**ORGB 672**

**ORGANIZATIONAL NETWORK ANALYSIS**

**Presented to Professor Roman Galperin**

**Zhicheng Zhong (261143399)**

**Tashfeen Ahmed (261145667)**

**Jingwen He (260897181)**

**Xinran Yu (260922576)**

**Apr 22, 2024**

**Introduction (Datasets, methodology)**

The study aims to uncover the organizational, social and demographic factors that impact patent processing time at The United States Patent and Trademark Office (USPTO). The USPTO dataset consists of 1537478 rows and 16 columns. The original variables are 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 10)patent\_number 11)patent\_issue\_date 12)abandon\_date 13)disposal\_type 14)appl\_status\_code (application status code) 15)appl\_status\_date (application status date) 16)tc (technology center). To study the social and demographic impact on patent processing time, we used the babyname package to infer gender from examiners’ first name, the rethnicity package to infer ethnicity from examiners’ family name, the lubridate package to infer tenure days by calculating the interval between the minimum filing\_date and maximum appl\_status\_date. The patent processing time is calculated by subtracting filing\_date from either patent\_issue\_date or abandon\_date. Since there are more than 153 million records of patent application, the process would be extremely computationally and time consuming for fitting a tree-based model on all observations. To fit a random forest model on the data, we randomly sample 50000 records from the 153 million observations that exist in the dataset. This number is believed to be significantly large to capture patterns representative of the whole dataset.

The second dataset at our disposal is the Advice Network dataset which consists of 4 columns: 1)application\_number 2)advice\_date 3)ego\_examiner\_id 4)alter\_examiner\_id. The ego and alter examiner id can be joined to the examiner\_id column in the USPTO dataset. We obtained an examiner attributes table from USPTO dataset after having inferred gender, ethnicity and tenure days and joined it to the network dataset on the same examiner id for network visualization. Based on the observed patterns in the visualized network we calculate centrality measures of choice that are utilized in the regression analysis, where we used interaction terms to control for factors such as gender and ethnicity to study the relationship between process time and social factors such as centralities, demographic factors such as gender and race, as well as organizational factors such as technology center.

**EDA**

The histogram of application processing time depicts the distribution of patent application processing times, which is right-skewed, indicating that the majority of applications are processed in fewer days (Figure 2.1.1). The typical processing time hovers around 1000 days. However, there is a long tail of applications taking longer, with outliers evident beyond 2000 days. These outliers should be excluded for a more accurate model.

The scatter plot of degree and processing time reveals a subtle upward trend, suggesting that as the degree of interaction increases, the processing time tends to rise for both genders (Figure 2.1.2).

The analysis of gender and processing time indicates that both genders exhibit a similar mean processing time of approximately 1,000 days for patent applications, with outliers present for both. These outliers are considered for removal in subsequent analyses (Figures 2.2.1 and 2.2.2). From the perspective of race, the graph of the average processing time by race indicates that the average time taken to process patent applications is consistent across all races, with a mean of approximately 1250 days (Figure 2.3.2). In terms of workload, the Number of Applications by Race graph (Figure 2.3.3) indicates that white examiners have processed the most applications overall. However, when examining the average number of applications processed per race (Figure 2.3.4), the figures appear to be relatively consistent, suggesting an unbiased distribution in application assignment across races. However, the higher number of white examiners might indicate potential disparities in hiring or examiner selection. The "other" category is identified as an outlier and is therefore excluded from the analysis.

The network graphs for different races reveal distinct patterns of interaction among examiners. The network graphs for the white and Asian examiners demonstrate a high level of interaction. The white group’s graph is characterized by a central clustering, indicating a core of examiners frequently engaging with each other, which may reflect a strong community bond or centralized influence (Figure 2.4.1). The Asian graph exhibits a circular formation with numerous interlinking connections, suggesting a broad pattern of interaction where members are actively involved in a more evenly distributed communication network (Figure 2.4.2).

Conversely, the network graphs for the Black, Hispanic, and other racial categories indicate a paucity of interaction among their examiners. The Black and Hispanic graphs exhibit scattered nodes with minimal clustering, suggesting isolated interactions or lower engagement in collective discussion. (Figure 2.4.3, 2.4.4). The "Other" category graph, with its scant nodes, provides limited insight. For gender, male examiners demonstrate a centralized network with dense clustering, particularly evident among Asian and white males, indicating a high degree of interaction within their respective racial groups (Figure 2.5.1). Female examiners, while overall more dispersed, also exhibit clustering, most notably among white and Asian females (Figure 2.5.2).

The Mean Centrality Measures by Race graph reveals that Asian examiners have considerably high betweenness centrality, which may indicate their crucial role in transferring information across different clusters. White examiners also display high betweenness after Asians. All races demonstrate approximately the same degree of centrality, with Hispanics slightly higher, suggesting their active involvement in interactions (Figure 2.3.1). When examining centrality measures, both genders display comparable mean degree centrality, indicating that male and female examiners are equally engaged in information sharing within the organization. However, males exhibit a higher betweenness centrality, suggesting that they play a more pivotal role in facilitating information transfer throughout the organization by acting as intermediaries between separate clusters (Figure 2.2.3).

**Linear Regression and Random Forest Feature Importance Analysis**

To understand which variables are affecting application processing time and how a linear regression model and Random forest model was built. Features used in models are degree and betweenness centrality, gender, race and tenure days. Table 3.1 in appendix shows the results from linear regression.

**Explanation Linear Regression Results:**

**(Intercept)**: Represents the baseline patent prosecution time when all other variables are zero. It is significantly high, indicating a substantial processing time when no other factors are taken into account.

**log\_degree**: Exhibits a positive coefficient, indicating that as an examiner’s network connectivity increases,so does the processing time for patents.

**log\_betweenness**: Also has a positive coefficient, suggesting that examiners who frequently bridge communication between other examiners are associated with longer patent processing times.

**gendermale**: Shows a negative coefficient, which means male examiners are associated with shorter patent processing times compared to their female counterparts.

**raceblack**: This positive coefficient signifies that examiners identified as black are associated with longer processing times compared to the baseline race category (Asian).

**raceHispanic**: Although this has a positive coefficient, it’s not statistically significant, suggesting no clear association with processing time for Hispanic examiners.

**raceother**: Has a very large positive coefficient, indicating a strong association with increased processing times for examiners categorized in the ‘other’ race group.

**racewhite**: With a negative coefficient, white examiners are associated with shorter processing times compared to the baseline race category (Asian).

**tenure\_days**: A negative coefficient here indicates that with each additional tenure day, an examiner’s patent processing time decreases, suggesting more experienced examiners work faster.

**tc1700, tc2100, tc2400**: Each of these technology center variables has a positive coefficient, with ‘tc2100’ having the largest, meaning patent applications in these centers take longer to process, particularly in ‘tc2100’.

**log\_degree:gendermale**: The negative interaction term suggests the increase in processing time due to network degree is less for male examiners compared to female.

**log\_betweenness:gendermal**e: A positive coefficient here, though smaller, indicates male examiners with high betweenness might experience a slight increase in processing times.

**gendermale:tenure\_days**: This interaction term is not statistically significant, indicating tenure does not affect processing times for male examiners differently than for females.

**log\_degree:raceblack, log\_degree:raceHispanic, log\_degree:raceother, log\_degree:racewhite**: These terms all have negative coefficients ), This indicates that while increased network connectivity generally leads to longer patent processing times, its impact is mitigated for examiners of these racial groups compared to the baseline (Asian). Specifically, black, Hispanic, experience a reduced delay in processing times as their connectivity increases

**log\_betweenness:raceblack, log\_betweenness:raceHispanic, log\_betweenness:racewhite**: Here we see varying effects. The coefficient for log\_betweenness:raceblack is 2.14462, but it is not statistically significant (p = 0.704796), indicating that increased betweenness centrality does not significantly affect the processing times for black examiners compared to the baseline. On the other hand, log\_betweenness:raceHispanic shows a significant positive coefficient of 21.44965 (p = 1.16e-05), suggesting that Hispanic examiners with higher betweenness centrality experience notably longer processing times, highlighting a substantial impact of this network role on efficiency. For log\_betweenness:racewhite, the coefficient is significantly negative (-12.05342, p < 2e-16), indicating that increased betweenness actually leads to shorter processing times for white examiners, contrasting with the effect observed for Hispanic examiners

**raceblack:tenure\_days, raceHispanic:tenure\_days, racewhite:tenure\_days**: The term raceblack:tenure\_days has a significantly negative coefficient of -42.02431, indicating that each additional day of tenure substantially decreases the processing time for black examiners. This suggests that experience has a pronounced effect on improving efficiency among black examiners compared to the baseline group (Asian). In contrast, the interaction term raceHispanic:tenure\_days shows an insignificant positive coefficient of 1.16061, implying that tenure does not significantly affect processing times for Hispanic examiners. For racewhite:tenure\_days, the coefficient is also significantly negative (-14.46995), indicating that white examiners, like their black counterparts, experience reduced processing times with increased tenure, though the effect is less pronounced than for black examiners

In the linear regression analysis provided, the variables raceHispanic, gendermale:tenure\_days, log\_betweenness:raceblack, and raceHispanic:tenure\_days are statistically insignificant. This conclusion is based on their p-values, which are higher than the conventional significance level of 0.05. In simpler terms, changes in these variables do not significantly affect the application processing time, when controlling for other factors in the model. Race other variable are unknown races that were not able to be found in our analysis so in our analysis it can be ignored

**Explanation Random Forest Feature Importance:**

To understand which feature is most important in deciding in application processing time Feature importance was extracted from random forest model (Figure 3.1). Random Forest is particularly valuable because it can handle complex, non-linear relationships and interactions between variables that linear regression may not capture effectively. This model calculates feature importance by assessing how changes in each feature affect the model's accuracy, thereby identifying the most influential factors affecting processing times

**Technology Center (tc)**: This feature has the highest impact on MSE when permuted, indicating that the specific technology center handling the patent application is the most significant predictor of processing time. Variations between centers could reflect differences in process efficiency, case complexity, or resource allocation.

**Network Metrics (log\_betweenness, log\_degree)**: These features also show a substantial effect on the MSE, highlighting the significant role that an examiner’s position within the advice-sharing network has on patent prosecution times. Examiners who are central in the network or have more connections might be handling more complex applications or be more integral to the examination process, affecting timeframes.

**Tenure (tenure\_days)**: Examiner tenure is another influential factor, suggesting that more experienced examiners have a notable impact on prosecution times, potentially due to better familiarity with the patent process or greater efficiency in handling applications.

**Demographics (gender, race):** These features have a lesser but still meaningful impact on MSE when permuted, implying that demographic characteristics of the examiners do contribute to variations in processing times, though their influence is not as strong as organizational factors like technology center designation or network metrics

**Conclusion**

The comprehensive analysis conducted as part of our study on the United States Patent and Trademark Office (USPTO) processing times offers significant insights into the social and demographic factors that influence these durations. By utilizing datasets and modeling techniques, including linear regression and Random Forest models, the study provides an understanding of the factors influencing patent examination durations.

The study's key findings indicate that while technology centers show the most significant impact on processing times, network metrics like degree of connectivity and betweenness also play crucial roles. These network characteristics suggest that more interconnected or centrally positioned examiners influence processing times (Increase in processing times with increase in degree or betweeness) due to their roles in handling complex or voluminous patent applications. Demographic factors, including race and gender, also show varying impacts on processing times. For instance, male examiners have slightly lower processing times compared to female examiners. Linear regression shows slightly variance in processing times for different races and specific interaction effects which are marginal

The methodology employed in the study, particularly the use of random sampling for only random forest model to manage the vast USPTO dataset and the application of statistical tools to infer examiner demographics. The influence of technology centers is particularly pronounced, with certain centers showing significantly longer processing times. Moreover, tenure and experience play a beneficial role, with more seasoned examiners demonstrating quicker processing times. This supports the potential for targeted training and experience-building programs to enhance efficiency. This approach not only aids in identifying key predictors of processing times but also highlights areas for potential efficiency improvements and further research.

Building on our findings, we recommend several strategic initiatives. First, incorporating tailored training can significantly expedite examiners' familiarity with patent processes, enhancing overall efficiency. Furthermore, as examiners with higher centrality in the advice-sharing network often handle more complex applications or are overloaded, aim to evenly distribute network engagement among USPTO examiners, manage collaborative loads effectively using streamlined tools and protocols that prevent over-reliance on specific key individuals for information, and provide targeted support to high-centrality examiners to prevent bottlenecks and enhance overall organizational productivity improving network connectivity can help manage processing times more effectively. Additionally, addressing demographic disparities through diversity and inclusion training can mitigate unconscious biases and ensure equitable work distribution, thereby fostering fairness and efficiency. Furthermore, address inefficiencies in technology centers with longer processing times through resource reallocation or process reengineering.

Moving forward, it would be beneficial to incorporate a longitudinal perspective to observe how these relationships evolve over time, especially with shifts in USPTO policies or demographic changes, could further enrich our understanding of the social dynamics affecting patent examination processes. This approach would not only enhance predictive accuracy but also contribute to a more equitable and efficient patent system.

Overall, by continuing to refine these analytical frameworks, stakeholders can better strategize resource allocation, training, and policy-making to enhance the overall efficiency and fairness of the patent examination process.

**APPENDIX**

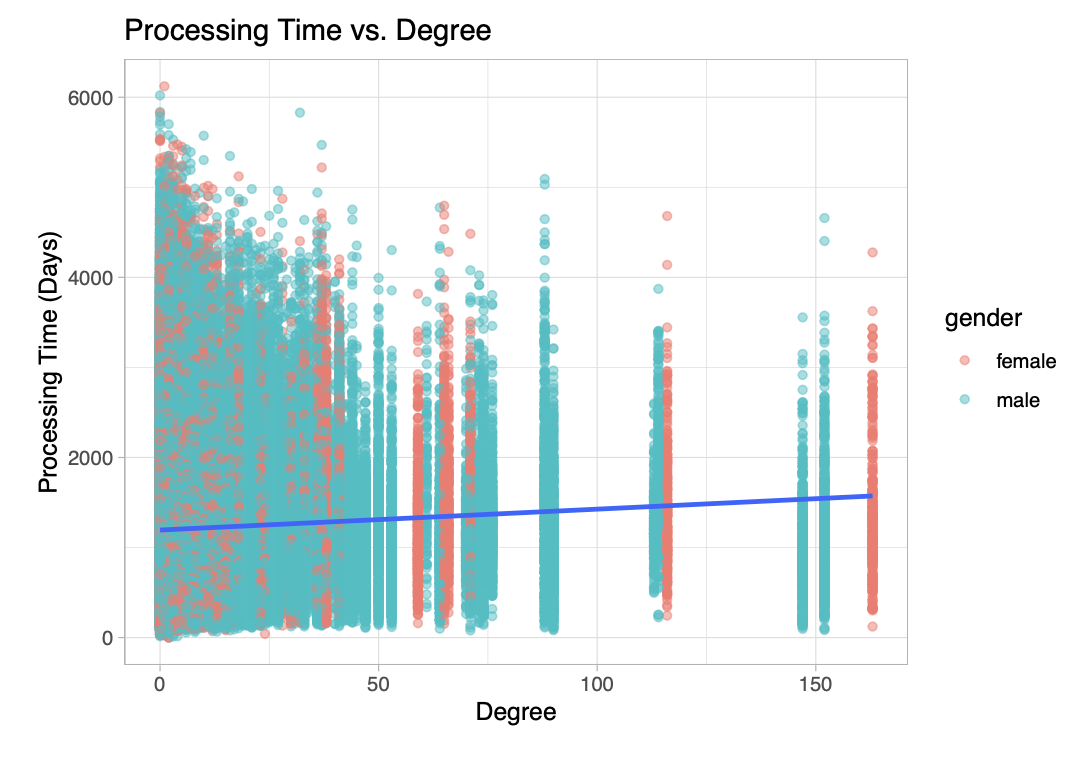
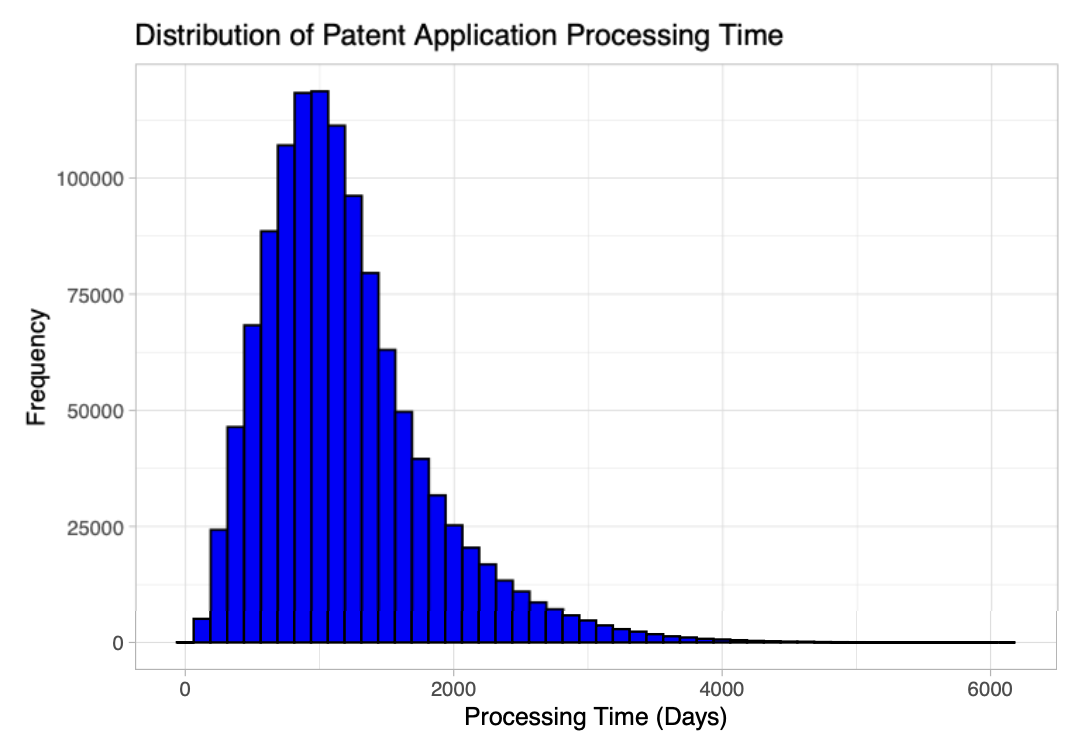


Figure 2.1.1 Histogram for Application Processing Time Figure 2.1.2 Scatter Plot and Trend for Application Processing Time and Betweenness

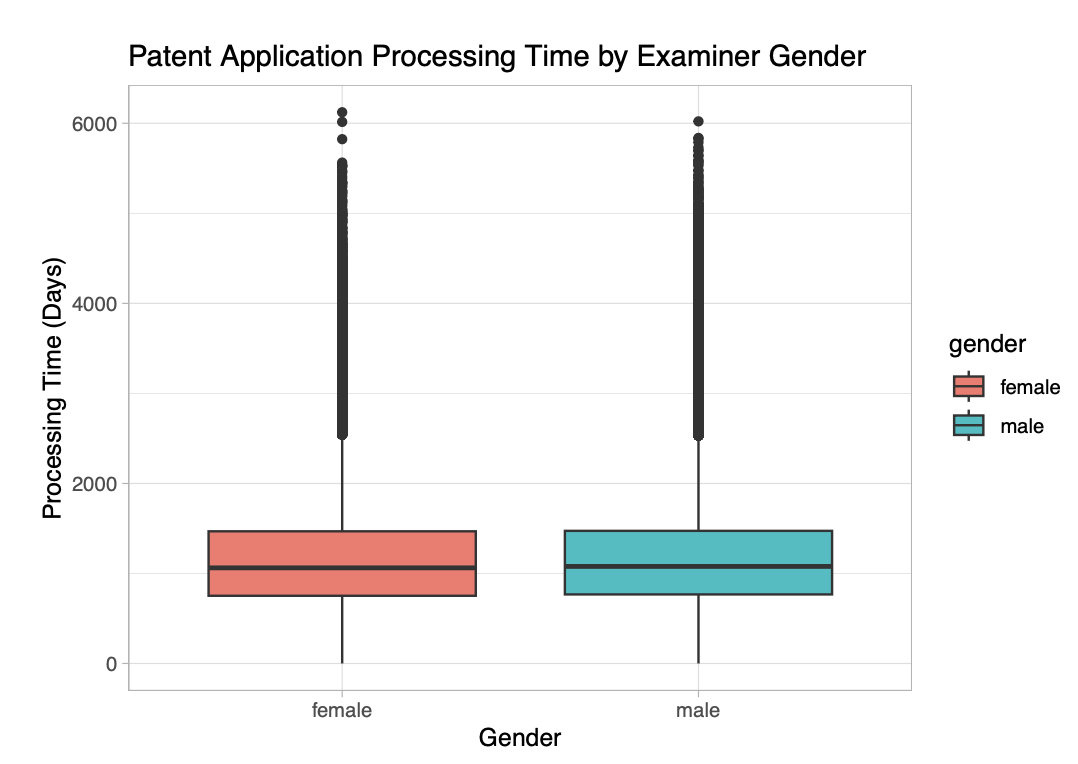
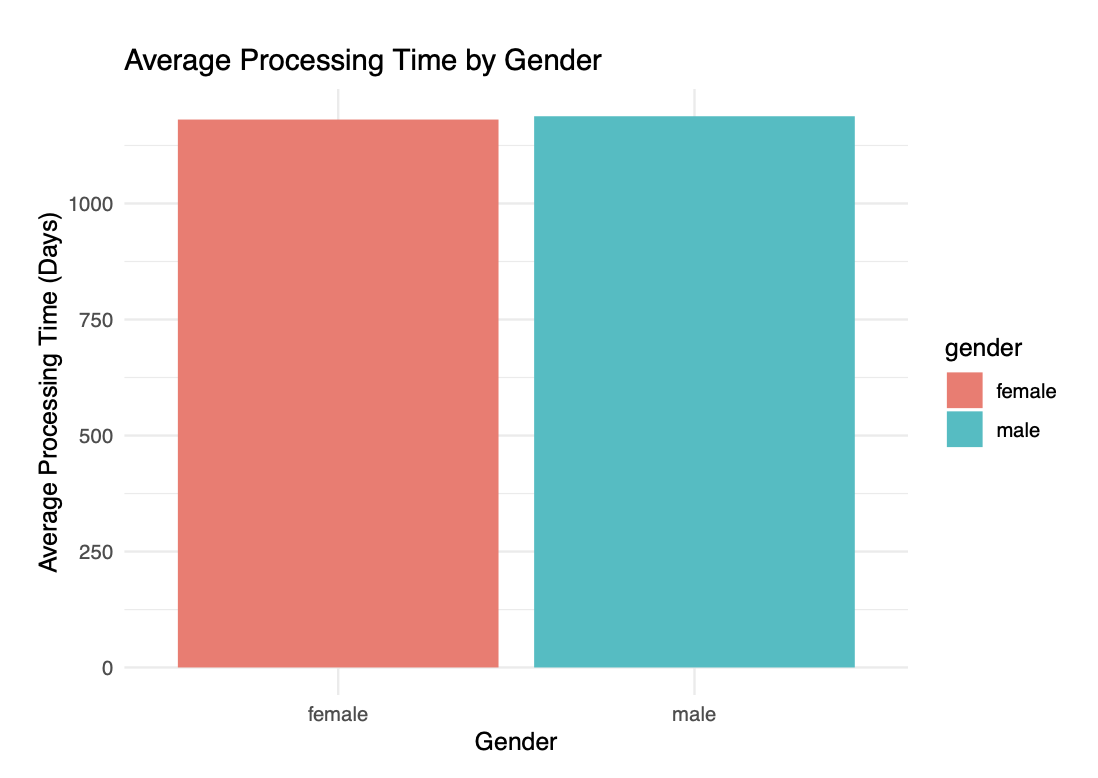


Figure 2.2.1 Boxplot for Patent Application Processing Time by Gender Figure 2.2.2 Average Processing Time by Gender

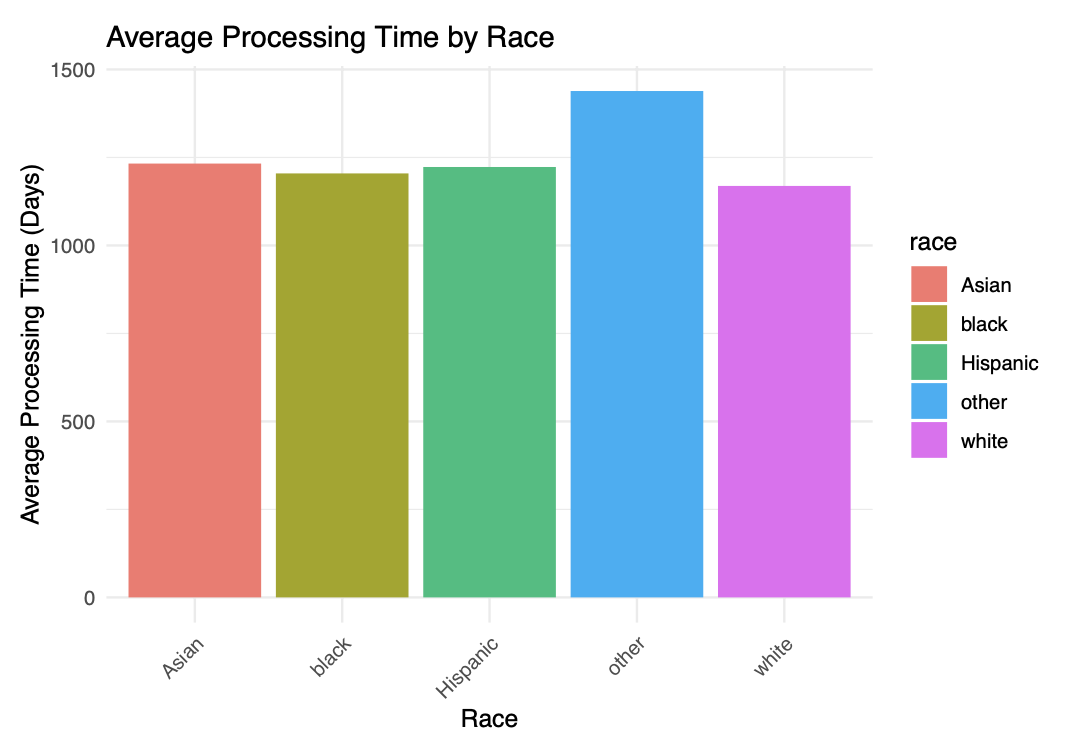
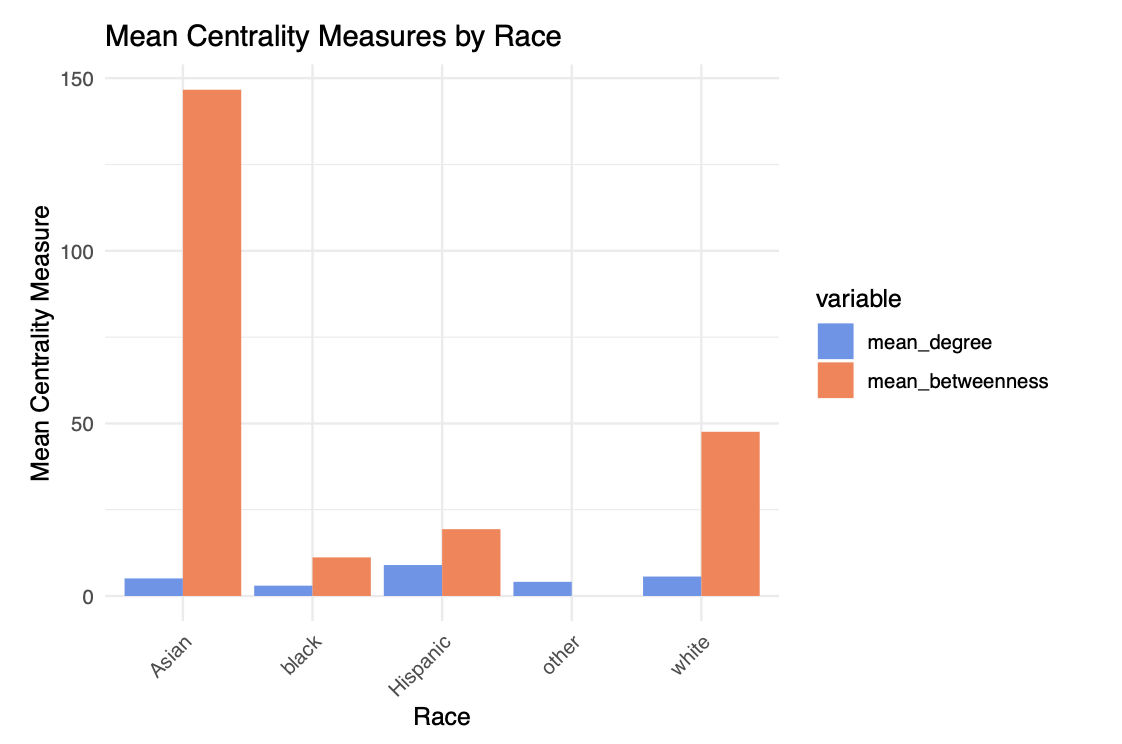
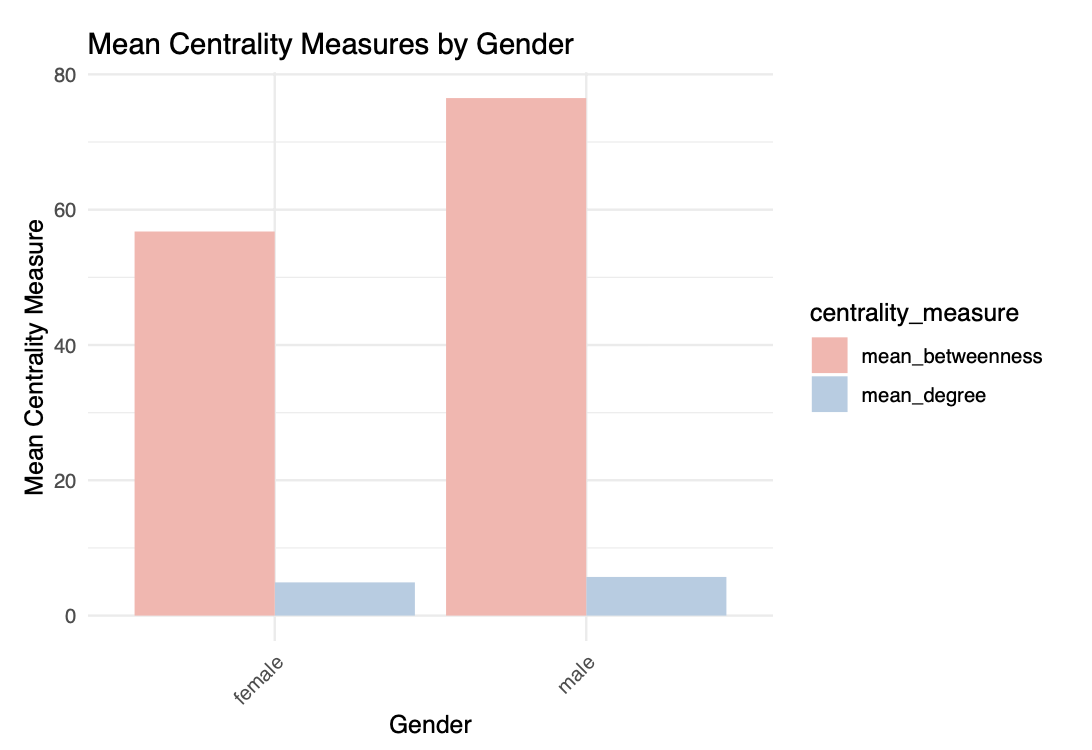


Figure 2.2.3 Mean Centrality Measures by Gender Figure 2.3.1 Mean Centrality Measures by Race Figure 2.3.2 Average Processing Time by Race

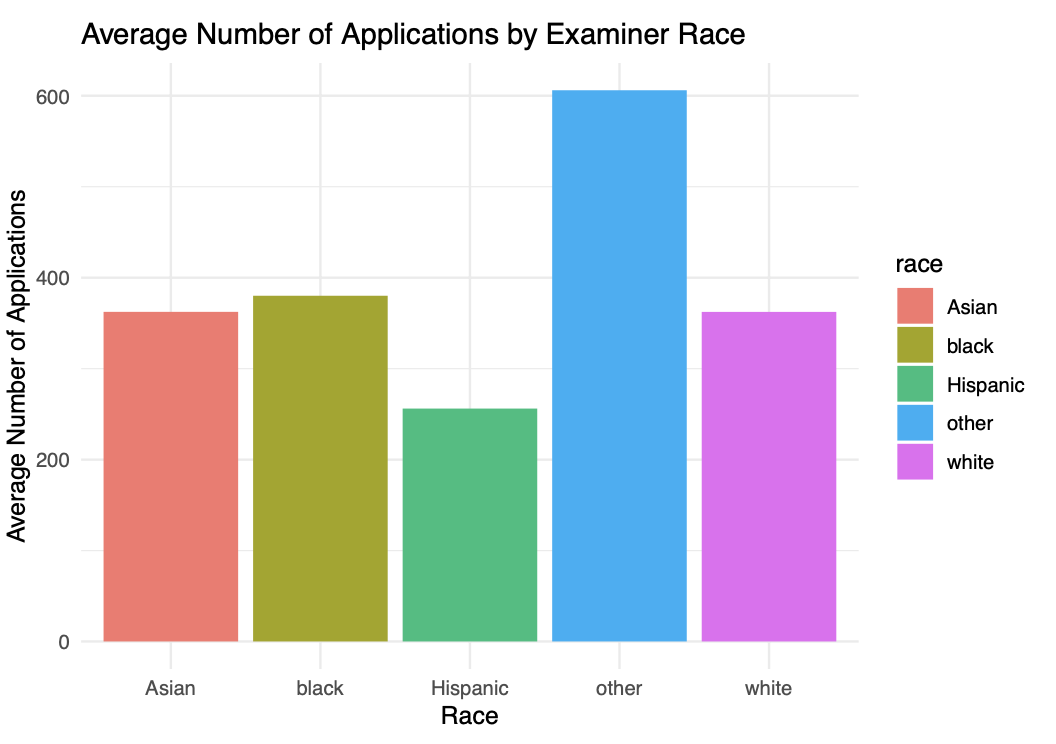
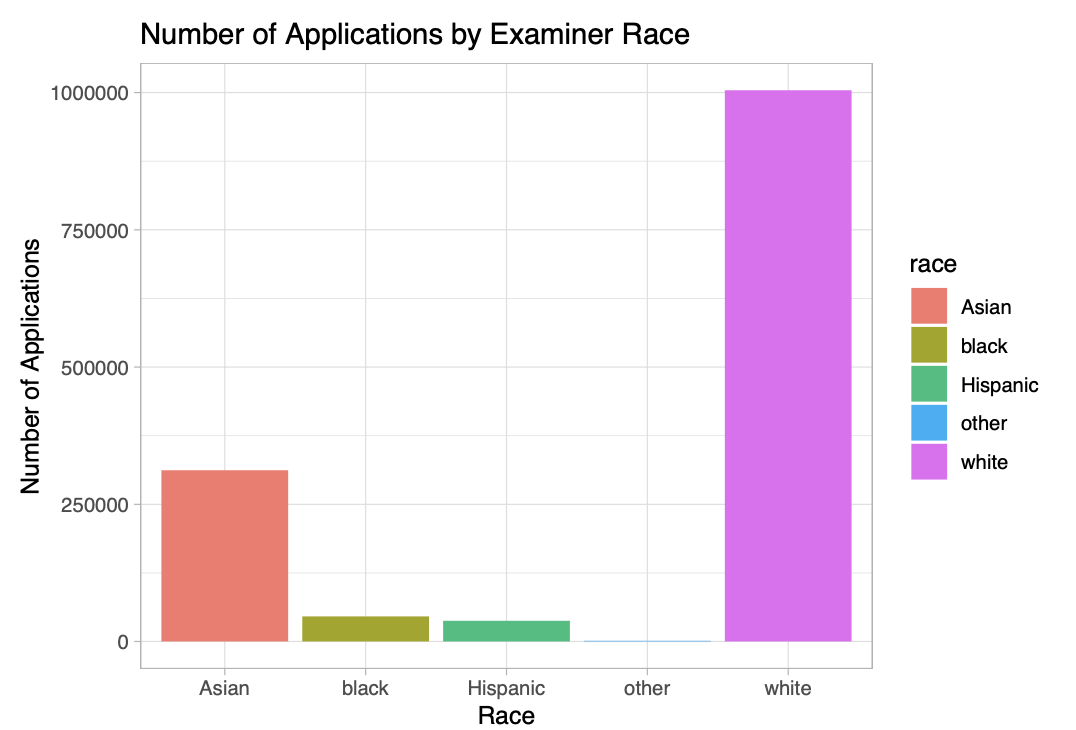


Figure 2.3.3 Number of Applications by Race Figure 2.3.4 Average Number of Applications by Race

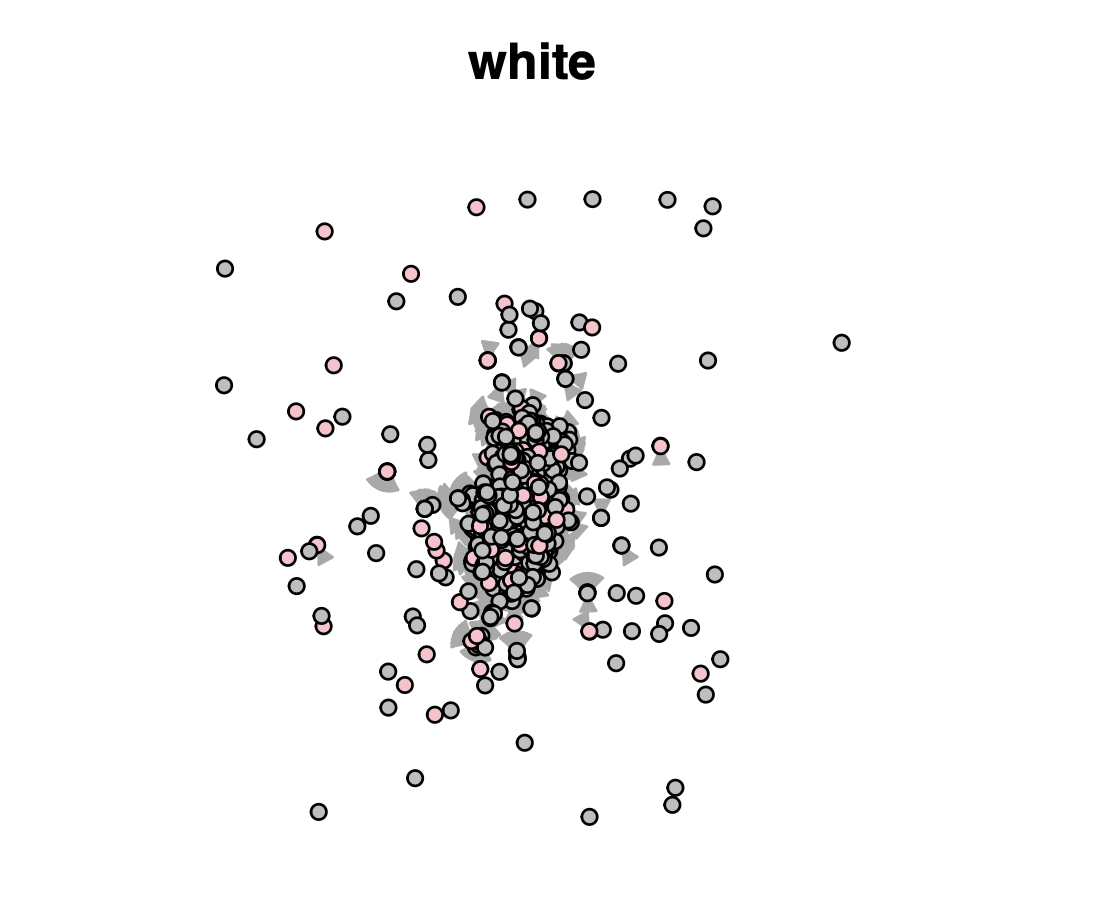
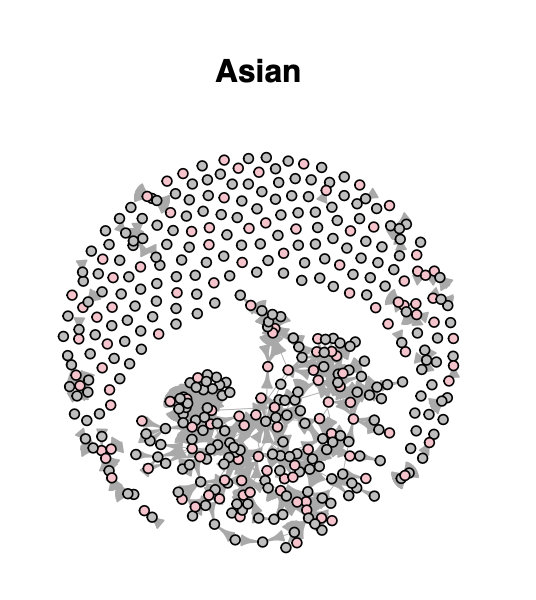
 

Figure 2.4.1 Network Graph for Race - White Figure 2.4.2 Network Graph for Race - Asian

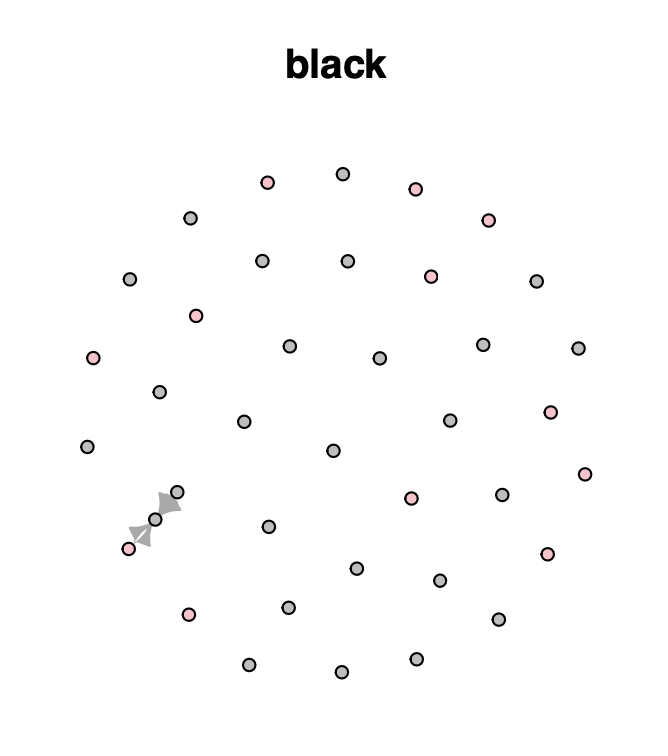
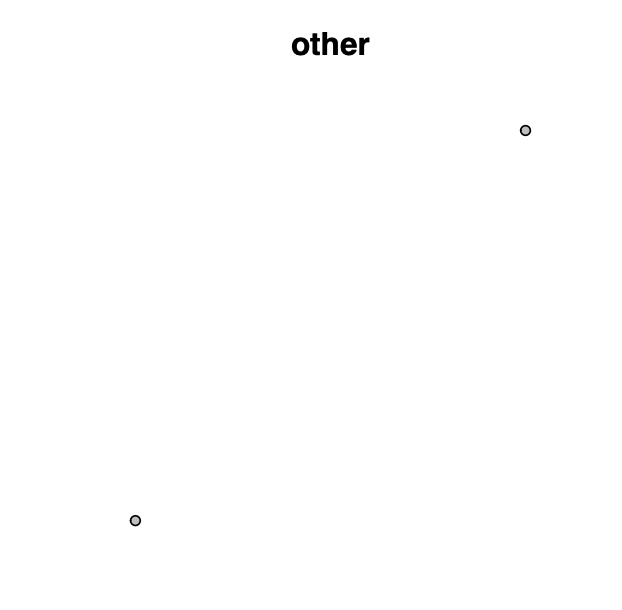
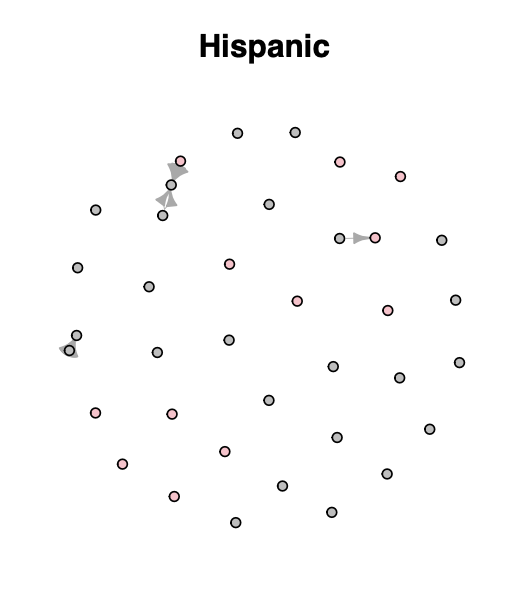
 

Figure 2.4.3 Network Graph for Race - Black Figure 2.4.4 Network Graph for Race - Hispanic Figure 2.4.5 Network Graph for Race - Other

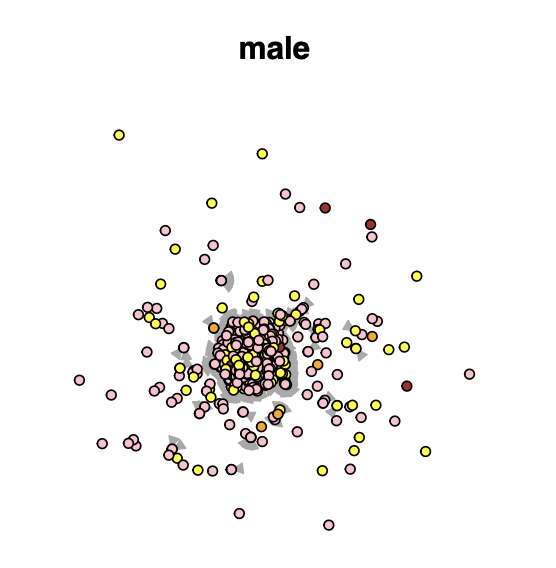
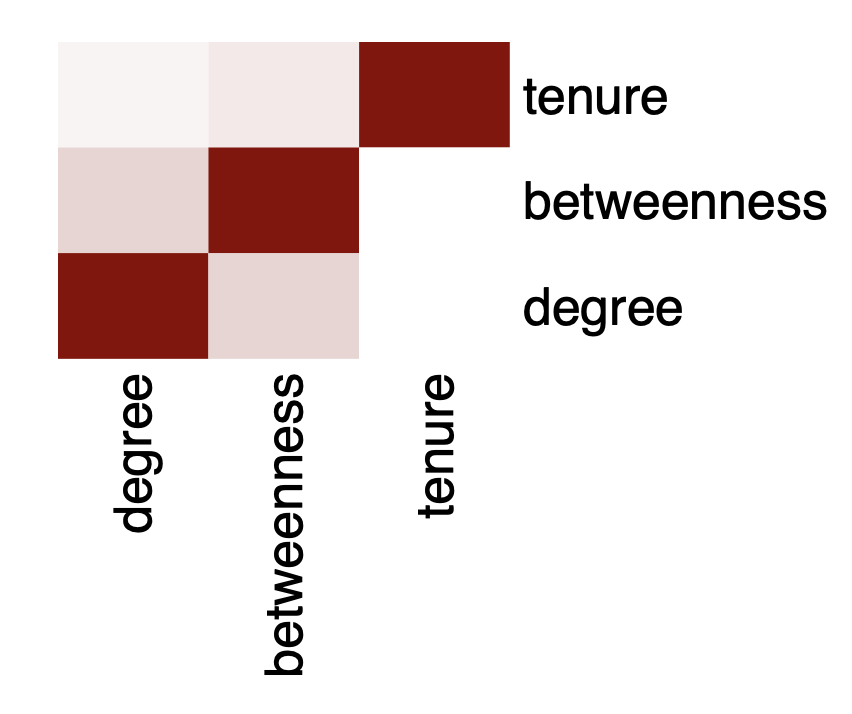
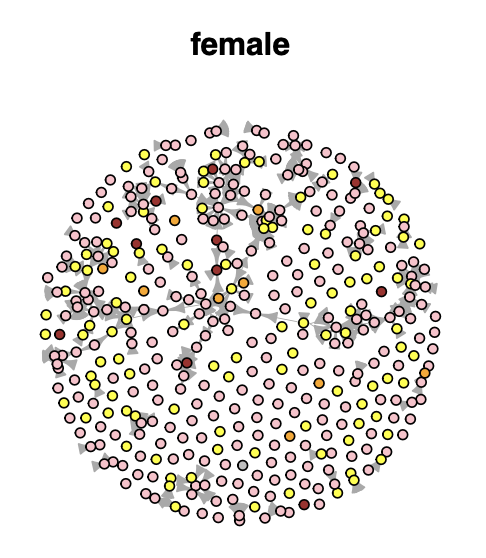
 

Figure 2.5.1 Network Graph for Gender - Male Figure 2.5.2 Network Graph for Gender - Female Figure 2.6.1 Correlation Matrix

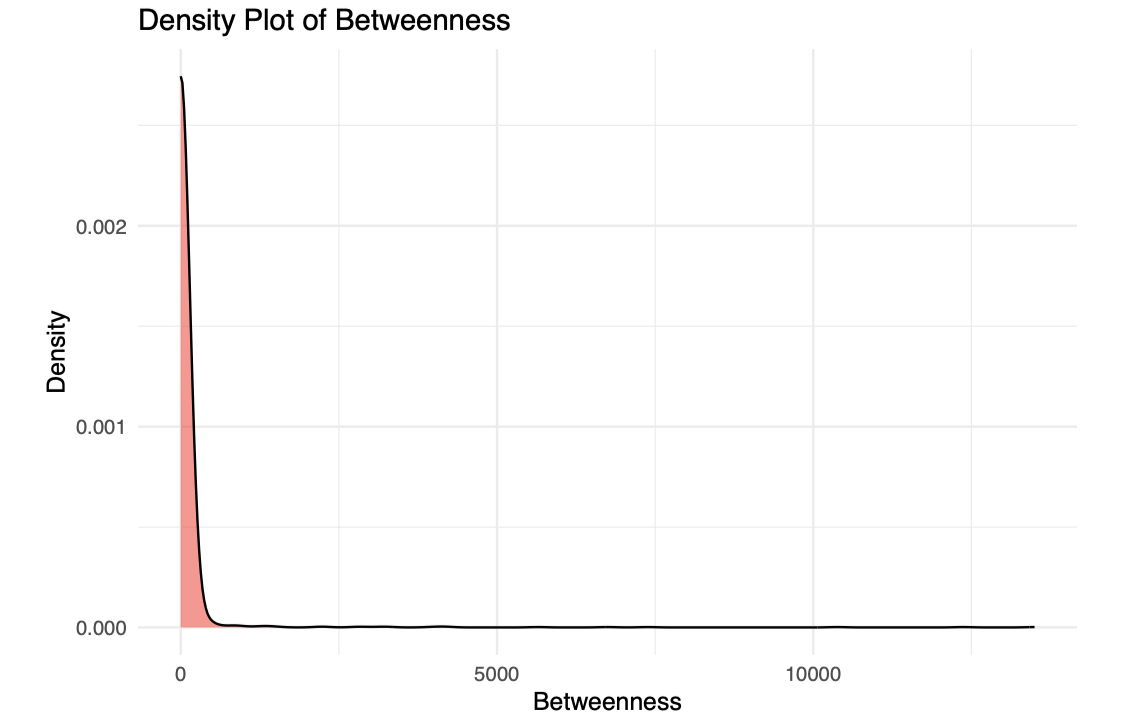
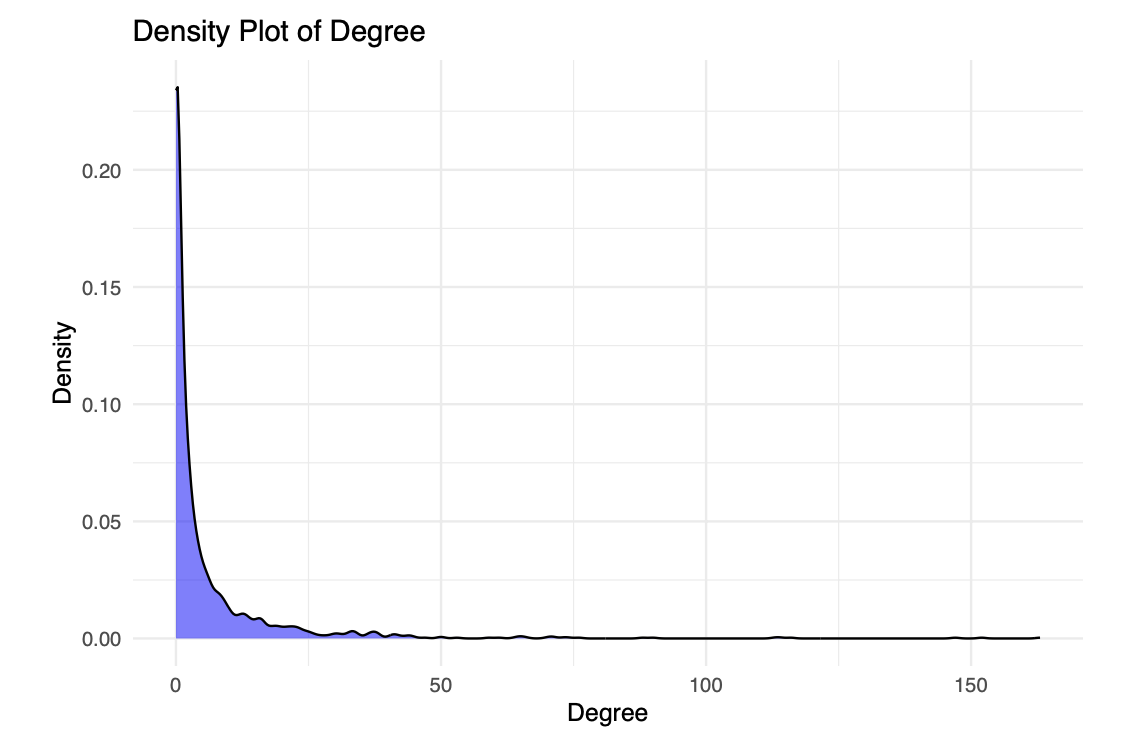


Figure 2.6.2 Density Plot of Degree Figure 2.6.3 Density Plot of Betweenness

| **Variable** | **Estimate** | **Std. Error** | **t-value** | **P-value** |
| --- | --- | --- | --- | --- |
| (Intercept) | 1009.5562 | 1.6959 | 595.291 | <2e-16 |
| log\_degree | 25.40401 | 1.59915 | 15.886 | <2e-16 |
| log\_betweenness | 12.72942 | 1.46346 | 8.698 | <2e-16 |
| gendermale | -6.3593 | 1.21307 | -5.242 | 1.59E-07 |
| raceblack | 11.23396 | 3.85076 | 2.917 | 0.00353 |
| raceHispanic | 1.13986 | 4.26892 | 0.267 | 0.789459 |
| raceother | 134.31215 | 13.94044 | 9.635 | <2e-16 |
| racewhite | -12.72897 | 1.32165 | -9.631 | <2e-16 |
| tenure\_days | -8.35162 | 1.34475 | -6.211 | 5.28E-10 |
| tc1700 | 22.9081 | 1.42356 | 16.092 | <2e-16 |
| tc2100 | 169.04821 | 1.71483 | 98.58 | <2e-16 |
| tc2400 | 95.73142 | 1.98808 | 48.153 | <2e-16 |
| log\_degree:gendermale | -4.96 | 1.33244 | -3.722 | 0.000197 |
| log\_betweenness:gendermale | 2.73935 | 1.34479 | 2.037 | 0.041649 |
| gendermale:tenure\_days | -0.06023 | 1.27766 | -0.047 | 0.962404 |
| log\_degree:raceblack | -11.77472 | 4.62745 | -2.545 | 0.010942 |
| log\_degree:raceHispanic | -25.95658 | 4.21688 | -6.155 | 7.49E-10 |
| log\_degree:raceother | -90.99896 | 14.55167 | -6.254 | 4.02E-10 |
| log\_degree:racewhite | -10.70987 | 1.42758 | -7.502 | 6.29E-14 |
| log\_betweenness:raceblack | 2.14462 | 5.66081 | 0.379 | 0.704796 |
| log\_betweenness:raceHispanic | 21.44965 | 4.89239 | 4.384 | 1.16E-05 |
| log\_betweenness:racewhite | -12.05342 | 1.25049 | -9.639 | <2e-16 |
| raceblack:tenure\_days | -42.02431 | 5.17635 | -8.119 | 4.73E-16 |
| raceHispanic:tenure\_days | 1.16061 | 5.0203 | 0.231 | 0.817173 |
| racewhite:tenure\_days | -14.46995 | 1.21934 | -11.867 | <2e-16 |
| Residual standard error | 410.8 | 603840 |  |  |
| Multiple R-squared | 0.03136 |  |  |  |
| Adjusted R-squared | 0.0313 |  |  |  |

Table 3.1 Linear Regression Model

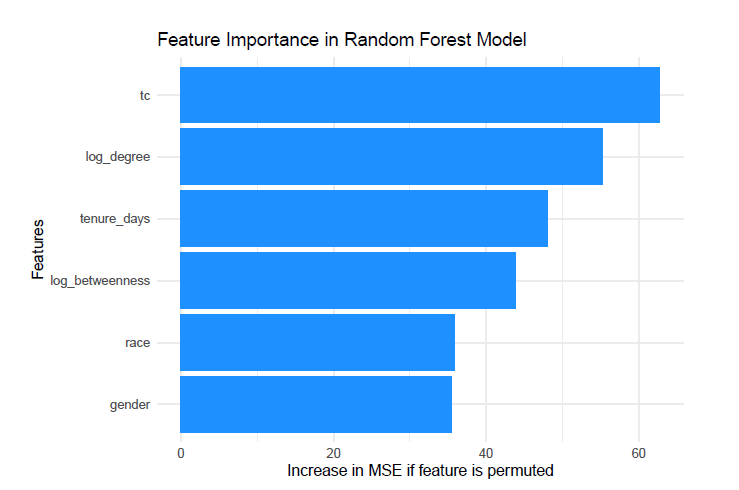


Figure 3.1 Random Forest Feature Importance