Homophily in Professional Advice Networks at the USPTO

Meriem Mehri

`r Sys.Date()`

Introduction

This analysis delves into the intricate web of professional advice networks within the US Patent and Trademark Office (USPTO), exploring the prevalence of homophily—the tendency of individuals to associate with similar others. By meticulously examining the relationships between patent examiners in terms of their gender, race, and tenure, this report sheds light on the dynamics of advice-seeking behavior and its potential implications for organizational knowledge flow and diversity.

Setup and Data Loading

The first step in our analysis involves setting up the R environment and loading the necessary libraries. These libraries provide a comprehensive toolkit for data manipulation (`tidyverse`), network analysis (`igraph`, `tidygraph`), and visualization (`ggraph`, `ggplot2`).

```
{r setup, message=FALSE, warning=FALSE} library(tidygraph) library(tidyverse) library(igraph) library(ggplot2) library(vroom) library(arrow) library(scales) library(ggraph) library(ggtext) library(ggrepel) library(ggforce) library(ggthemes) library(patchwork) library(qualpalr) library(gender) library(wru) library(skimr)
```

Data Import

Here, we import the USPTO patent examiner data and the advice network edges. The app_data_sample.parquet file contains examiner demographic information and patent details, while the edges_sample.csv file details the advice-seeking interactions among examiners.

```
{r import-data} app_data <- read_parquet("C:/Users/mehri/Downloads/app_data_sample.parquet") edge_data <- vroom("C:/Users/mehri/Downloads/edges_sample.csv", delim = ",")
```

Data Preparation

Data preparation is crucial for ensuring accuracy in the subsequent analysis. This section outlines the steps taken to estimate demographic variables, calculate tenure, and prepare the edge data for network construction.

Estimating Gender and Race

Understanding the demographic composition of the USPTO advice network is essential for analyzing homophily. While this analysis assumes direct availability of gender and race data, practical applications would entail estimating these variables using specialized R packages based on name data.

Note: Assuming gender and race data are directly available in the dataset. In practice, methods like `gender` and `wru` packages can be used for estimation based on names.

Calculating Tenure

Tenure, indicative of an examiner’s experience, is calculated as the time span between their first and last patent application. This variable offers insights into the potential influence of experience on advice-seeking patterns.

```
{r calculate-tenure} app_data <- app_data %>%   group_by(examiner_id) %>%   mutate(tenure = max(filing_date) - min(filing_date)) %>%   ungroup()
```

Preparing Edge Data

The edge data, representing advice-seeking interactions, is refined to ensure compatibility with the network construction process.

```
{r prepare-edge-data} edge_data <- edge_data %>%   select(ego_examiner_id, alter_examiner_id, advice_date)
```

Network Construction and Analysis

The construction and analysis of the advice network are pivotal to understanding the structural dynamics within the USPTO. This process involves validating the data, ensuring that all referenced nodes are present, and then constructing the graph for analysis.

Validate and Prepare Data for Network Construction

This preparatory step is designed to identify and rectify any discrepancies between the edge and node datasets, ensuring a smooth network construction process.

Ensure Compatibility Between Edge and Vertex Data

Before constructing the network, it’s crucial to ensure that all IDs referenced in the edge data have corresponding entries in the vertex data. This step will prevent errors related to missing vertex names during the network construction process.

Preparing Data for Network Construction

“‘{r preparing-data-for-network} ## Identify and Add Missing Examiners

Convert all IDs to character to ensure consistent matching

app_data$examiner_i d<-as.character$¿examiner_id) edge_data $ego_e xaminer_i d<-as.character$¿ego_examiner_id) edge_data $alter_e xaminer_i d<-as.character$¿alter_examiner_id)

Identify unique examiner IDs in edge data

all_edge_ids <- unique(c(edge_data$ego_e xaminer_i d, edge_d ata$alter_examiner_id))

Identify IDs in edge data not present in app data

missing_ids <- setdiff(all_edge_ids, app_data$examiner_id)

Create 'missing_examiners' with the same columns as 'app_data'

missing_examiners <- tibble(examiner_id = missing_ids, gender = NA_character_, # Adjust based on actual data type in 'app_data' race = NA_character_, # Adjust based on actual data type in 'app_data' tenure = NA_real_ # Adjust based on actual data type in 'app_data' # Add more columns here as placeholders to match the structure of 'app_data')

Ensure 'missing_examiners' has all the columns present in 'app_data', filled with NA

additional_columns <- setdiff(names(app_data), names(missing_examiners)) for (col in additional_columns) { missing_examiners[[col]] <- NA }

Ensure no duplicate examiner entries

app_data <- app_data %>% distinct(examiner_id, .keep_all = TRUE)

~~~~
```{r network-construction}
# Attempt to construct the network with the updated app data
g <- igraph::graph_from_data_frame(d = edge_data, vertices = app_data,
directed = TRUE)
~~~~

Demographic Analysis

Exploring the demographic attributes within the advice network provides insights into patterns of homophily. This section compares the gender, race, and tenure distributions across the network.

Gender Distribution

Gender diversity within the advice network is analyzed to understand the balance and potential biases in advice-seeking behavior.

```
{r gender-dist} gender_dist <- app_data %>%   group_by(gender) %>%   summarise(count = n())
```

Race Distribution

Race distribution analysis sheds light on the racial diversity within the network, highlighting areas of homogeneity and diversity.

```
{r race-dist} race_dist <- app_data %>%   group_by(race) %>%   summarise(count = n())
```

Tenure Distribution

Examining the distribution of tenure among the examiners reveals insights into the experience levels present within the advice network.

```
{r tenure-dist} tenure_dist <- app_data %>%   summarise(mean_tenure = mean(tenure), sd_tenure = sd(tenure))
```

Network Visualization

Visualizing the network by demographic characteristics allows us to visually assess patterns of homophily and the structural positions of various groups within the network.

```
{r network-visualization, fig.cap="USPTO Advice Network by Gender"} set.seed(123) ggraph(g, layout = "fr") +   geom_edge_link() +   geom_node_point(aes(color = gender)) +   scale_color_manual(values = c("Male" = "blue", "Female" = "pink")) +   theme_graph() +   labs(title = "USPTO Advice Network", color = "Gender")
```

Conclusion

This analysis highlights the role of homophily in shaping the professional advice networks at the USPTO. By understanding these patterns, organizations can foster more inclusive and effective knowledge-sharing environments. The insights derived from this study underscore the importance of considering demographic diversity in organizational network analysis to promote a more collaborative and equitable workplace.