ORGB 672: Exercise 4

Impact of Examiner Centrality and Gender on Patent Processing Times

Meriem Mehr

2024-04-07

Introduction & Data Preprocessing

This section outlines the steps undertaken in the preprocessing of the dataset used for our analysis. In the preceding assignment, we addressed the issue of missing values within the dataset, specifically focusing on the gender and race columns, and estimated the tenure of examiners. These preparatory measures were critical for ensuring the integrity and reliability of our dataset, resulting in a comprehensively preprocessed file named "applications.csv". For the purposes of this assignment, this file serves as the foundation upon which our subsequent analyses are built.

Data Loading and Initial Setup

Upon commencing this assignment, the first step involved loading the "applications.csv" dataset. This dataset is pivotal as it contains the preprocessed data required for our analysis, encompassing information on patent applications managed by the examiners. The significance of loading this dataset promptly cannot be overstated, as it provides the basis for all subsequent analytical procedures.

```
applications <- read.csv("C:/Users/mehri/Downloads/applications.csv")
attach(applications)
```

Calculating Application Processing Time

An essential aspect of our analysis was to accurately calculate the application processing time. This calculation necessitated the removal of any rows lacking concrete dates for when patents were issued or abandoned, thereby preventing any potential inaccuracies stemming from null values. To streamline this process, we ingeniously amalgamated the columns indicating the issue and abandonment dates into a singular variable termed "decision date." This variable represents the conclusive day on which an application was either approved (patent issued) or dismissed (patent abandoned). Utilizing the "decision date," we then calculated the processing time by determining the interval between the application's filing date and its decision date.

```
# Installing & Loading Relevant Packages Here:
library(readr)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v purrr     1.0.2
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## -- 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(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.3.3
```

```
##
## 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(gender)
```

```
## Warning: package 'gender' was built under R version 4.3.3
```

```
library(wru)
```

```
## Warning: package 'wru' was built under R version 4.3.3
```

```
##
## 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(lubridate)
library(dplyr)
library(gtsummary)
```

```
## Warning: package 'gtsummary' was built under R version 4.3.3
```

```
library(arrow)
```

```
## Warning: package 'arrow' was built under R version 4.3.2
```

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

```
library(tidyr)
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(purrr)
library(babynames)
```

```
## Warning: package 'babynames' was built under R version 4.3.3
```

```
library(rethnicity)
```

```
## Warning: package 'rethnicity' was built under R version 4.3.3
```

```
## == WARNING! ========================================================================
## The method provided by this package has its limitations and anyone must use them cautiously and respo
## ====================================================================================
```

Geting rid of rows without NA in patent_issue_date and abandon_date

```
# Removal of rows with NA in patent_issue_date and abandon_date and creation of decision_date
applications <- applications %>%
  filter(!(is.na(patent_issue_date) & is.na(abandon_date)))
applications <- applications %>%
  mutate(
    decision_date = coalesce(patent_issue_date, abandon_date),
    filing_date = ymd(filing_date),
    decision_date = ymd(decision_date),
    app_processing_time = as.numeric(difftime(decision_date, filing_date, units = "days"))
  )
```

This initial phase of our analysis was critical for setting a solid foundation for our study. By meticulously preparing our dataset and establishing a robust methodology for calculating application processing times, we ensured the accuracy and reliability of our subsequent analyses. These steps reflect a thoughtful and systematic approach to addressing the research questions posed by our study, highlighting our commitment to rigor and precision in academic research.

Addressing Null Values

A fundamental step in ensuring the integrity of our dataset involved scrutinizing it for null values, particularly in the columns earmarked for analysis. Given the pivotal role of gender in our study, identifying and addressing any null values in this column was imperative. To this end, a systematic check was conducted to ascertain the presence of null values across the dataset.

```
# Identification of null values in each column
na_count_by_column <- colSums(is.na(applications))
print(na_count_by_column)
```

```
##   application_number          filing_date   examiner_name_last
##                    0                    0                    0
##  examiner_name_first examiner_name_middle          examiner_id
##                    0                 1904                   21
##    examiner_art_unit           uspc_class        uspc_subclass
##                    0                    0                    7
##        patent_number    patent_issue_date         abandon_date
##                 2943                 2940                 5393
##        disposal_type     appl_status_code     appl_status_date
```

```
##                   0                   0                   0
##                  tc              gender                race
##                   0                1236                   0
##        earliest_date         latest_date         tenure_days
##                   0                   0                   0
##        decision_date app_processing_time
##                   0                   0
```

Despite previous preprocessing efforts, we identified 1236 instances where gender data was missing. Acknowledging the criticality of this variable for our analysis, it was decided that rows with null gender values would be excluded from the dataset. This decision was made to ensure the completeness and accuracy of our analysis, especially since gender plays a crucial role in our subsequent examinations.

```
# Exclusion of rows with null values in the gender column
applications <- applications %>%
  filter(!is.na(gender))

# Reassessment of null values post-exclusion
na_count_by_column <- colSums(is.na(applications))
print(na_count_by_column)
```

```
##   application_number          filing_date   examiner_name_last
##                    0                    0                    0
##  examiner_name_first examiner_name_middle         examiner_id
##                    0                 1284                    0
##    examiner_art_unit          uspc_class       uspc_subclass
##                    0                    0                    0
##        patent_number    patent_issue_date         abandon_date
##                 2512                 2510                 4587
##        disposal_type     appl_status_code     appl_status_date
##                    0                    0                    0
##                   tc              gender                race
##                    0                    0                    0
##        earliest_date         latest_date         tenure_days
##                    0                    0                    0
##        decision_date app_processing_time
##                    0                    0
```

Aligning Application and Examiner Data

With the dataset now refined, attention was turned towards ensuring consistency between the applications data and the examiner network data. This alignment is crucial for the accurate calculation of centrality measures. Therefore, we proceeded to filter the examiner network data to retain only those records that correspond to the application numbers present in our refined applications dataset.

```
edges_sample <- read.csv("C:/Users/mehri/Downloads/edges_sample.csv")
# Ensuring consistency between application numbers in both datasets
filtered_edges <- edges_sample %>%
  filter(application_number %in% applications$application_number)
```

Centrality Measures Calculation

To explore the network dynamics within the dataset, the preparation for calculating centrality measures commenced. Centrality measures are instrumental in understanding the positions and influence of examiners within the network. A directed graph, which requires the assignment of weights to each examiner pair, was proposed as the methodological framework for this analysis. The weights, determined by the application number and grouped by the examiners' IDs, serve to quantify the connections within the network, thereby laying the groundwork for a nuanced analysis of centrality.

The steps taken in addressing null values, ensuring data consistency, and preparing for centrality measures calculation underscore the rigorous approach adopted in this study. These preparatory actions are critical for the integrity of the analysis, enabling us to proceed with confidence into the centrality measures calculation and further statistical examinations.

Network Construction with Directed Graphs

The construction of the examiner network marked a pivotal phase in our analysis. Utilizing the igraph package, we embarked on creating a directed graph that encapsulates the intricate relationships between patent examiners. This approach necessitated the computation of weights for each examiner pair, which were derived from the distinct application numbers associated with each pair. These weights serve as a quantitative measure of the connections between examiners, reflecting the intensity of their interactions within the patent examination process.

```
# Installation and loading of the igraph package
install.packages("igraph")
```

```
## Warning: package 'igraph' is in use and will not be installed
```

```
library(igraph)

# Calculation of weights for the directed graph
edges <- edges_sample %>%
  group_by(ego_examiner_id, alter_examiner_id) %>%
  summarise(weight = n_distinct(application_number), .groups = "drop") %>%
  rename(ego = ego_examiner_id, alter = alter_examiner_id)

# Construction of the directed graph
g <- graph_from_data_frame(d=edges, directed=TRUE, vertices=NULL)
```

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

With the directed graph established, the calculation of centrality measures ensued. These measures—namely, degree (both in-degree and out-degree), betweenness, and closeness centrality—offer insights into the roles and positions of individual examiners within the network. The inclusion of weights in the calculation of these measures enhances the precision of our analysis, enabling a nuanced understanding of the examiners' influence and connectivity.

```
# Calculation of centrality measures
in_degree <- degree(g, mode="in")
out_degree <- degree(g, mode="out")
betweenness <- betweenness(g, directed=TRUE)
```

```r
in_closeness <- closeness(g, mode="in", weights=E(g)$weight)
out_closeness <- closeness(g, mode="out", weights=E(g)$weight)

# Compilation of centrality measures into a dataframe
centrality_measures <- data.frame(
  examiner_id = V(g)$name,
  in_degree_centrality = in_degree,
  out_degree_centrality = out_degree,
  betweenness_centrality = betweenness,
  in_closeness_centrality = in_closeness,
  out_closeness_centrality = out_closeness
)
```

Linear Regression Analysis Without Gender

In anticipation of conducting linear regression analysis to explore the impact of centrality on application processing times, it was imperative to integrate the centrality measures with the examiner data. This integration ensures a cohesive dataset that aligns the centrality measures with the corresponding examiners, thereby facilitating a comprehensive analysis. The transformation of examiner IDs to character types and the subsequent merging of datasets are critical steps in this preparatory process.

```r
# Integration of centrality measures with examiner data
applications <- applications %>%
  mutate(examiner_id = as.character(examiner_id))
centrality_measures <- centrality_measures %>%
  mutate(examiner_id = as.character(examiner_id))
joined_data <- applications %>%
  left_join(centrality_measures, by = "examiner_id")
```

Regression Model Deployment

Utilizing the integrated dataset comprising applications and their corresponding centrality measures, we employed a linear regression model. This model aims to quantify the relationship between five centrality measures (in-degree, out-degree, betweenness, in-closeness, and out-closeness) and the application processing time.

```r
linear <- lm(app_processing_time ~ in_degree_centrality + out_degree_centrality + betweenness_centrality
```

A summary of this model provided insights into the significance and magnitude of each centrality measure's impact on processing time.

Interpretation of Regression Coefficients

- **Intercept (1,212 days)** indicates the estimated processing time when centrality measures are absent. This figure is notably significant, underscoring the baseline processing period in the absence of network effects.
- **In-degree Centrality** suggests a slight increase in processing time (approximately 6.07 days) per unit increase in centrality. Despite its marginal significance, this finding hints at the potential for more central examiners to experience longer processing times, possibly due to a higher workload.
- **Out-degree Centrality** shows a negligible decrease in processing time, though the effect lacks statistical significance, indicating minimal impact on efficiency.

- **Betweenness Centrality** reveals an almost imperceptible increase in processing time per unit increase, again lacking statistical significance. This measure's minimal effect suggests that being a conduit between examiner groups does not substantially affect processing duration.
- **In-closeness Centrality** also does not significantly affect processing times, with a slight increase per unit increase in centrality.
- **Out-closeness Centrality**, contrastingly, demonstrates a significant reduction in processing time (113 days) per unit increase, highlighting the efficiency benefit of being closely connected to numerous network members.

Assessment of Model Fit

- The **Residual Standard Error** of 642.3 days illustrates the considerable variation in processing times not accounted for by the model, indicating other factors at play.
- The **Multiple R-squared** value of 0.007282 reveals that a very small portion of the variance in processing time is explained by the centrality measures, suggesting the model's limited explanatory power.
- Despite these limitations, the **F-statistic** value suggests a statistically significant relationship between centrality measures and processing time, albeit with a low overall impact.

Key Takeaways

The linear regression analysis unveils nuanced insights into the relationship between examiners' network centrality and patent application processing times. Notably, out-closeness centrality emerges as a significant efficiency enhancer, suggesting that examiners who maintain close connections across the network can expedite processing times. However, the overall explanatory power of the centrality measures is modest, indicating that other unexamined factors likely play a substantial role in determining processing times. This analysis sets the stage for further exploration into the myriad factors influencing patent processing efficiency, offering a foundational understanding of the role of network dynamics within the USPTO.

Linear Regression Analysis With Gender

Incorporating Gender into the Regression Model

To explore the potential impact of gender on the interactions between patent examiners and application processing time, we augmented our linear regression model to include gender as a variable. This addition allows for a nuanced examination of how gender, both independently and in conjunction with centrality measures, may influence the processing times of patent applications.

```
linear_gender <- lm(app_processing_time ~ in_degree_centrality*gender + out_degree_centrality*gender +
    betweenness_centrality*gender + in_closeness_centrality*gender + out_closeness_centrality*gender,
    data = joined_data)
```

A comprehensive summary of this model shed light on the intricate dynamics between gender, centrality, and processing times.

Unveiling the Impact of Gender and Centrality

- **Baseline Processing Time**: The intercept indicates an average processing time of 1150 days under baseline conditions (all centrality measures at zero), assuming the examiner is of the baseline gender category. This finding underscores the significant duration required for patent processing in the absence of network effects.

- **Centrality Measures**: The effects of centrality measures on processing times appear nuanced, with most showing no significant impact when gender interactions are considered. Notably, out-closeness centrality's influence on reducing processing times in the initial model becomes statistically insignificant upon the introduction of gender, suggesting the complexity of these interactions.
- **Gender Differences**: The model introduces a gender variable, highlighting a non-significant difference in processing times between male and female examiners. This finding prompts further investigation into the role of gender within the organizational dynamics of the USPTO.

Interaction Effects: Gender and Centrality

- The interaction terms between gender and centrality measures offer insights into how gender may modify the influence of network position on processing times. Although these interactions are not statistically significant, they hint at underlying patterns that merit further exploration. For instance, the interaction between in-degree centrality and gender suggests that the relationship between an examiner's network position and processing efficiency could differ across genders.

Analyzing Model Fit and Significance

- The **Residual Standard Error** remains consistent with the previous model, indicating persistent unexplained variability in processing times.
- The **Multiple R-squared** value sees a slight increase, suggesting a marginally improved explanation of variance in processing times through the inclusion of gender. However, the overall explanatory power remains limited.
- The **F-statistic** and its associated p-value do not reach conventional levels of statistical significance, indicating that while the model introduces interesting variables for consideration, it does not conclusively demonstrate the combined effect of centrality measures and gender on processing times.

Implications and Reflections

This expanded analysis introduces gender as a crucial variable in understanding the dynamics of patent processing times within the USPTO. Although the direct impact of gender and its interactions with centrality measures on processing times is not statistically significant in this model, the exploration opens avenues for deeper investigation into organizational behaviors and processes. It suggests that factors beyond the scope of this study, possibly including workload distribution, examiner experience, and application complexity, play critical roles in shaping processing times. This insight encourages a broader examination of the USPTO's operational efficiency, emphasizing the need for comprehensive strategies that address the multifaceted nature of patent examination processes.

Concluding Insights and Strategic Recommendations for the USPTO

Overview of Findings

Our investigation into the US Patent and Trademark Office (USPTO) patent processing times, through the lens of organizational network analysis, unveiled nuanced relationships between examiner centrality within the patent citation network, gender, and processing efficiency. Key findings indicate:

- **Centrality's Impact**: A modest correlation exists between certain centrality measures, particularly out-closeness centrality, and processing efficiency. This implies that examiners well-integrated within the network may leverage their positions for more effective information dissemination and acquisition, potentially streamlining the application review process.

- **Gender Dynamics**: The inclusion of gender as a variable, along with its interaction with centrality measures, did not significantly alter processing times. This suggests that operational efficiency at the individual level might be influenced more by experience, workload, and accessibility to resources rather than the examiner's gender.

Implications for Operational Enhancements

These insights offer the USPTO actionable intelligence on optimizing processing times and enhancing examiner efficiency. Recommended strategies include:

1. **Fostering Network Integration through Training**: Encouraging deeper integration of examiners into the patent citation network could be instrumental. Tailored training programs emphasizing knowledge sharing and collaborative practices could enrich examiners' network positions, fostering a culture of efficiency and mutual support.
2. **Focus on Expertise and Technological Specialization**: The absence of gender-related disparities in processing efficiency underscores the importance of focusing on skill development and specialization. Investing in specialized training that aligns with the complexity and technological demands of patent applications can ensure examiners are well-equipped to handle their caseloads effectively.
3. **Leveraging Technology for Enhanced Connectivity**: The USPTO stands to benefit from technology solutions that bolster examiners' access to prior art and other pertinent information. By enhancing the technological infrastructure to facilitate easier access to essential resources, examiners can potentially expedite the processing times without compromising the quality of patent examination.

Moving Forward

The findings from this study provide a foundation for strategic improvements within the USPTO, highlighting the potential of network centrality as a lever for enhancing operational efficiency. While the direct influence of gender on processing times was not substantiated, the analysis reinforces the value of focusing on broader organizational and procedural factors that contribute to processing efficiency. By adopting a holistic approach that includes targeted training, technological enhancements, and an emphasis on examiner specialization, the USPTO can aim to not only improve processing times but also maintain the integrity and quality of patent examination processes.