x + betweenness_centrality.x +
closeness_centrality.x) * gender.x + race.x + tenure_days.x,
data = applications_clean
)
summary(combined_gender_interaction)

##
## Call:
## lm(formula = app_proc_time ~ (degree_centrality.x + betweenness_centrality.x +
##         closeness_centrality.x) * gender.x + race.x + tenure_days.x,
##         data = applications_clean)
##  Residuals:
##        Min        1Q    Median       3Q       Max
##  -2555.1   -441.7       -118.3      305.9    4999.7
##
##  Coefficients:
##                                         Estimate      Std. Error    t value  Pr(>|t|)
##  (Intercept)                            1.297e+03     2.625e+00    494.104   < 2e-16
##  degree_centrality.x                    3.374e-01     5.395e-02      6.254  4.00e-10
##  betweenness_centrality.x              -1.719e-03     2.218e-04     -7.750  9.21e-15
##  closeness_centrality.x                -1.134e+02     4.269e+00    -26.555   < 2e-16
##  gender.xmale                           4.337e+01     2.697e+00     16.083   < 2e-16
##  race.xblack                            1.800e+01     4.774e+00      3.770  0.000163
##  race.xHispanic                         1.955e+01     5.766e+00      3.390  0.000700
##  race.xother                            2.451e+01     3.611e+01      0.679  0.497290

##  race.xwhite                           -6.226e+01   1.922e+00 -32.391          < 2e-16
## tenure_days.x                          9.373e-05  9.076e-06  10.327   < 2e-16
## degree_centrality.x:gender.xmale      -6.815e-01  6.158e-02 -11.066   < 2e-16
## betweenness_centrality.x:gender.xmale  3.078e-03  2.641e-04  11.657   < 2e-16
## closeness_centrality.x:gender.xmale   -2.452e+01  5.132e+00  -4.779  1.77e-06
##
## (Intercept)                           ***
## degree_centrality.x                   ***
## betweenness_centrality.x              ***
## closeness_centrality.x                ***
## gender.xmale                          ***
## race.xblack                           ***
## race.xHispanic                        ***
## race.xother
## race.xwhite                           ***
## tenure_days.x                         ***
## degree_centrality.x:gender.xmale      ***
## betweenness_centrality.x:gender.xmale ***
## closeness_centrality.x:gender.xmale   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 646 on 598619 degrees of freedom
##   (231685 observations deleted due to missingness)
## Multiple R-squared:  0.009132,	Adjusted R-squared:  0.009112
## F-statistic: 459.7 on 12 and 598619 DF,  p-value: < 2.2e-16
```

Explanation on Combined-Gender interaction

The model that combines all centrality measures and their interactions with gender reveals complex effects, with gender moderating these impacts, yet it still fails to significantly enhance explanatory power.

Conclusion

In summary, although gender modifies the influence of centrality on processing times, the low adjusted R-squared values across all models indicate that these variables alone do not effectively predict processing times, highlighting the need for a more comprehensive model to fully grasp the dynamics affecting processing times at the USPTO.