Paper prepared for Social Network Analysis Project at TUDublin.

Date: 16th December 2021

Yelp Social Network Analysis of IKEA Reviewers

Brendan Kent

Abstract. Recommendation systems such as Yelp use friendship networks to pass on review information about businesses. These friendship networks also provide a way for the yelp community to grow. This friendship networks contain a lot of information which could be exploited by businesses in order to find out what is important for a user. In this project the friendship network of IKEA reviewers is explored to see if it follows some well known social network analysis principles and whether communities can be detected in the network and furthermore to see if a prediction model can be developed to quantify the chance of a friendship in the network. The results of this project show the IKEA friendship network follows the power law, does not have a “Rich-club” effect, contains multiple different communitites which tie in with the friends location and by completeing ERGM models the most important attribute that helps form friendships is the location(“state”).

1 Introduction

Social networks have been widely utilized by a variety of online review websites. They help propagate reviews along on online network to the people that may like to see them. They usually mange this by having influential users, which many other users follow. The aim of this project is to study the social network of a friendship network of IKEA reviewers using an dataset given by Yelp. Yelp is a geosocial platform that helps people find local businesses, like dentists, restaurants, hair stylists, and many more, it relies on crowd-sourced reviews. In this project the network of reviews giving to IKEA business was analyzed.

On Yelp, links between users are termed as “friends”, though it may be a stretch since many might never met or know each other. The friendship is quite practical since a user review who lives close can be very useful. In many cases, the friendship network are automatically added when user sign-in via larger social network sites such as Facebook.

The Yelp data set(Yelp, 2021) is large in it’s dimensions and the complete collections contains over 10 Gigabytes of information. The data set is grouped into the following object types:

- 8,635,403 reviews
- 160,585 businesses
- 200,000 pictures
- 8 metropolitan areas
- 2,189,457 users
- 1,162,119 tips
- 138,876 check-ins

Because of the size of the dataset, it was important to set a delimitation on the size of the dataset, to be able to filter it so as to make it usable for this project’s work. As mentioned above, this study concentrated only on reviews given to IKEA businesses, narrowing down the scope to 1,505 users spread across 8 IKEA businesses in both the USA and Canada.

This project’s aim is answer a set of research questions as outlined here:

Q1. Does the friendship network follow the power law?

Q2. Does a rich club exist in the friendship network?

Q3. Do friendship networks differ by geographical location in terms of average degree,average shortest path, Transitivity and Assortativity?

Q4. Can the network be split using Modularity-based community detection? what is the Modularity score?

Q5. Can the Girvan–Newman method split the IKEA friendship network into it's geographical parts or communities? Which split is best based on modularity scores?

Q6. Do IKEA users who give poor ratings tend to make friendships with other uses who also give poor ratings?

Q7. Which attribute is the most important for making friendships in the IKEA friendship network?

2 Related work

There exists a large amount of work which have used the Yelp dataset in order to study the networks contained within it.

(Pranata and Susilo, 2016) - Analyzed the Yelp users network for untrustworthy reviews by collecting and comparing the ratings of the most popular users, proving that users should not rely on popular users.

(Amato et al., 2017) - Studied the Yelp social network to find social communities and capture all the relationships that exist between them, as well as performing some ranking of the users. These authors concentrated on the multimedia features of the dataset.

(Cervellini et al., 2017) - This paper analyzed the YELP networks communities and the bridges between them, in particular the characteristics of the bridges and how they influence their friends and power users. They present the term "k-bridge" which is a user who connects to k sub networks. Their finding was that these k-bridges could be used to find the best targets are a marketing campaign.

(Chen et al., 2018) - The authors propose a recommendation algorithm based on density-distance dynamic clustering model using Yelp data. They use bipartite networks of the Yelp users based on clustering communities to build their algorithm.

(Berkani, 2020) - The authors used the Yelp social network to develop a recommendation system to facilitate finding friends for a user based on the similarity between the active user and their friends. This recommendator system also uses the credibility of the user during the evaluation.

(Yu et al., 2018) - Analyzed triangle motifs in order to investigate the gender impact on the Yelp friendship network and found that users prefer to have friends of opposite gender, Additionally it was found that the circle of friends with tolerant users are more closed than the circle of friends using the clustering coefficient.

(Rahimi et al., 2018) - An interesting study where the authors try to distinguish neighbourhoods of similar taste based on Natural Language Processing of the reviews along with other attributes of the businesses, they then find these neighbourhoods to be good indicators of the socioeconomic status of the population of those neighbourhoods.

3 Methodology, data gathering and processing

This project is based on secondary data fetched from Yelp, this empirical evidence is then analyzed using quantitative research methods. All the reasoning comes from an inductive perspective, starting from observations in the user networks and working to form theories.

In order to prepare the data for Social Network Analysis, the first step involved finding all users who have completed reviews for an IKEA stores as shown in Fig 1 below.

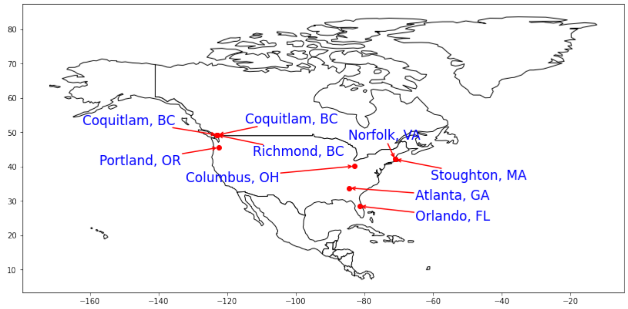


Fig. 1: Map of IKEA Businesses in the Yelp dataset

After this filtering, the user data was then enriched with location information, adding attributes called "state" and "west" to each user node. The attribute "state" is inferred from the rating the user had given, for example if the user rated a IKEA in Georgia, then that user was assigned "GA". The attribute "west" is a simple Boolean to make whether the state is int he east or the west. At this point in the preprocessing the data set contained 3,092 users, a brief overview of them in show in Fig @ref(fig:east _west) below, where it can be seen that most users are from eastern states and MA(Massachusetts) having the highest number. The section "find users who reviewed IKEA businesses" in the accompanying jupyter notebook file (`yelp_data.ipynb`) shows how this was accomplished.

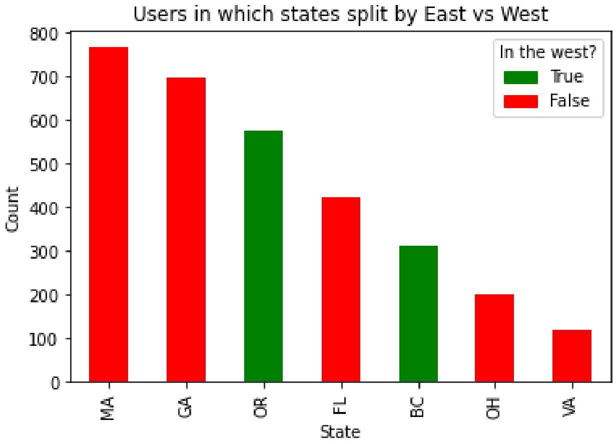


Fig. 2: Break down of users who reviewed IKEA stores

The next step was to set up the friendship network where an edge list was setup by querying each user for friends who had also given IKEA reviews. In this way, each node on my edge list would have been a reviewer of an IKEA store. This step also involved removing users who didn't have any friends. As shown below this removed over half the users in the network, since they would have no edge to connect.

Now the barchart to break down the users and the states looks different as seen below.

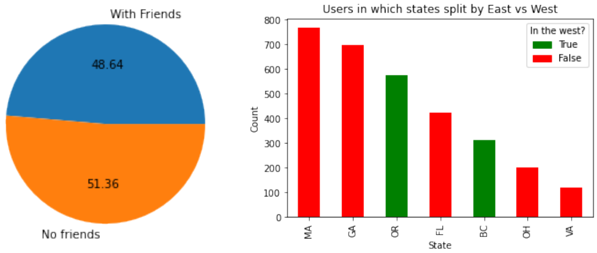


Fig. 3: Break down of users who reviewed IKEA stores who have friends who also reviewed it

At this point everything is ready to read the data into python's networkx(NetworkX, 2021) and R's SNA(Butts, 2020) packages, as well as others such as ERGM (Krivitsky et al., 2021) and Tidyverse (Wickham et al., 2019) to conduct the research questions outlined in the Introduction.

The dataset has be refined down to two files, one which contains the edge list in a CSV format, which shows all the relationships ion the network and the other is a CSV of the users detailing all the different attributes for each user/node/friend.

attribute	value
user_id	N4mIzFm7Qs9yKcRFXg4-1g
name	Scottie
review_count	357
yelping_since	2012-05-25 13:01:18
useful	836
funny	434
cool	198
elite	2013,2014,2015,2016,2017,2018
friends	MuGxIIrcGJ7FnxU7kTPNBw,v0KmYNybWmLXDZdL5BcKug,DxeNa6k3TEekeiF4wM-C9w
fans	30
average_stars	3.3
compliment_hot	4
compliment_more	6
compliment_profile	0
compliment_cute	0
compliment_list	0
compliment_note	20
compliment_plain	24
compliment_cool	7
compliment_funny	7
compliment_writer	6
compliment_photos	2
west	False
state	MA
friends_count	3

Table 1: Example of one friend in the network showing the list of attributes attached

4 Experimentation and Results

4.1 Q1. Does the friendship network follow the power law?

Like many other networks, the IKEA friendship network should show a power law distribution, that is a network whose degree distribution follows a power law. These types of network are known as scale-free networks. In python this is quite easily calculated by sorting the degree measure of each node and fitting those counts of each degree to the power law. In the below figure, a power law Distribution has been created, with the fitted parameter alpha(the law's exponent) and its standard error sigma.

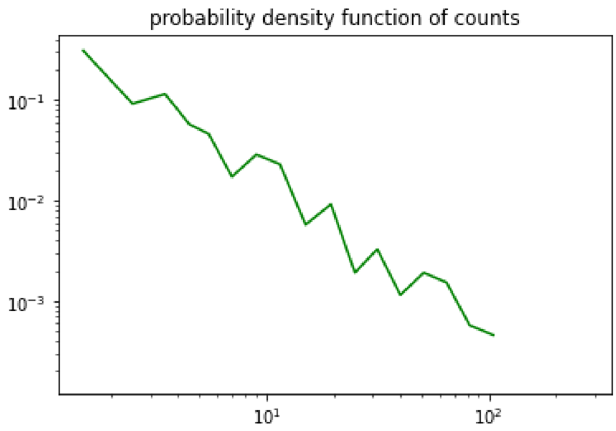


Fig. 4: Showing the power law distribution of the edge degree of each friend

```
Calculating best minimal value for power
    law fit
alpha 1.7455240062565691
xmin 3.0
signma 0.10338557814028446
```

Using the powerlaw package in python, the fit was compared against two other well known distributions, both the log-normal and the exponential, both comparisons showed that the power law was a better fit. A much better fit than exponential but however only a slightly better fit than log-normal. The results vs lognormal and exponential are given below.

	ratio	significance
lognormal	0.171	0.865
exponential	2.736	0.006

Table 2: Power law results for the IKEA friendship network

4.2 Q2. Does a rich club exist in the friendship network?

The rich-club coefficient is a metric which is designed to measure the extent to which nodes with high degree also connect to other nodes with high degree. A normalized rich-club coefficient greater than 1 shows the presence of the rich-club phenomenon with respect to the null case. In the case of the IKEA friendship network, we can see that for every degree, the coefficient is below 1. Hence it can be said there is no rich club present. Looking at the Figure below, the end of the degree plot is not as fine due to the low count of big nodes on the network.

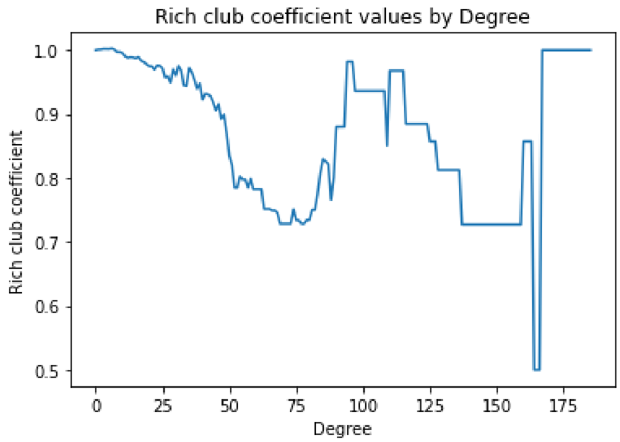


Fig. 5: Showing the rich club coefficient as a function of degree for the IKEA network

4.3 Q3. Do friendship networks differ by geographical location in terms of average degree, average shortest path, Transitivity and Assortativity?

Average degree, average shortest path, Transitivity and Assortativity are all important attributes of a social network.

- **Average degree** is the most basic measure that captures the average number of links each node the network has.
- **Average shortest path** is as it says, the average of all the shortest paths for all possible pairs of the network nodes.
- **Transitivity** is the fraction of all possible triangles present in the friendship network. It refers to the extent to which a link between two nodes in a network are connected. It is also known as the clustering coefficient.
- **Assortativity** measures the preference for a networks nodes to attach to others that are alike in some way. This coefficient is the Pearson correlation coefficient of the node's degree between pairs of linked nodes, so in general is lies between 1 and -1, -1 meaning there is no Assortativity at all and 1 meaning there is perfect Assortativity.

In order to complete this task, a python panda's dataframe was completed to store the values for each state. The results are shown below and the code used is found in the attached file `yelp_ networkx.ipynb` where many more networks metrics were calculated such as "outsider_friendships", "percentage of network which is the Giant Component"(perc_gcc) and "average_clustering".

state	avg_degree	perc_gcc	edges	nodes	transitivity	average_shortest_path_GC	assortativity
OH	4.341232	100.000000	458	211	0.093584	2.501061	-0.475013
FL	9.100000	99.615385	2366	520	0.202849	2.831617	-0.331114
MA	6.329571	98.645598	1402	443	0.129132	3.017561	-0.170801
GA	9.931777	98.025135	2766	557	0.261351	2.973815	-0.214499
BC	7.777778	99.305556	1120	288	0.219885	2.804417	-0.320016
OR	6.985294	96.078431	1425	408	0.196946	2.962133	-0.237391
VA	4.319527	98.816568	365	169	0.140507	2.694611	-0.541086
ALL	11.571809	98.537234	8702	1504	0.215305	3.296666	-0.169191

Table 3: Network metrics for each state

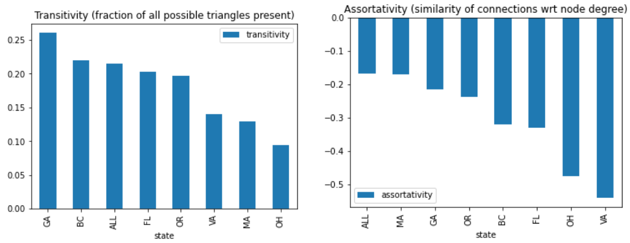
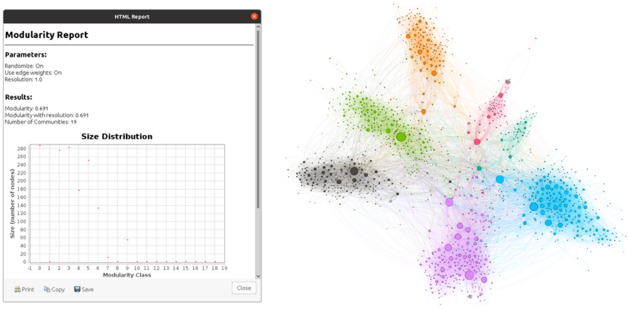


Fig. 6: Transitivity and Assortativity for the friendship networks in each state with an IKEA store (”ALL” refers to all states)

Again like with the rich club question, the Assortativity goes negative for this network for each state meaning the large degree nodes tend to attach to low-degree nodes. As for the Transitivity, it can be seen that Canda's state BC(British Culumbia) and GA(Gerogia) are above average for connected the triangle of friendship.

4.4 Q4. Can the network be split using Modularity-based community detection? what is the Modularity score?

To answer this question, Gephi (https://gephi.org/) was used, Gephi has an implementation of the Louvain method (Blondel et al., 2008) to break down the network into communities and find the modularity score. The method is a greedy optimization method that tries to optimize the "modularity" of a partition of the network. The optimization is done in two steps. First, the method looks for "small" communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. This method is much faster than the Girvan–Newman method which is used in the next research question.



HTML Report

Modularity Report

Parameters:

Randomize: On
Use edge weights: On
Resolution: 1.0

Results:

Modularity: 0.691
Modularity with resolution: 0.691
Number of Communities: 19

Fig. 7: Gephi's Louvain method for finding communities.

Gephi's Louvain method found a total of 19 clusters/communities. However as can be seen in the Chart on the left above, the size of communities 10-19 is very small. The resulting Modularity score was 0.691. The larger the variation from 0 the better the fit, so 0.691 would suggest there exists communities inside this network.

4.5 Q5. Can the Girvan–Newman method split the IKEA friendship network into it's geographical parts or communities? Which split is best based on modularity scores?

This was answered by using python's networkx package, the Girvan–Newman method was used to find the first 6 split in the IKEA friendshsip network. The highest modularity score was found with split number 7. Each split with this algorithm is very slow because it must calculate inbetweeness at each step, the execution to find split number 7 took 1 hour 52 minutes and 45.92 seconds. Below is a table showing the split, modularity and times for each split.

splits	modularity	time
2	0.18547764268933936	3788.9
3	0.3697724721570822	5099.97
4	0.5169992827419141	6145.51
5	0.5544311866151118	6575.19
6	0.5545102641304193	6761.96
7	0.5545892820694971	6765.92

Table 4: Girvan–Newman method results on the IKEA friendship network

Below is a depiction of the split at 7, which forms 7 communities,along with a printout of the executed code that show the breakdown of those communitites in term of users and whihc states they have reviewed in, it can clearly be seen that each community is based on location.

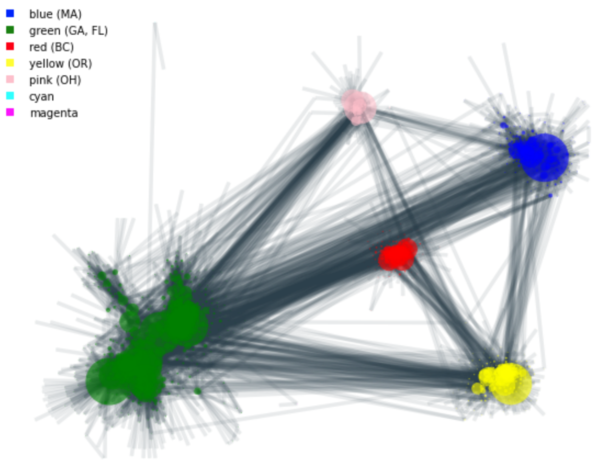


Fig. 8: Communities of the IKEA friendship network after split 7. (showing the states which are majority of each community)

The following text displays the details of the Figure above:

```
285 blue {'MA': 272, 'GA': 4, 'OR': 4, 'FL': 3, '
    BC': 2} {False: 279, True: 6}
659 green {'GA': 318, 'FL': 263, 'VA': 46, 'MA':
    15, 'OR': 14, 'OH': 3} {False: 645, True: 14}
178 red {'BC': 171, 'OR': 5, 'GA': 2} {True: 176,
    False: 2}
241 yellow {'OR': 235, 'MA': 3, 'BC': 2, 'OH': 1}
    {True: 237, False: 4}
115 pink {'OH': 99, 'GA': 6, 'FL': 5, 'MA': 3, 'VA
    ': 1, 'OR': 1} {False: 114, True: 1}
2 cyan {'GA': 2} {False: 2}
2 magenta {'GA': 2} {False: 2}
modularity 0.5545892820694971
```

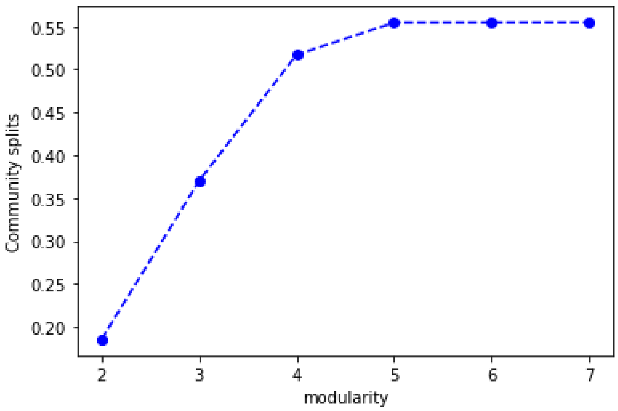


Fig. 9: Modularity score for each split using the Girvan–Newman method in networkx

The modularity for this split (0.55) using the Girvan–Newman method doesn't compare well to Gephi's Louvain method(0.69). The Louvain method is more accurate at breakdown networks and is also much faster to execute.

4.6 Q6. Do IKEA users who give poor ratings tend to make friendships with other uses who also give poor ratings?

In order to answer this question, a ERGM model will be used. ERGM stands for Exponential Random Graph Model, ERGMs work similar to logistic regressions, they predict the probability that a pair of nodes in a network will have a tie between them. The basic idea behind the method is a comparison of an observed network to Exponential Random Graphs, therefore to setup these random graphs, execution on a single personal computer can be slow as the network grows.

The data needs to be imported into R in order to use an ERGM model, to do this the edge list and user attribute csv are loaded. All of the code to run this section of the project can be found in `IKEA_ERGM.R` Below is a view of the network showing the states which each user placed a review in, it shown in contrast to the depictions above, there is no community detection happening in this view. IN this view, it can be seen that some friendships don't follow state lines and some don't have friends at all in their own state, perhaps because they moved address.

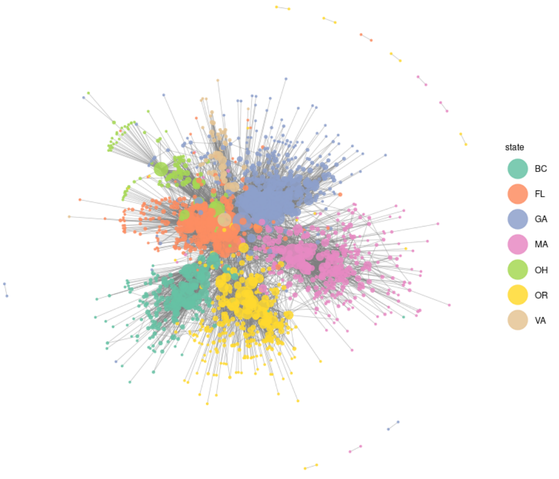


Fig. 10: Depictation of the network using R's ggnet(https://briatte.github.io/ggnet/) package

Model with only edges and average_stars:

```
Call:
ergm(formula = IKEA.net ~ edges + nodecov("
    average_stars"))

Maximum Likelihood Results:

                          Estimate Std. Error MCMC % z
                               value Pr(>|z|)
edges                     -6.10276    0.15565      0
    -39.208    <1e-04 ***
nodecov.average_stars   0.16288    0.02027      0
      8.034    <1e-04 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05
    ’.’ 0.1 ’ ’ 1

     Null Deviance: 1566868  on 1130256  degrees
          of freedom
 Residual Deviance:  101970  on 1130254  degrees
     of freedom

AIC: 101974  BIC: 101998  (Smaller is better. MC
    Std. Err. = 0)
> plogis(coef(m2)[[’edges’]])
[1] 0.002231699
> plogis(coef(m2)[[’edges’]] + coef(m2)[[’nodecov.
    average_stars’]])
[1] 0.002625435
```

The results of this model show that adding average stars is statistically significant, however the coefficient is very small, so that model is not any better using the average stars attribute as an indicator of the chance of a link between 2 random friends. The complete model only shows a 0.2 % chance of a link when both these attributes are known. Therefore poor ratings are not a good indicator of friendship.

4.7 Q7. Which attribute is the most important for making friendships in the IKEA friendship network?

A ERGM model was created using all attributes of the friend. Results are shown below:

```
Call:
ergm(formula = IKEA.net ~ edges + nodematch("west
    ") + nodecov("funny") +
    nodecov("cool") + nodecov("useful") +
        nodematch("state"))

Maximum Likelihood Results:

                  Estimate Std. Error MCMC %  z
                      value Pr(>|z|)
edges            -7.686e+00  5.535e-02      0
    -138.862   <1e-04 ***
nodematch.west   5.921e-01  6.491e-02      0
    9.122   <1e-04 ***
nodecov.funny   -9.459e-05  4.931e-06      0
    -19.182   <1e-04 ***
nodecov.cool    -1.740e-04  6.122e-06      0
    -28.414   <1e-04 ***
nodecov.useful   2.470e-04  4.973e-06      0
    49.671   <1e-04 ***
nodematch.state  3.574e+00  4.276e-02      0
    83.589   <1e-04 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05
    ’.’ 0.1 ’ ’ 1

     Null Deviance: 1566868  on 1130256  degrees
          of freedom
 Residual Deviance:   75949  on 1130250  degrees
     of freedom

AIC: 75961  BIC: 76033  (Smaller is better. MC Std
    . Err. = 0)
> plogis(coef(m4)[[’edges’]] + coef(m4)[[’
    nodematch.west’]]  + coef(m4)[[’nodecov.funny
    ’]]  + coef(m4)[[’nodecov.cool’]] + coef(m4)
    [[’nodecov.useful’]])
[1] 0.0008292781
> plogis(coef(m4)[[’edges’]] + coef(m4)[[’
    nodematch.west’]]  + coef(m4)[[’nodecov.funny
    ’]]  + coef(m4)[[’nodecov.cool’]] + coef(m4)
    [[’nodecov.useful’]] + coef(m4)[[’nodematch.
    state’]] )
[1] 0.0287529
```

When the coefficients of all attributes are compared, it is clearly visible that the "state" attribute has the most influence on establishing a friendship link in this network, which shouldn't come as any surprise since we could split the network earlier based on users locations. Before using "state", there is only a 0.08% chance of a link using the other attributes, increasing to 2.8% with "state". In addition to that, using all attributes, the model is still not successful at predicting a link between two friends, only managing a 2.8% chance of a link if "state","funny","west","cool" and "useful" attributes are all known. Since there was a lot of filtering performed on the dataset at the beginning of this project, the model could be missing many important links that could only be explored if a more boarder view of the yelp data is explored.

5 Conclusions

During this project, the social network of the IKEA reviews in a yelp dataset was analyzed. The network was shown to follow the power law, it did not have the characteristics of a rich club, Girvan–Newman and Louvain methods were used to detect communities which confirmed that location is the most important identity of the community. ERGM models were also created to see if any attributes could predict the likelihood of a successful friendship, the attribute "state" which defines the location of the user is the best predictor.

Future work could expand the data section to include perhaps all stores in one location so as to not put such an importance on the "state" variable. In addition to the two clustering methods used in this project, it could be interesting to see how they compare against more traditional machine learning methods such as k-nearest neighbors and Support-vector machines. Another idea for future work could involve developing a weighted network of the businesses in Yelp, the weight would be the amount of user reviews in common.

References

Amato, F., Colace, F., Cozzolino, G., Moscato, V., Picariello, A., and Sperlí, G. (2017). Analysis on yelp social network.

Berkani, L. (2020). A semantic and social-based collaborative recommendation of friends in social networks. *Software - Practice and Experience*, 50:1498–1519.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.

Butts, C. T. (2020). Tools for social network analysis [r package sna version 2.6].

Cervellini, P., Menezes, A. G., and Mago, V. K. (2017). Finding trendsetters on yelp dataset. *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016.*

Chen, J., Lin, X., Wu, Y., Chen, Y., Zheng, H., Su, M., Yu, S., and Ruan, Z. (2018). Double layered recommendation algorithm based on fast density clustering: Case study on yelp social networks dataset. *2017 International Workshop on Complex Systems and Networks, IWCSN 2017*, 2018-January:242–252.

Krivitsky, P. N., Hunter, D. R., Morris, M., and Klumb, C. (2021). ergm 4.0: New features and improvements.

NetworkX (2021). Networkx — networkx documentation.

Pranata, I. and Susilo, W. (2016). Are the most popular users always trustworthy? the case of yelp. *Electronic Commerce Research and Applications*, 20:30–41.

Rahimi, S., Mottahedi, S., and Liu, X. (2018). The geography of taste: Using yelp to study urban culture. *Canadian Historical Review*, 7.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., Franc¸ois,R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Yelp (2021). Yelp dataset.

Yu, S., Wang, J., Zhang, D., and Shen, B. (2018). Analysis of gender motifs in online yelp friendship network. *2017 International Workshop on Complex Systems and Networks, IWCSN 2017*, 2018-January:279–283.