USPTO Project Report

Submitted to: Professor Roman Galperin

By: Keel Scruton (260433121), Yulin Hong (260898713), Carlos Fabbri Garcia (261018821), Justine Nadeau-Routhier (260483869)

Introduction The United States Patent and Trademark Office ("USPTO") is the federal agency for granting U.S. patents and registering trademarks. The USPTO's mandate is to "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." [1], as per the U.S. Constitution (Article I, Section 8, Clause 3). Hence, the USPTO promotes the nation's technological progress and achievement through the protection of new ideas and investments in innovation and creativity. To achieve that, the USPTO employs more than 10,000 patent examiners. The USPTO faces challenges arising from long review times for some patent applications, which makes the examination time within inventors inconsistent. Since there are imbalances in the timing of the USPTO's decisions on patent applications, it is becoming a sensitive issue for inventors. To keep up with pressure from policymakers and regulators, the USPTO is considering looking more deeply into the patterns of examiners' work, including differences in network position and application processing time based on gender, race, and workgroup categorizations. The central research questions were to understand the organizational and social factors associated with the length of patent application prosecution, as well as the role of network structure, race, and ethnicity in this process. The goal of this report is to present and discuss the main findings of the analysis of the USPTO's data and to give objective recommendations with respect to the challenges experienced by the USPTO, notably on improving the patent application process and reducing variations in the examination time for patent applications. This report includes a detailed description of the methodology used to conduct the analysis, a demonstration of the analysis and results, a discussion of the results, and conclusions and recommendations drawn from them.

Methodology By following a series of steps beginning with selecting the data and the characteristic variables, creating advice networks, carrying out calculations of centrality score and application processing time, our consulting team was able to develop a linear regression model and to provide recommendations to the USPTO for improving the patent application process. The linear regression model was developed to estimate the relationship between network centrality and the application processing time, while also taking examiner demographic attributes into account. These steps are further elaborated in the following sections. Selection of data For the purpose of this analysis, two workgroups were chosen taking into consideration the fact that in order to compare the effects of examiners' demographics and other applications data on the patent processing time, the workgroups should have some baseline similarities in terms of occupation and type of work involved. The two workgroups of examiners selected come from the same technology centre: the workgroup 161 is focusing on organic compounds patents, whereas the workgroup 162 is focusing on organic chemistry patents. Because of this, we expect that there will be advice seeking between both groups on cross departmental patent issues. In fact, due to the similar nature of the work, any differences in demographic that appear in the workgroups could be potentially used as explanation for why one group may be able to process applications faster than another. Selection of characteristic variables A close look at Figure 1 clearly shows that the workgroup 161 has more male than female examiners, while the opposite trend is observed for the workgroup 162, in which the gender diversity is greater and with the greatest difference from the average demographic of the entire USPTO examiners. From the histograms, we can observe that: 36% of examiners in the workgroup 161 are females and 49% of examiners in the workgroup 161 are males; 42% of examiners in the workgroup 162 are females and 36% of examiners in the workgroup 162 are males; and in average, 28% of examiners are females and 57% of examiners are males.

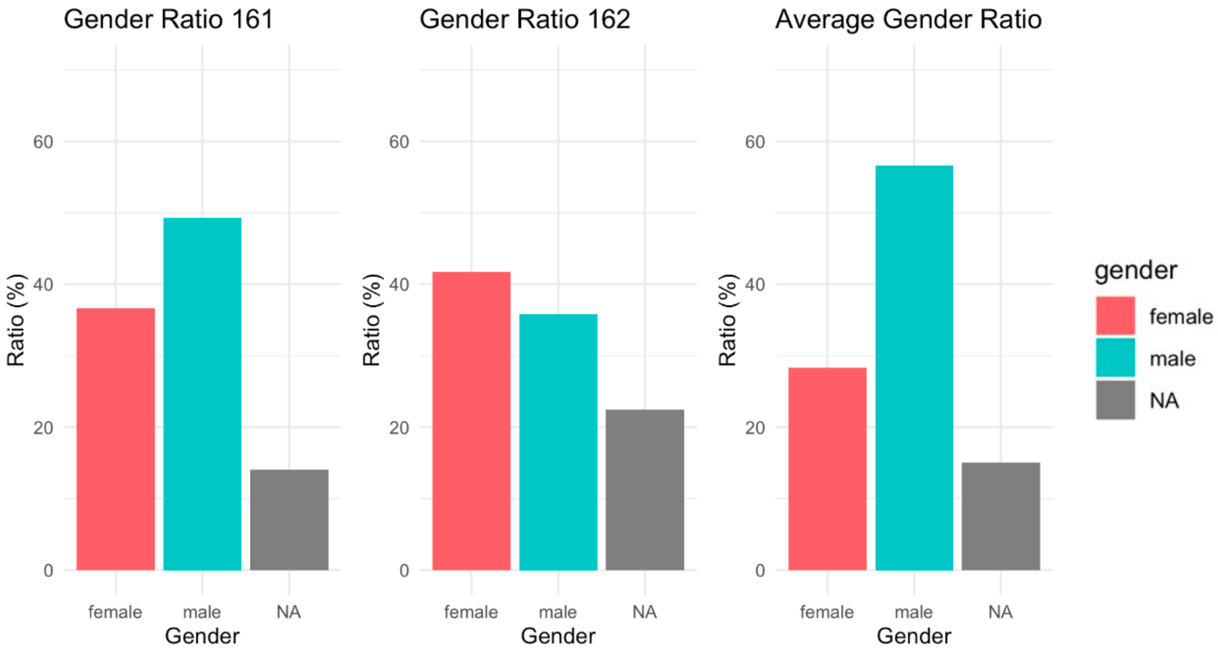
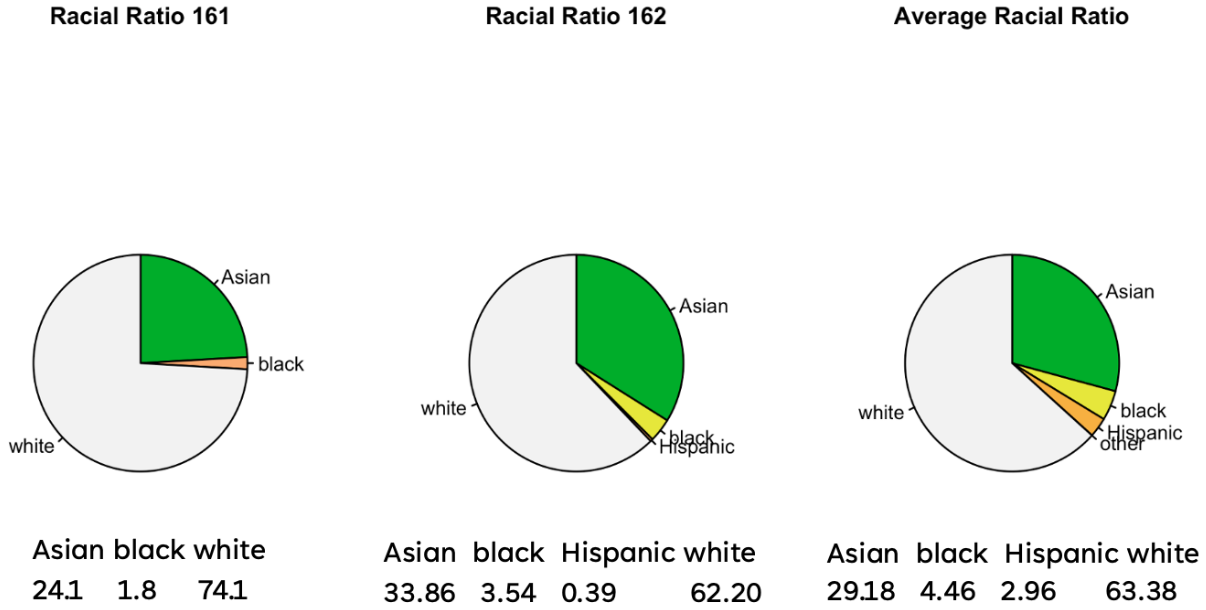


Figure 1: Gender Ratios of Selected Groups

Continuing the analysis of the workgroups 161 and 162, this time observing race demographics, it can be observed that both groups are very similar from a race perspective and do not differ much from the average of the USPTO. White and Asian examiners are the most common group, while black and Hispanic examiners make up a smaller minority. Pie charts containing results associated with each workgroup can be observed in Figure 2, along with the percentages allocated to each race.



Asian	black	white
24.1	1.8	74.1

Asian	black	Hispanic	white
33.86	3.54	0.39	62.20

Asian	black	Hispanic	white
29.18	4.46	2.96	63.38

Figure 2: Race Ratios of Selected Groups

Creation of advice networks In order to understand the interconnected advice relationships between examiners we have created graphs that allow us to investigate the network structure taking into account both the direction of advice, as shows by the arrow's direction in the graph, and degree centrality which has been represented by the size of the nodes. From Figure 3, we can see that those with larger nodes are mostly those examiners who are asking a lot of questions. That is the arrows are facing away from large nodes towards others. Furthermore, the examiners that the arrows are pointing towards tend to be smaller, i.e. they do not tend to have many connections, and there tend to be many of them for each advice seeker. From these observations, we can conclude that examiners are seeking highly specialized information that requires visiting different experts for each of their applications. We can also see that there is a fair mix of advice seeking within each group with questions also going back and forth between the two groups. This is because the topics covered by the two working groups are very similar, both groups belonging to the same technology center. We would expect some patents to potentially overlap the two topics and require expert opinion from both. By observing individual nodes, we can see that there are examples of examiners from one group seeking advice from an examiner in their own group, and that examiner in turn seeking advice from the other group. This could be representing instances of informational brokerage with certain examiners acting as the broker between the two groups.

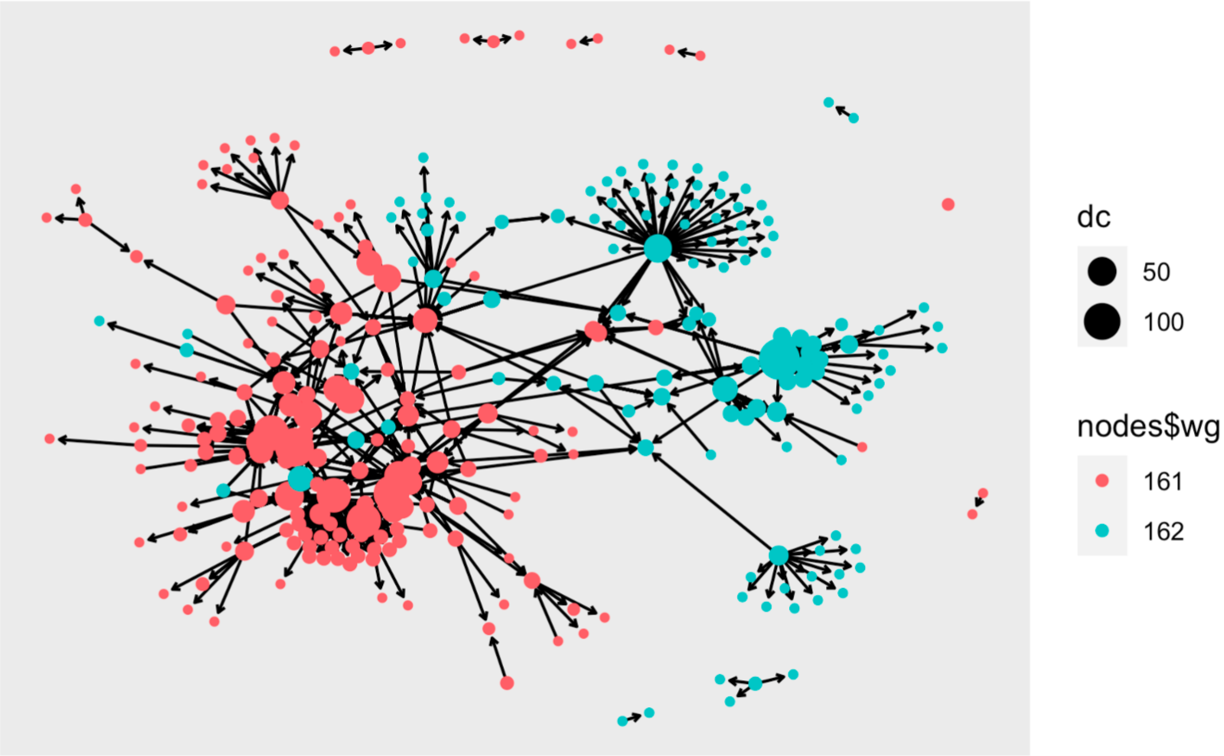


Figure 3: Directed Advice Network Graph with Node Size Showing Degree Centrality

When re-graphing the advice network taking into consideration betweenness centrality represented by the node size, we see on Figure 4 that the number of larger nodes decreases. We can also observe that the examiners who tend to ask for a lot of advice have very low betweenness centrality scores. This is because those with high betweenness centrality are acting as informational brokers who promote informational flow throughout the entire group. The issue that we notice here is that there is a very small number of examiners who are acting as brokers helping the flow of information for both group and the combined network of both workgroups. It is now possible to see that those examiners identified in the previous graph as having a high degree centrality and lots of connections, but only seeking advice do not contribute to efficient informational flow.

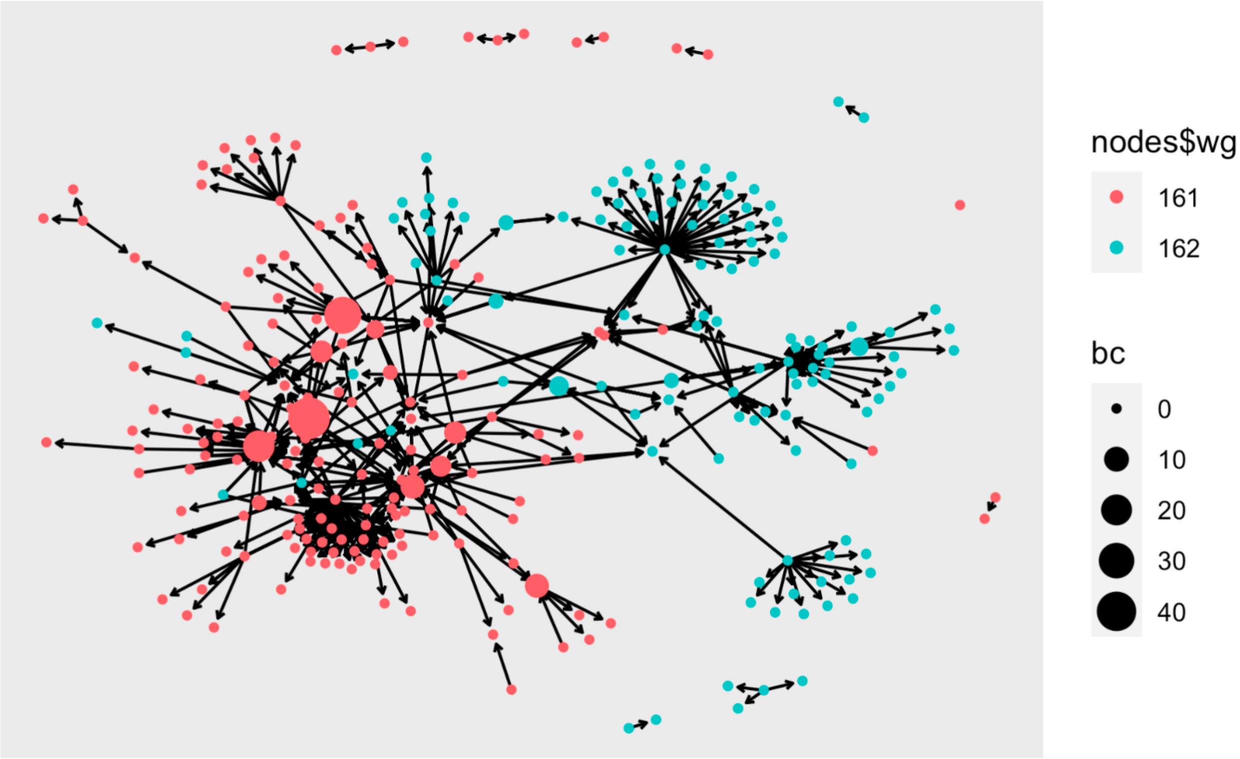


Figure 4: Directed Advice Network Graph with Node Size Showing Betweenness Centrality

Development of a Linear Regression Model To study the organizational and social factors associated with the length of the patent application processes, we fit a linear regression including sociodemographic variables and network derived variables. Once the model was trained, we were able to print a report of the coefficients to perform a careful evaluation of the relationship with each factor and the application processing time, attempting to offer a reasonable interpretation for each of the significant variables detected. Next, we will proceed to discuss in detail the steps that lead up to the creation of the model. First, we maintained the same filters that we applied for the network analysis part of the work, meaning that we focused on analyzing the characteristics of the workgroups 161 and 162 for reasons that have been explained previously. When filtering applications for these two workgroups, we obtain 204,832 distinct applications, and a network composed of 283 different connected vertices by 1116 edges.

For each row in the application data frame, we calculate the application processing time, which is the difference between the date of initial request and that day of acceptance or refusal of the patent. We filter our data to keep only the rows that have information for one of these two dates, and we create the new column for application processing time.

We then use our network graph object to call the igraph methods that efficiently compute the centrality scores we intend to measure. In this case, as has been shown in preceding sections, we will explore four characteristics from each node: degree centrality, betweenness centrality, eigenvector centrality, and closeness centrality. Because closeness centrality turned out to hold null values for most of the nodes analyzed, we had to discard it from our variable selection for the model.

Afterwards, we joined the information we had gathered from each of the 283 nodes (each one representing a different examiner) back to the applications table. This left us with a final dataset size of 62,787 rows of application information to fit in the linear model. The designed model was a multivariate linear regression with the following formula:

App Proc Time ~ f(

- Gender
- Race
- Tenure Days
- Work groups

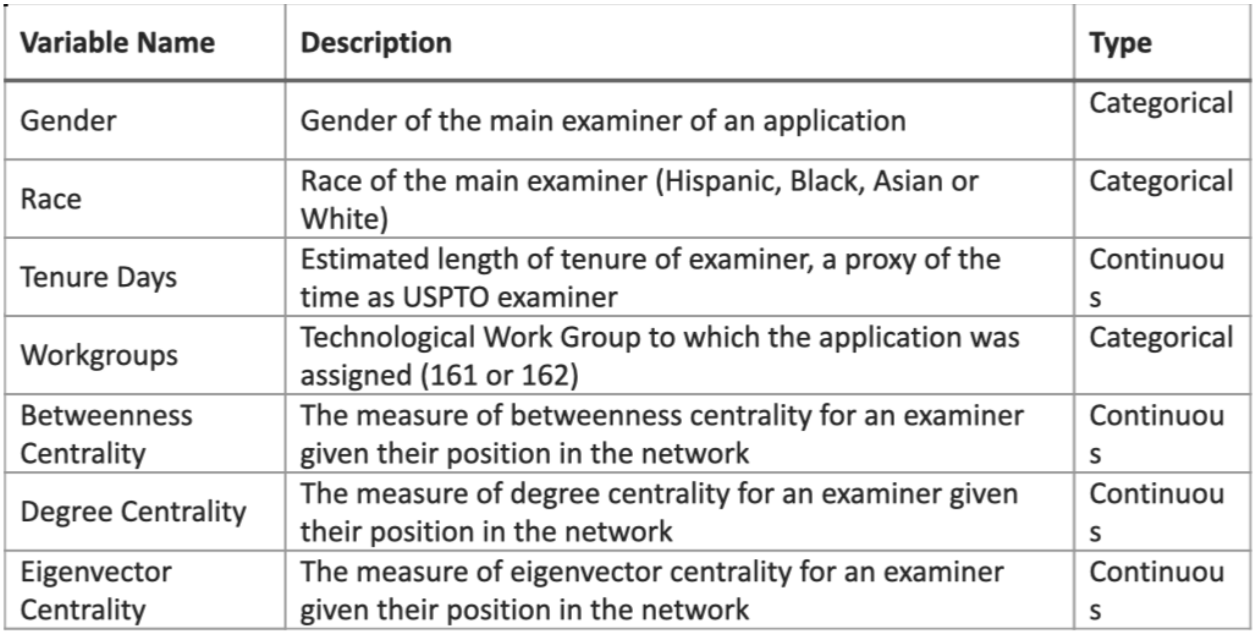
+

- Betweenness Centrality
- Degree Centrality
- Eigenvector Centrality

)

Figure 5: Linear Regression Model to Estimate the Relationship Between Variables and Application Processing Time

In total, there were 7 variables that we tested for significance in the model, as show in the table below:



Variable Name	Description	Type
Gender	Gender of the main examiner of an application	Categorical
Race	Race of the main examiner (Hispanic, Black, Asian or White)	Categorical
Tenure Days	Estimated length of tenure of examiner, a proxy of the time as USPTO examiner	Continuous
Workgroups	Technological Work Group to which the application was assigned (161 or 162)	Categorical
Betweenness Centrality	The measure of betweenness centrality for an examiner given their position in the network	Continuous
Degree Centrality	The measure of degree centrality for an examiner given their position in the network	Continuous
Eigenvector Centrality	The measure of eigenvector centrality for an examiner given their position in the network	Continuous

Figure 1: table 1:

Table 1: Independent Variables Tested for Significance on the Model

The results obtained from fitting the linear model are reported in section 4, Results. Some additional graphs were plotted to verify the relationship between independent and target variables and solidify the interpretations derived from the coefficients' magnitude and value.

Results Moving on to the modeling section of the project, we begin by presenting the target variable. The distribution of the target variable, application processing time, was plotted as a histogram to understand the trends of said metric.

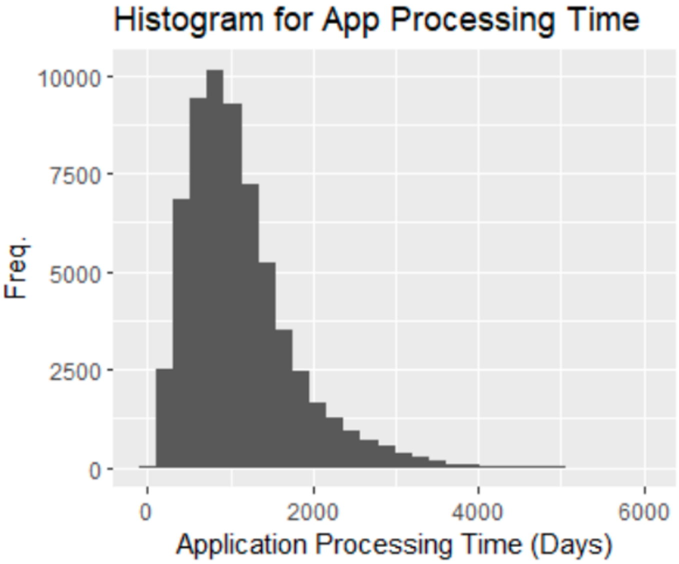
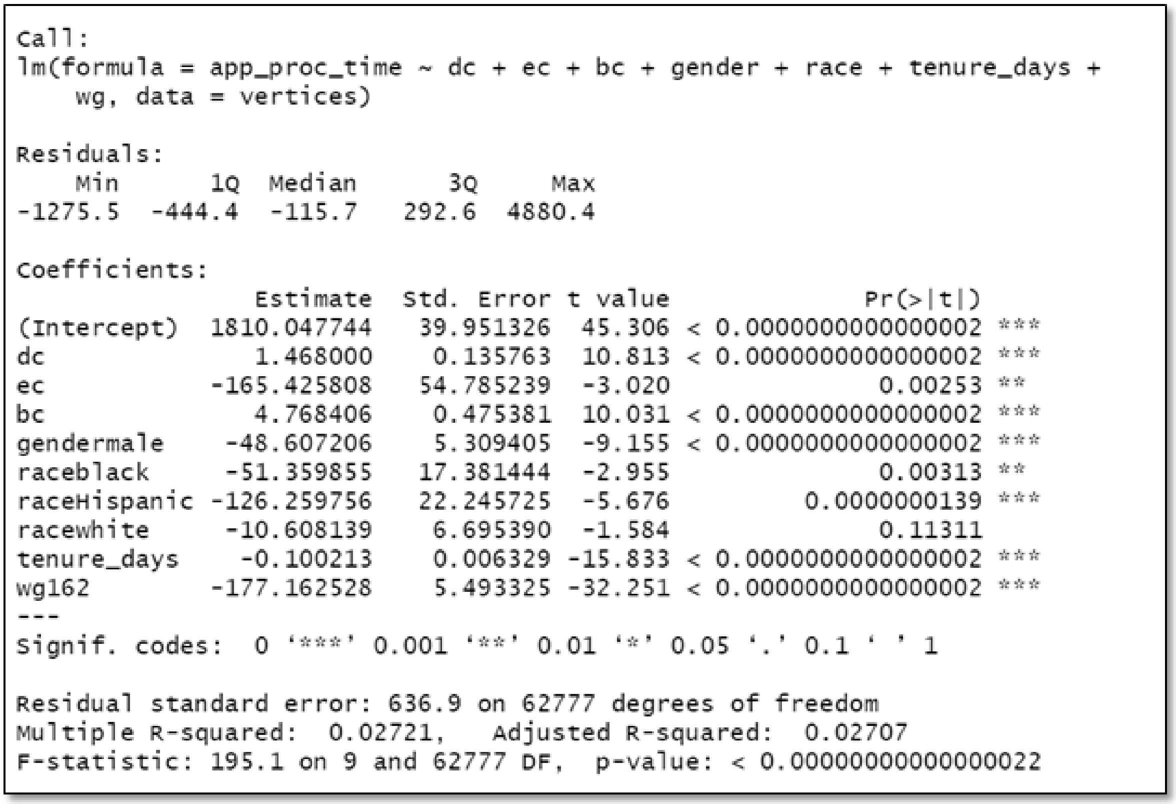


Figure 6: Histogram for Application Processing Time (in days)

The distribution of application processing times for the files belonging to workgroups 161 and 162 is shown in the plot to the left. The distribution peaks at 1,000 days of processing time and afterwards becomes skewed to the right, reaching some outliers of 4,000 or more days of processing time

We begin our reporting of model results by warning that the R-squared in our model landed in the range considered a fairly weak model. This means that the implication from our model could be inconsistent with a new sample or with the statistical universe. Still, the model reported some significant relationships between variables with a 95% confidence, thus we will present the total summary of our model.



```
Call:
lm(formula = app_proc_time ~ dc + ec + bc + gender + race + tenure_days +
    wg, data = vertices)

Residuals:
    Min      1Q  Median      3Q     Max
-1275.5  -444.4  -115.7   292.6  4880.4

Coefficients:
                Estimate  Std. Error t value           Pr(>|t|)
(Intercept)  1810.047744   39.951326  45.306 < 0.0000000000000002 ***
dc              1.468000    0.135763  10.813 < 0.0000000000000002 ***
ec           -165.425808   54.785239  -3.020            0.00253 **
bc              4.768406    0.475381  10.031 < 0.0000000000000002 ***
gendermale    -48.607206    5.309405  -9.155 < 0.0000000000000002 ***
raceblack     -51.359855   17.381444  -2.955            0.00313 **
raceHispanic -126.259756   22.245725  -5.676       0.0000000139 ***
racewhite     -10.608139    6.695390  -1.584            0.11311
tenure_days    -0.100213    0.006329 -15.833 < 0.0000000000000002 ***
wg162        -177.162528    5.493325 -32.251 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 636.9 on 62777 degrees of freedom
Multiple R-squared:  0.02721,   Adjusted R-squared:  0.02707
F-statistic: 195.1 on 9 and 62777 DF,  p-value: < 0.00000000000000022
```

Figure 7: R Generated Report of the Linear Model Fitted for Workgroups 161 and 162 Combined

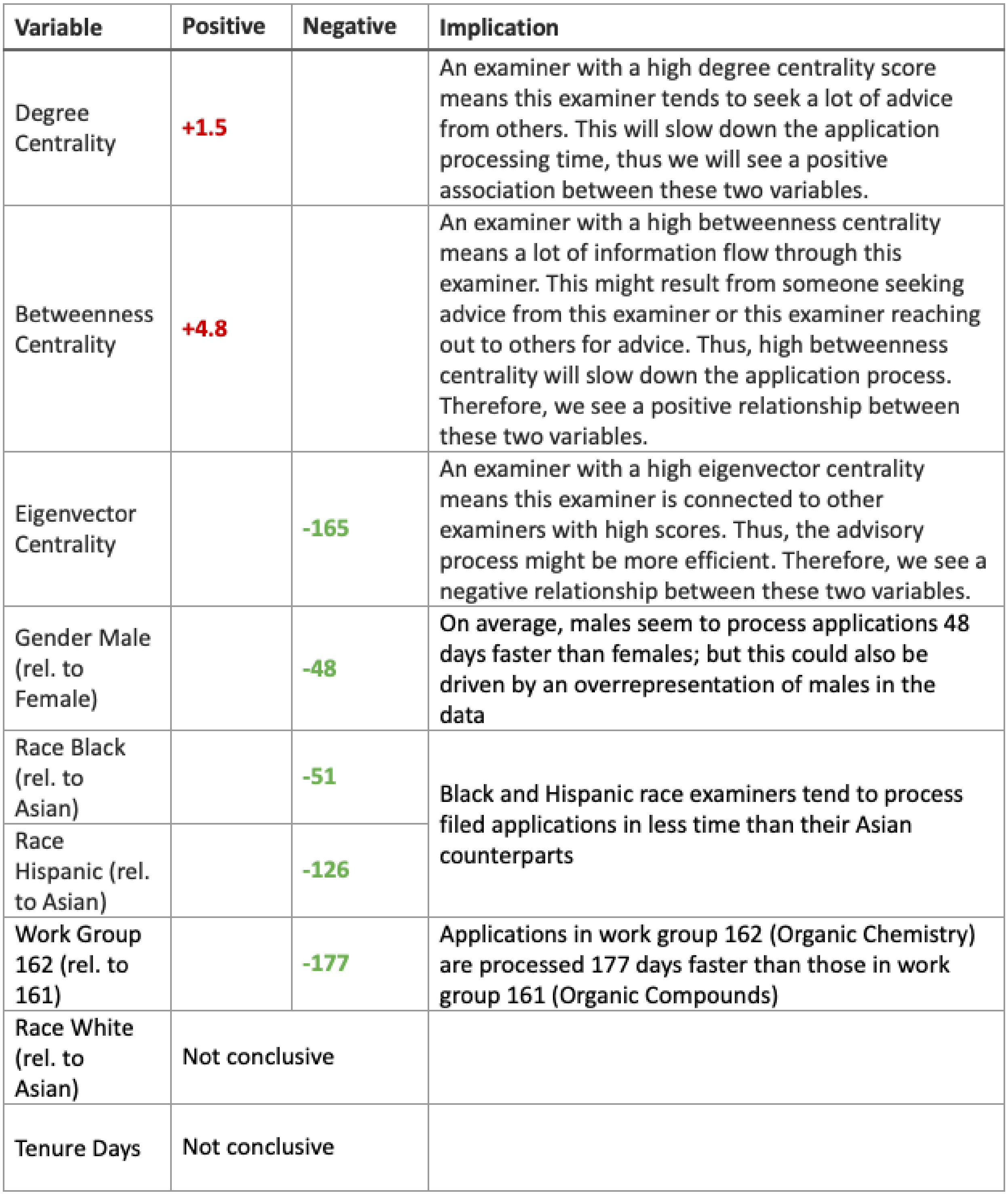
As we can see from Figure 7, the R-squared of the trained model is of 0.02 which is considered very low. From the eight coefficients computed (seven variables, plus the intercept), seven came out to be significant when predicting the value of the processing time for a given application. The coefficient values are reported in the column “Estimate” of Figure 7, where one can detect that certain variables have a direct relationship with the application processing times (coefficient is positive) and others have an indirect relationship (coefficient is negative). In the Discussion section, we will contextualize the numbers reported and discuss the implications they can have for the USPTO.

Discussion The data contributes a clearer understanding of the network structure and dynamics between examiners, even though the methodological choices were constrained by the selection of a linear regression model. The following Table 2 summarizes the regression coefficients for each variable in the model and their implications. It appears that most links and high centrality clusters are formed by people asking for advice.

Table 2: Implication of the Variables Regression Coefficients for Assessing their Effects on the Model

Based on the statistical summary, we have discovered some interesting findings. Before we dive into the findings, we would like to state the definition of different centrality score measurement method. Degree centrality assigns an importance score based simply on the number of links held by each node [2]. Betweenness centrality measures the number of times a node lies on the shortest path between other nodes [2]. Like degree centrality, Eigenvector Centrality measures a node’s influence based on the number of links it has to other nodes in the network [2]. Eigenvector Centrality then goes a step further by also considering how well connected a node is, and how many links their connections have, and so on through the network. We then interpret the linear regression statistic summary results based on these definitions.

We have found that eigenvector centrality is negatively associated with application processing time, while



Variable	Positive	Negative	Implication
Degree Centrality	+1.5		An examiner with a high degree centrality score means this examiner tends to seek a lot of advice from others. This will slow down the application processing time, thus we will see a positive association between these two variables.
Betweenness Centrality	+4.8		An examiner with a high betweenness centrality means a lot of information flow through this examiner. This might result from someone seeking advice from this examiner or this examiner reaching out to others for advice. Thus, high betweenness centrality will slow down the application process. Therefore, we see a positive relationship between these two variables.
Eigenvector Centrality		-165	An examiner with a high eigenvector centrality means this examiner is connected to other examiners with high scores. Thus, the advisory process might be more efficient. Therefore, we see a negative relationship between these two variables.
Gender Male (rel. to Female)		-48	On average, males seem to process applications 48 days faster than females; but this could also be driven by an overrepresentation of males in the data
Race Black (rel. to Asian)		-51	Black and Hispanic race examiners tend to process filed applications in less time than their Asian counterparts
Race Hispanic (rel. to Asian)		-126	
Work Group 162 (rel. to 161)		-177	Applications in work group 162 (Organic Chemistry) are processed 177 days faster than those in work group 161 (Organic Compounds)
Race White (rel. to Asian)	Not conclusive		
Tenure Days	Not conclusive		

Figure 2: Table 2

degree centrality and betweenness centrality are positively associated with application processing time. Degree centrality is easy and straightforward to interpret. Examiners with a high degree of centrality are connected to a lot of people in the advisory network. Thus, this means that they took a lot of time consulting others or answering others' questions. This will slow the application process down. Examiners with high betweenness centrality have a great impact on the information flow in the advisory system, which is like a marked brokerage position in the network. They would spend a lot of time connecting separate groups together and helping the information flow, hence they do not have enough time to work on their own application. Lastly, the examiners with high eigenvector centrality scores are connected to other examiners with high scores. They will be able to access the information and knowledge that their high-score neighbors have. This will make the advisory process efficient and fasten the application process. As for demographic variables, we have found that on average, male examiners seem to process the application faster than female examiners. However, because there are a lot more male examiners than female examiners in the dataset, this finding might be driven by an overrepresentation of male examiners. The black and Hispanic examiners, on average, tend to process the application faster than the Asian examiners. And the interaction between white and Asian examiners is not statistically significant form our analysis. For workgroup, the organic chemistry workgroup (162) processed the application 177 days faster than those in the organic compounds workgroup (161). The interaction between tenure days and application processing times is statistically insignificant.

Recommendations Based on the analysis mentioned above, we came up with three suggestions for the USPTO on how to make their patent application process more efficient.
Firstly, the USPTO should promoting the creation of broker agents in the network. Through initiatives activities such as buddy programs, the organization could help alleviate the current situation of having only a few examiners with high brokerage of information. Overall, this would help information flow faster among teams and stop high betweenness centrality from being an issue for certain key examiners. Secondly, promoting more diverse groups (in terms of race and gender) is a safe bet for the USPTO, since we have noticed that more diverse workgroups tend to have faster processing times. Also, a study [3] has shown that companies in the top quartile for ethnic and racial diversity in management were 35% more likely to have financial returns above their industry mean, and those in the top quartile for gender diversity were 15% more likely to have returns above the industry mean. Thus, promoting a more diverse workplace should help the USPTO built a more dynamic and efficient workplace. Thirdly and lastly, the organisation should focus on undermining social bubbles in order to diminish polarization of workgroups and on sharing examiner's knowledge, skills, and experience on a large scale. We believe that this could be done through the effects of social contagion, for example with increased social cooperation, broader diversity, and modified spatial structure. While the characteristics of examiners play a role in influence dynamics, diversity can amplify the likelihood of behaviour change.

References [1] United States Patent and Trademark Office (2022). Available at: https://www.uspto.gov/about-us [Accessed 1 June 2022]. [2] Cambridge Intelligence (2020). Social network analysis 101: centrality measures explained. Available at: https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/ [Accessed 1 June 2022]. [3] McKinsey. Diversity Matters (2015) Available at: https://www.mckinsey.com/~/media/mckinsey/business%20functions/people%20and%20orgorganizatio%20performance/our%20insights/why%20diversity%20matters/diversity%20matters.pdf [Accessed 1 June 2022].