Capstone project: Yelp social network analysis

Panos Christidis

20 November 2015

1.Title

The capstone project analyzed the social networks formed among Yelp users by creating a graph based on their friend lists. The overall question was whether friends influence each other in choices and evaluations (i.e. whether groups of friends have different behaviour from the rest of the users). The methodology includes the identification of "cliques" and influential users (by estimating centrality and betweeness indicators) and the comparison of the group dynamics between different cities in the dataset.

2.Introduction

The data for this capstone project came from Yelp, which is a business founded in 2004 to "help people find great local businesses like dentists, hair stylists and mechanics." As of the second quarter of 2015, Yelp had a monthly average of 83 million unique visitors who visited Yelp via their mobile device and written more than 83 million reviews.

3.Methods and data

Yelp data was available in five distinct datasets, all in json format:

- review_data: a dataset including all review texts with the id of the corresponding Yelp user, the id of the business for which the review was made, the review id, the number of stars awarded to the business by the reviewer, the date of the review and the votes the review got from other users
- business_data: a dataset with the characteristics of each business, including its id, full address, coordinates, city, state, category of business and stars
- tip_data: a dataset of the "tips" given by users to specific businesses on a specific date
- user_data: data on each user, including the time since they are Yelp members, their number of reviews, average stars, compliments, as well as a complete list of their Yelp friends
- checkin_data: data on the number of checkins for a particular business on each specific date in the dataset

While there is a wealth of information in these datasets, the analysis focuses on the variables that can give answers to the main social network analysis issues. The first steps of data processing before the main analysis were:

- Extract a dataset of all users with their corresponding friends
- Extract dataset with businesses reviewed by each user and stars awarded in each review
- Identify where (which state) each business is located and make a different dataset for users who have made a review in each state
- Join datasets and create a different dataset for each state that includes users with reviews in that state and their list of friends

An important step is to create a data frame with all the friendship relations in the dataset for each state. Each row represents one friendship connection, which is extracted from each user's list of friends. In parallel, data on each user's number of friends, number of reviews and average stars is collected.

```
# calculate average stars given, number of friends per user and create a graph from data on friends
gg<-NULL
d<-user_friends_clean_EDH
nfriend<-NULL
for (i in 1:nrow(d)) {
aa<-as.data.frame(d[i,5])
ff<-as.character(aa[,1])
bb<-as.data.frame(rep(d[i,1],nrow(aa)))
nfriend1<-cbind(d[i,1],nrow(aa))
nfriend<-rbind(nfriend,nfriend1)
cc<-cbind(bb,ff)
gg<-rbind(gg,cc)}

### from igraph package: create a graph from data frame we created based on the friendship relations
g<-graph_from_data_frame(gg, directed=FALSE)
```

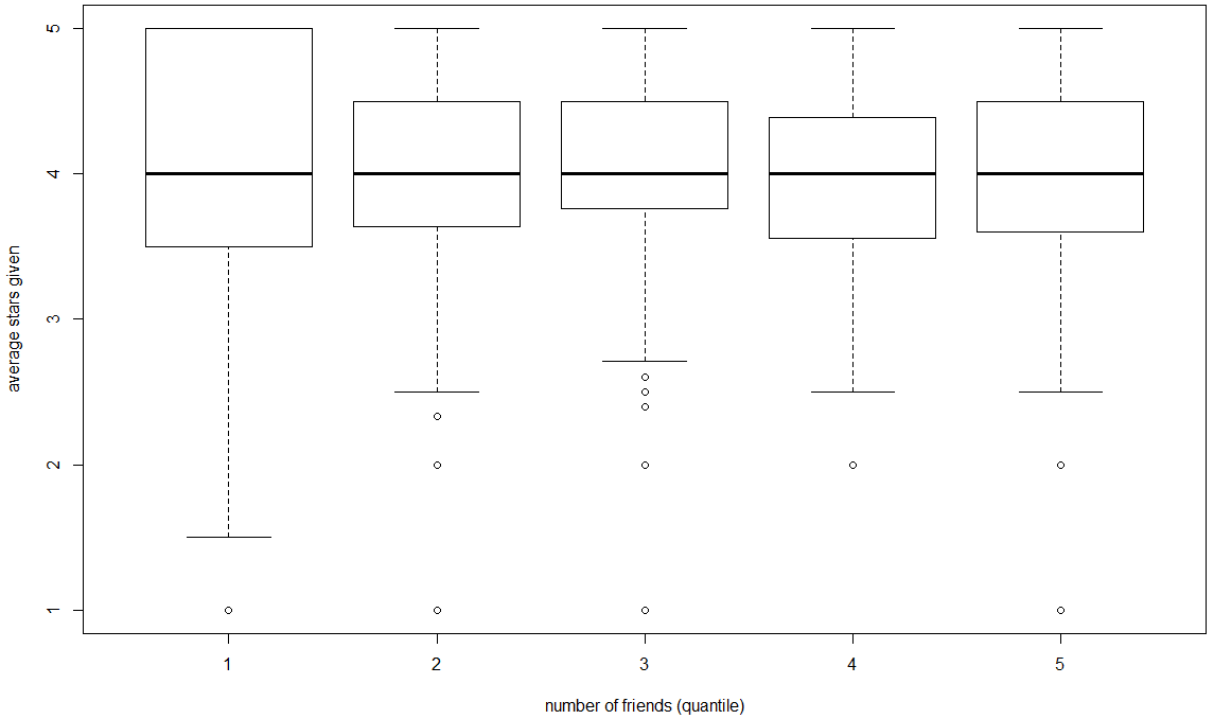
With the help of the **igraph** package, our data is now in the form of a graph.
We can also get some overall statistics.

```
nfriend<-as.data.frame(nfriend)
colnames(nfriend)<-c("user_id","nfriends")
nfriend$user_id<-as.character(nfriend$user_id)
nfriend$nfriends<-as.integer(nfriend$nfriends)
```

Joining all relevant fields, we have a full picture of each user's basic statistics:

```
head(nfriend)
##              user_id                    nfriends     nreviews  av_star gr_    fr
## 1 --4fX3LBeXoE88gDTK6TKQ       55                5         3.400000    3
## 2 --EBmNaY9_XSLqmT7ObsBw    1                    1         1.000000    1
## 3 -_JYArEMuIqOcuZfay_A8w       2                 2         2.500000    1
## 4 -0DRa69RLdSSNrc57P_5ew       117               9         3.666667    6
## 5 -4sXlp6iJxEzA9TnRTGcFw       1                 7         3.714286    1
## 6 -6JkSBDWaJqxPbbyz3hRSA       1                 1         5.000000    1
```

The boxplot of average stars for each quintile of the users according to number of friends implies that there is no significant difference between people with many friends and people with fewer friends in terms of average stars given. The mean for each quintile is almost identical and the main difference is variance, larger in the first quintile.



We add three graph metrics that are important in order to identify how important the role of each user is in the social network of each state (using **igraph**)

Degree: the number of connections of each user, in practice equal to each users number of friends. We normalise on the basis of the total connections in the graph of each State to remove the potential impact of the size of each graph.

```
V(g)$degree_norm <- degree(g, normalized = T)
```

Closeness: how "close" is each user to all other users in the State (high closeness means that a user would need to go through fewer common friends in order to reach anyone else in the system):

```
V(g)$closeness_norm <- closeness(g, normalized = T)
```

Betweenness: an indicator of a node's centrality in a network. A user with high betweenness has in theory a large influence on the connections between users through the network, under the assumption that the connections follow the shortest paths:

```
V(g)$betweenness_norm<- betweenness(g)
```

A first model to try is one using only the simple user statistics for Edinburgh:

```
model_lm<-lm(av_star~nfriends+nreviews,data=nfriend)
Call: lm(formula = av_star ~ nfriends + nreviews, data = nfriend)
Coefficients:     (Intercept)       nfriends          nreviews
                  3.9168367         0.0008949        -0.0009484
```

```
Analysis of Variance Table
Response: av_star Df Sum Sq Mean Sq F value Pr(>F)
nfriends 1 3.4 3.4395 3.1514 0.07596 .
nreviews 1 3.0 3.0205 2.7675 0.09630 .
Residuals 3095 3377.9 1.0914
--- Signif. codes: 0 '' 0.001 '' 0.01 '' 0.05 '.' 0.1 ' ' 1
```

According to this simple model, people with more friends tend to give more stars, while people with more reviews tend to be more negative. The result makes sense, but the two variables are only statistically significant at the 0.1 level. Additional combination tried on the simple user statistics did not reveal any particular effect.
A second model uses the graph statistics derived, again for Edinburgh:

```
model_lm2<-lm(av_star~betweenness_norm+nreviews,data=dt)
Call: lm(formula = av_star ~ betweenness_norm + nreviews, data = dt)
Coefficients:     (Intercept)       betweenness_  norm nreviews
                  3.994e+00         8.450e-09         -1.011e-03

Analysis of Variance Table
Response: av_star Df Sum Sq Mean Sq F value Pr(>F)
betweenness_norm 1 0.11 0.1130 0.1620 0.68739
nreviews 1 3.41 3.4072 4.8837 0.02724
Residuals 1774 1237.67 0.6977
--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We expected betweenness to play an important role, but this is not the case for Yelp social networks. However, closeness does play a role, together with degree (the number of connections:

```
model_lm3<-lm(av_star~degree_norm+closeness_norm-1,data=dt)
Call: lm(formula = av_star ~ degree_norm + closeness_norm - 1, data = dt)
Coefficients:     degree_norm    closeness_norm
                  -0.4588           2201.3280
Analysis of Variance Table
Response: av_star Df Sum Sq Mean Sq F value Pr(>F)
degree_norm 1 2788.3 2788.3 1139.5 < 2.2e-16  ***
closeness_norm 1 22330.3 22330.3 9125.6 < 2.2e-16  ***
Residuals 1775 4343.4 2.4
--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both degree and closeness are significant at the 0.001 level. The direction of the coefficients suggest that at least in Edinburg users closer to the centre of the community give a higher average star value, but this is counterbalanced by the people with many connections, which tend to be more negative.
A final check on the cliques:

```
mxcl<-max_cliques(g,min=17)
head(mxcl,1)
[[1]] + 17/14351 vertices, named: [1] Oefcy4KQN9wVpIxGe_zraA h4tfHGODErFCFrg6KEZ3xw
LhcGcnw4eHaxzg58y2YYUA OA9I8dKT9Uims5rp7YFKzA [5] WCTHtHnAdyQVCOkftfMgcQ -gg-
WvyzPXVjmbqMIVBP4w DNXdfzyuN64b38b3CbumXQ PmXejC0y5pT7ZS0x5s19aQ [9] 9tAQkCB7Ok-
```

SHPpMhIxWeg CAgh-C_qcQ2MIcGy0qiLFA geywmKDjKrdl_pzeGoLcqg 02lE5Clq84ZyxrdmMN3MvA [13] -fyIkHBKXXHUrUjW1Lit-Q JCkzKmJYVfA6MEhsShpfsg -f-PipG4HL0nORgWzjc7qw 2mBin9oV_a3QnbjidmbVaA [17] wtp5CmMVic44bYTBKLjkDg

The size of the largest clique (the largest subset with all its members connected between them) is only 17, out of a total of 14351 users, which in practice makes it irrelevant as regards influencing the results of the whole system in Edinburgh.

4.Results

The methodology described in the example of Edinburgh (section 3) was applied to another seven states in the dataset. Model 2 (based on betweenness) did not produce any significant results in any state, but model 1 (based on simple connection statistics) and model 3 (based on closeness) did produce significant results that can provide some insights for social network analysis.

State	model1 nfriends	model1 nreviews	model3 degree_norm	model3 close_norm	max_clique
EDH	0.0008949	-0.0009484	**-0.459**	**2201.328**	17
AZ	**0.0020658**	-0.0007658	**-2.557**	**9181.679**	14
NC	**0.002576**	-0.0006403	**1.232**	**4272.807**	15
NV	**0.0013204**	0.0009842	**4.541**	**8063.476**	15
IL	**0.002617**	-0.0046430	**-0.085**	**2727.490**	14
PA	**0.0025077**	-0.0009903	**2.721**	**3102.713**	15
QC	**0.002063**	-0.0007510	**4.844**	**3020.761**	17
WI	**0.002876**	-0.0030120	**0.539**	**3431.677**	18

Figures in **bold** are statistically significant at the 0.001 level

5.Discussion

The results for each state analysed show some differences but still imply that the methodology followed is robust and that some conclusions can be generalised:

- Users with higher closeness have a higher impact on the local Yelp user network. They are users who connect different parts of the local Yelp community and tend to give higher stars than the average user.
- The number of friends a user has is positively correlated with the stars given. But if closeness is taken into account the correlation may turn negative. This suggests that what is important is the centrality of the user in the network and not necessarily the user's number of friends. Although these may be correlated, it is not always the case (one may have many friends but be isolated in a disconnected part of the network).
- Cliques have a negligible role in the dynamics of Yelp users. Users tend to have an average of about 10 connections (but variance is very high) and do not tend to form small (closed) groups.
- The results suggest that if we want to identify most influential user we should look into people having contacts with diverse subgroups in the community of Yelp.