Yelp Reviews: Analysis of the Impact of Yelp 'Influencers' on the Yelp User Community.

1. Introduction

The aim of our project was to decipher social interactions between Yelp users. In the first part of our analysis we have quantified the impact of influencers, or 'elite users' as per Yelp. We assessed the importance of the influencer's nodes in the interaction graphs. We then try to quantify the factors that make a user an 'influencer'.

2. The Dataset

The dataset consisted of over 1.2 million business, 6,685,900 reviews and 1,637,138 users. The JSON data was split into 3 main files - business, review and user information files. We included only those users who had left at least one review. The dataset majorly consists of restaurant reviews.

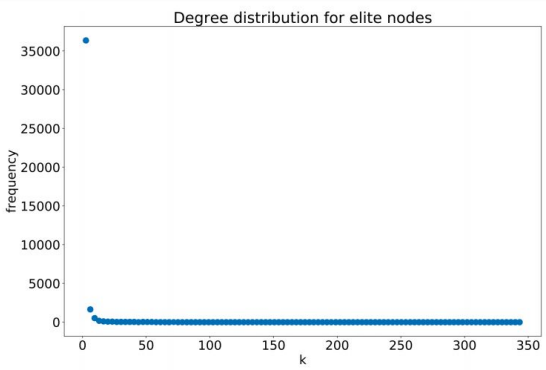
.

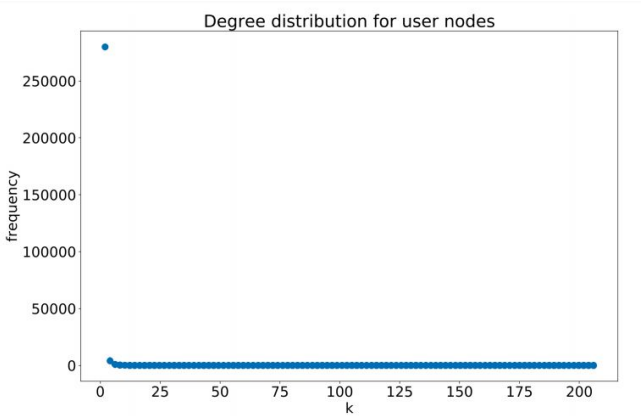
3. Network of Yelp Users

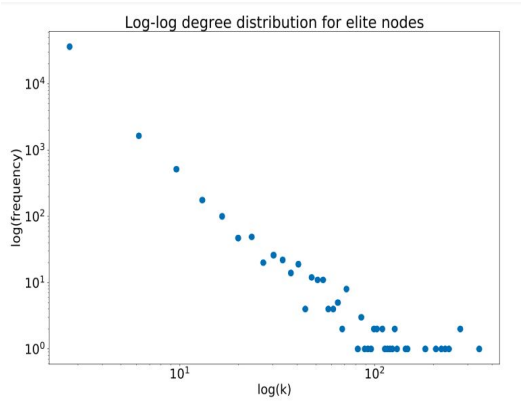
Yelp users are connected in various ways, it is a social network similar to Facebook. We constructed graphs from the datasets that model their relationships. Users represent nodes and are connected to each other if they have reviewed the same business or are each others friends.
Total Number of Nodes: 324751 Regular User Nodes: 285663
Total Number of Edges: 225475 Elite User Nodes: 39088

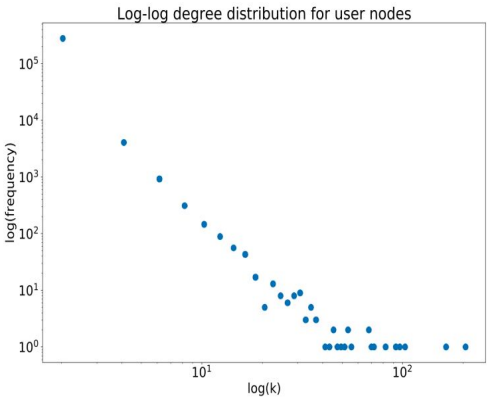
a. Degree Distribution:

The degree distribution for different types of users clearly illustrates that most users have very few friends. Influencers generally tend to have more friends as compared to regular users.



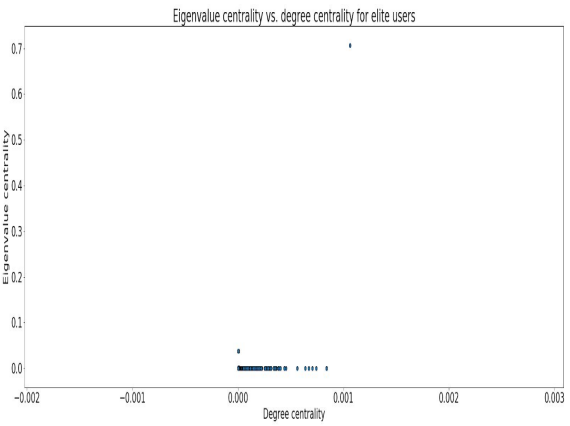


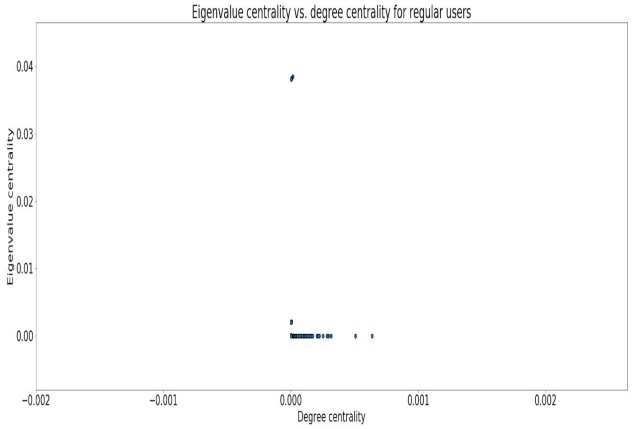




b. Centrality Measures:

We used degree centrality and eigenvalue centrality to determine the importance and role of nodes in the network. The graph shows a correlation between the number of friends and number of popular friends for nodes.





c. Robustness Analysis:

We conducted an analysis of the impact of influencer nodes on the graphs by recursively removing nodes from the graph and computing the largest connected sub components for the graphs. We removed 25% of the influencer nodes at a time and computed the largest connected sub components. Our results showed a reduction of around 63% of the original size of the network when the influencer nodes were removed. On the contrary, we saw a reduction in network size of only around 10% on removing 25% of the regular nodes.

Network Size: 15100741

Average Clustering: 0.0002901350223878054

Average **Eigenvector** centrality values:

Elite Users: 3.8758926177562565e-05

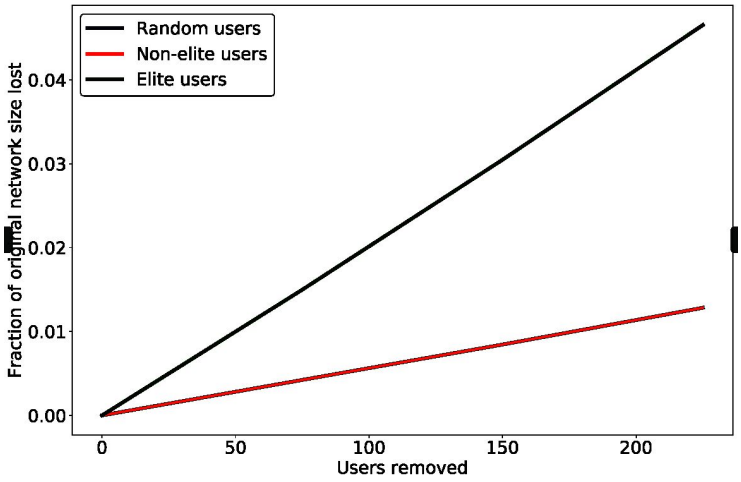
Regular Users: 1.336028861347182e-05

All Users: 2.280645489195958e-05

Average **Degree** Centrality Values

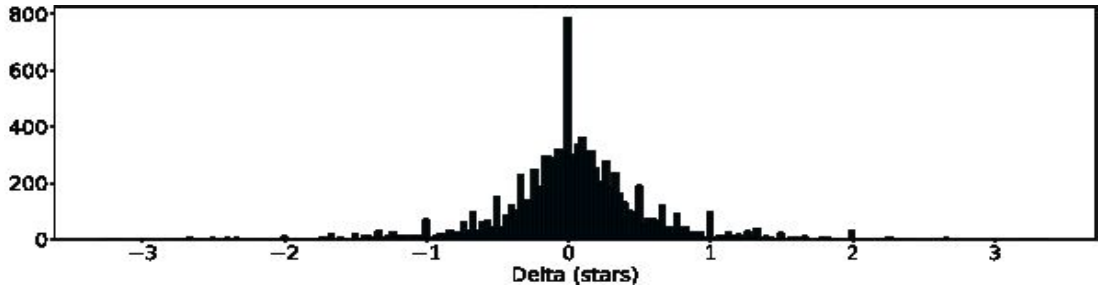
Elite Users: 6.814020167911525e-06

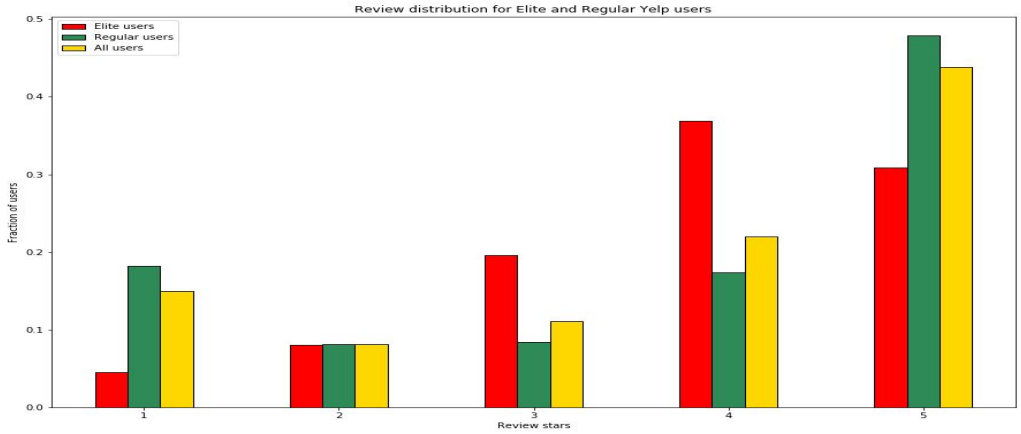
Regular Users: 3.9286158870906085e-06
All Users: 4.275911761556671e-06



d. Differences in Ratings between Influencers and Regular Users.

We discovered that influencers are representative of the reviews for the community of users of which they make up only 31%. This is helpful in making a trend analysis when assessing trends using all users is difficult, we can focus on the influencers. We observed that influencers rated mostly 3 or 4 stars to businesses although they reviewed higher number of businesses. Contrarily, regular users rated more on the extremes i.e. 1's and 5's. We concluded that influencers give more balanced reviews than regular users who tend to leave fewer reviews and review a business mostly when they either love it or hate it enough to bother leaving a review.



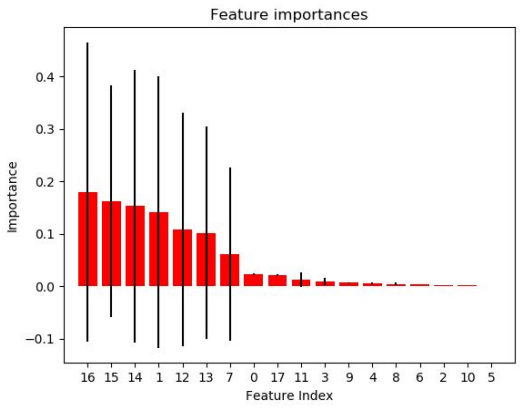


Quantifying a measure for influence:

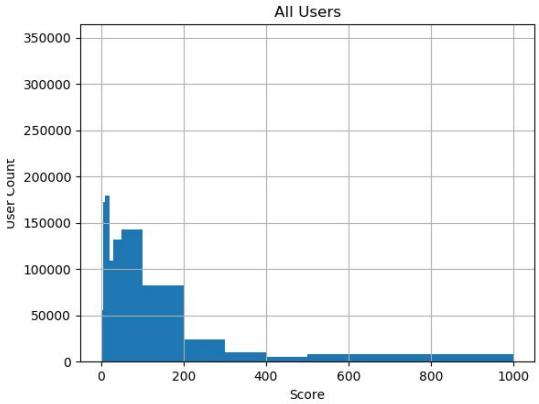
We wanted to see if we can somehow quantify a measure of influence like a score. Since elite users have a great influence on others, we decided to use being elite or not as our target feature. We started by training the randomforest classifier. Since the data set is very imbalanced, we used class weights to train the model properly. During the training we got a predictive score of 97% but since 95 of data is just non-elite users, we used only the elite user data and predicted on it. we got a predictive score of 93% which is not bad and it also tells us that the model is not overfit. Form the model we gathered weights of each feature.

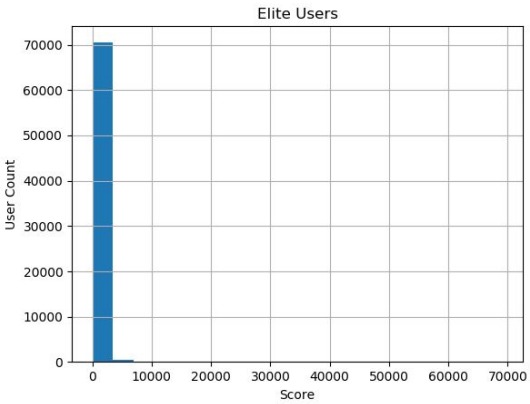
The weights make sense with how the people become influential. We made a score for each user by taking the dot product of the features and their weights.

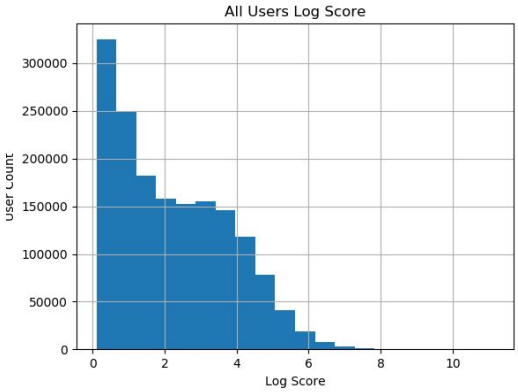
To get a better understanding of the distribution we took log of the scores and plotted it again.We can see in **All User Log Score** plot that its stays in line with the data, as we can see that number influencer is very low in the data. So, as the score increases the count of users decreases. The distribution of elite users is a normal distribution.

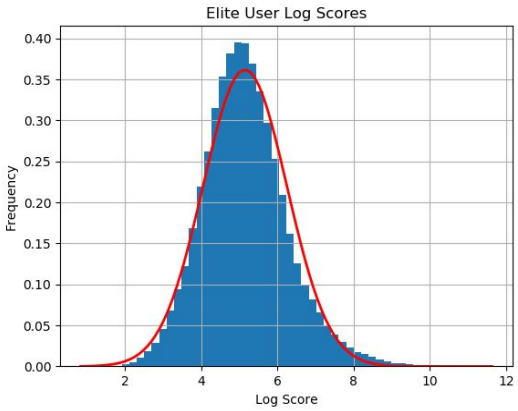



```
useful (0.179282)
review_count (0.162832)
funny (0.153242)
compliment_cool (0.141443)
cool (0.108123)
fans (0.102173)
compliment_note (0.061187)
average_stars (0.023516)
number of Friends (0.020809)
 compliment_writer (0.012704)
 compliment_funny (0.009124)
 compliment_plain (0.006890)
 compliment_hot (0.006187)
 compliment_photos (0.004670)
 compliment_more (0.003711)
 compliment_cute (0.001807)
 compliment_profile (0.001677)
 compliment_list (0.000622)
```









Predicting the probable Elite Users:

With the trained model and the score we derived. We tried to predict user who have more than 80% chance to become elite user and are not already elite users. We conclude that there are 15286 potential users who could become elite users.

4. Conclusions:

We have quantified how users influence each other on the Yelp social network. We determined the impact that influencers have on the network both the social and reviews network. We can conclude that the influencers have a large impact on the community they belong to and are representative of the overall opinion of the community. We also determined features that successfully predict which users influence others. This can be used to determine which users will have an impact on the community and businesses.