Analysis of Yelp Social Networks - Finding my friends influence on me!!

Authors:
1) Harish Kumar Ravichandran
2) Saisruthi Sathyanarayanan
3) Deepak Kumar Balasubramaniam

Analysis of Yelp Social Networks - Finding my friends influence on me!!

Team Members: Harish Kumar Ravichandran
Saisruthi Sathyanarayanan
Deepak Kumar Balasubramaniam

Contents:

1. Abstract
2. Introduction
3. Influence in Social Networks
4. Yelp Network and dataset
5. Hypotheses
6. Implementation
   6.1 Data parsing and cleaning
   6.2 Hypothesis 1
   6.3 Hypothesis 2
   6.4 Hypothesis 3
7. Results and Analysis
   7.1 Hypothesis 1
   7.2 Hypothesis 2
   7.3 Hypothesis 3
8. Future scope
9. Conclusion

1. Abstract

Study of social networks' influence on a node and its connections has been a trending topic in recent days. Online reviews intertwined with the social networks have a huge impact on the business of the products. One such social network is Yelp, where you can from friends and post reviews about businesses. This work focuses on studying the behavior of social influence on a user's contribution at a particular time. For this purpose we study three hypotheses proposed in the base paper and analyze the results.

2. Introduction

In a world of ever-growing technology, almost everything is available online. Online reviews about products, services, stores have become a dominant source of information these days. Before purchasing a product, it has become very common for people to see online reviews about the product in the first place. Hence, online reviews about a product or a service is of utmost importance these days. In a recent survey conducted, it was found that more than three-quarters of online review readers reported that the reviews had a significant influence on their purchases. Thus, we can see that online reviews have a significant impact on the purchase but unfortunately, there are not as many review contributors compared to review readers. Thus, each online review platform promotes its users to post reviews in different ways. Yelp has a really promising approach. Yelp does not pay users to write reviews but adopts a community strategy where a user is allowed to make friends, upvote for another review, send compliments to other users. Though this is unpaid, its success suggests that social interactions among reviewers play an important role in review generation.

All social networks primarily involve interactions with friends. Some friends may be highly influential while some friends may not be that influential. Some users can be easily influenced by their friends while some may not get influenced by any of their friends. A user will generally have friends in more than one social circle. A user will be influenced more by friends belonging to the social circle with which the user has strong ties than by his other friends belonging to other social circles. The users' behavior based on friends can have any of the above possibilities and analyzing this in a social networking platform is one the most interesting topics in today's world of increasing communication where it is very easy to stay in touch with someone in any part of the world.

Yelp is a social networking platform, allowing its users to post reviews about businesses and these reviews are openly available to everyone so that other users can read about it and get to know the opinions of the co-users. Do the reviews of co-users influence me? Do the reviews of my friends influence me? This project analyzes these questions given the Yelp dataset collected from 2001 until 2014. The project mainly quantifies the number of reviews written by the user and his friends. Hence, the major amount of information is retrieved from the social circle of the user and the review contribution from him and his social circle.

This analysis is different from previous studies in 2 ways. One primary difference is that, in this analysis, the primary focus is on the information and social environment surrounding users of Online Review Platforms (ORPs) and examining the roles of different information sources and the factors that moderate their importance. Also, unlike previous studies, rather than modeling social relations as a first-order source of influence, this analysis focuses on the information flow between social actors and model social relations as moderators of the value of information. The actual flow of interactions between users, now widely accessible to researchers are carriers of social influence and hence should be the focus of social influence studies.

3. Influence in Social Networks

Social influence occurs when one's emotions, opinions, or behaviors are affected by others. Social influence takes many forms and can be seen in conformity, socialization, peer pressure, obedience, leadership, persuasion, sales and marketing. The base paper studies the yelp social network and proposes three hypotheses. The hypotheses are based upon on the following theories of influence:

Awareness - A user gets aware of a business from the reviews made by his friends. Hence the user is likely to visit or review that business.

Reciprocity - A user may contribute as a reciprocal activity to his friends contribution. Reciprocal activity is considered to be an instance of influence.

Social Norm - A user may contribute in order to abide to the social norm of his friends.

4. Yelp Network and dataset

The dataset for this project was taken from the academic dataset challenge organized by Yelp. This dataset consisted of rich research data. It consists of dataset collected at Seattle since it was believed that the total number of reviews and the number of reviews per restaurant in Seattle are close to the mean of 21 popular cities listed on front page of Yelp during that data collection period.

Each file is composed of a single object type, one json-object per-line. The main objects of the dataset include business, review, user, check-in and tip. The entire dataset was represented in 5 files which included all information about users, reviews, businesses, tips, checkin information in the yelp. It represented a social network of 366K users and 1.5M reviews for a total of 2.9M social edges. Dataset processing was difficult because of the enormity of the information provided, like 1.5 Gb of data. Of the enormous amount of information provided, this project necessitated the use of the following datasets - 1. user dataset 2. review dataset 3. business dataset (to identify similar business)

The user dataset consisted of the following information:

```
'type': 'user',
'user_id': (encrypted user id),
'name': (first name),
'review_count': (review count),
'average_stars': (floating point average, like 4.31),
'votes': {(vote type): (count)},
'friends': [(friend user_ids)],
'elite': [(years_elite)],
'yelping_since': (date, formatted like '2012-03'),
'compliments': {
    (compliment_type): (num_compliments_of_this_type),
    ...
},
'fans': (num_fans)
```

The review dataset consists of:

```
'type': 'review',
'business_id': (encrypted business id),
'user_id': (encrypted user id),
'stars': (star rating, rounded to half-stars),
'text': (review text),
'date': (date, formatted like '2012-03-14'),
'votes': {(vote type): (count)}
```

Business dataset:

```
'type': 'business',
'business_id': (encrypted business id),
'name': (business name),
'neighborhoods': [(hood names)],
'full_address': (localized address),
'city': (city),
'state': (state),
'latitude': latitude,
'longitude': longitude,
'stars': (star rating, rounded to half-stars),
'review_count': review count,
'categories': [(localized category names)]
'open': True / False (corresponds to closed, not business hours),
'hours': {
    (day_of_week): {
        'open': (HH:MM),
        'close': (HH:MM)
    },
    ...
},
'attributes': {
    (attribute_name): (attribute_value),
    ...
}
```

There are various other information presented in the dataset such as :

Check-in dataset

```
'type': 'checkin',
    'business_id': (encrypted business id),
    'checkin_info': {
        '0-0': (number of checkins from 00:00 to 01:00 on all
Sundays),
        '1-0': (number of checkins from 01:00 to 02:00 on all
Sundays),
        ...
        '14-4': (number of checkins from 14:00 to 15:00 on all
Thursdays),
        ...
        '23-6': (number of checkins from 23:00 to 00:00 on all
Saturdays)
    }, # if there was no checkin for a hour-day block it will not be
in the dict
```

Tip dataset:

```
'type': 'tip',
    'text': (tip text),
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'date': (date, formatted like '2012-03-14'),
    'likes': (count),
```

5. Behaviors to be analyzed

This project focuses on analyzing the influence of contributing friends, non-contributing friends and redundancy. There may be various kinds of users in a social network. There may be users who are easily influenced by other users/friends or there may be users who are not easily influenced by other users (elite users). Also, a user may not be easily influenced by other users in general, but he may have friends who are highly influential that the user ends up being influenced by his friends. But elite users, no matter how influential their friends are, remain uninfluenced by his friends. There are various behaviors that could be analyzed in this perspective some of which have been analyzed as a part of this project.

Only if a user is aware of a particular store, there is a possibility that he may visit the store and review it. Reading others' reviews is a dominant source of awareness about different kinds of stores. The more the number of reviews he reads, the more awareness he gets about the different kinds of stores. Given that, in a social network of many number of users, a user is most likely to read reviews posted by his friends whom he trusts rather than reading the reviews of some unknown users. A user may have both contributing as well as non-contributing friends. Friends' contributions may affect the user (through reciprocity) and vice-versa. Furthermore, the users of Yelp may be either regular users or elite users. Regular users may get easily influenced by friends who review dissimilar stores rather than similar stores. But elite users will not get influenced by any of his friends, both of those who review similar and dissimilar stores. Elite users are uninfluential users whose behaviors are different from those of normal users. There are many such behaviors that can be analyzed in a social networking platform, three of which have been analyzed as a part of this project and they are listed below.

- Whether a user's contribution in period t is positively related to the number of contributing friends in period t-1.
- Whether a user's ratings in period t is influenced by the contribution of his friends in t-1 or not.
- A user's elite status dampens the negative effect of redundancy, that is, an elite user is more positively affected by her contributing friends who review similar stores than non-elite users are.

6. Implementation

6.1 Data:

6.1.1 Data Cleaning:

The yelp dataset is really huge especially user dataset and user’s reviews dataset. The dataset contained details about 366,000 users and 1,500,000 reviews and had about 2,900,000 edges. Since the dataset contained some information which we did not need like “text” in reviews dataset which contained what the user gave as a review. Instead we used similar information from the “Stars” field, which had the number of stars out of 5, a particular user gave for a business. Similarly all these datasets contained unique pseudo-random ID for users and reviews. These were replaced with our generated ID for easy parsing and tracking.

Original: {"votes": {"funny": 0, "useful": 2, "cool": 0}, "user_id": "H1kH6QZV7Le4zqTRNxoZow", "review_id": "RF6UnRTtG7tWMcrO2GEoAg", "stars": 2, "date": "2010-03-22", "text": "Unfortunately, the frustration of being Dr. Goldberg's patient is a repeat of the experience I've had with so many other doctors in NYC -- good doctor, terrible staff. It seems that his staff simply never answers the phone. It usually takes 2 hours of repeated calling to get an answer. Who has time for that or wants to deal with it? I have run into this problem with many other doctors and I just don't get it. You have office workers, you have patients with medical needs, why isn't anyone answering the phone? It's incomprehensible and not work the aggravation. It's with regret that I feel that I have to give Dr. Goldberg 2 stars.", "type": "review", "business_id": "vcNAWiLM4dR7D2nwwJ7nCA"}

Pruned: {"votes": {"funny": 0, "useful": 2, "cool": 0}, "user_id": "H1kH6QZV7Le4zqTRNxoZow", "review_id": "RF6UnRTtG7tWMcrO2GEoAg", "stars": 2, "date": "2010-03-22", "business_id": "vcNAWiLM4dR7D2nwwJ7nCA"}

Modified: {"votes": {"funny": 0, "useful": 2, "cool": 0}, "user_id": "2", "review_id": "2", "stars": 2, "date": "2010-03-22", "business_id": "1"}

This brought the size of files from MB to Bytes and enabled faster reading and processing.

6.1.2 Data Processing:

For dealing with large datasets, Python’s (Pandas) dataframe is faster and helps in tracking all our data. Plus all types of data can be kept in a single table-like structure.

Sample Rows in a dataframe:

```
user_id             average_stars                         compliments  \
7                     3.02            {u'note': 1, u'photos': 1, u'funny': 1, u'plai...
723                   3.02            {u'note': 1, u'photos': 1, u'funny': 1, u'plai...

user_id      elite   fans  friends   name       review_count    type  user_id  \
7            [2014]    3   [00723]   Deepak        119            user     7
```

723	[2014]	3	[007]	Scarlett	11	user	723

user_id	**votes**	**yelping_since**	**Influence**
7	{u'funny': 52, u'useful': 209, u'cool': 33}	2007-03	Yes
723	{u'funny': 52, u'useful': 209, u'cool': 33}	2007-03	Yes

Since these data frames are volatile and computing these takes lot of time, we used ‘Python pickles’ which are used to serialize and de-serialize python object structures. All the calculated data frames are converted to pickles which are then stored locally for later use.

Because hypothesis 3 requires an approach by which review dataset has to be parsed once and user dataset twice for each review (O(n^4)) we reduced the problem to a smaller dataset of ~400 users (users and friends) and their reviews.

6.2 Hypotheses 1: Whether a user’s contribution in period t is positively related to the number of contributing friends in period t-1.

We have previously seen in the section 3 about the various theories behind the influence in social networks.Based on the functioning of the social networks the above hypothesis has been formulated.Generally a participant in a social network abides by the social norms and behaviors of social circle.Hence it is more likely that a user’s contribution in a time period is positively related to the number of contributing friends previously.Let us look into the methodology that we followed to evaluate this hypothesis in the Yelp social network.

Methodology:

The structure of the Yelp Academic and our data pruning techniques has been explained in the Section 5.The user dataset of Yelp,has an attribute called ‘yelping since’ for every user.This specifies the date from which the user has been contributing in yelp.We split this time frame into two by median (t-1,t) and compare his contribution in time period t with respect to his friends contribution in time t-1.The strategy followed is:

1. For each user take his review contribution timeline(list of reviews made from the first review till now) and find the median. The no of reviews contributed by the user in time t (i.e: after the median) is found.
2. For the user selected, identify the list of his friends and select the reviews made by them in the time period t-1.
3. The list obtained from steps 1 and 2 gives us the contribution of a user in time period t and the contribution made by his friends in time period t-1.
4. Repeat this process for the list of users in the social network.

Hypothesis 2:
Influence of friends' ratings

Every user has a unique liking in following businesses, i.e., a user may follow restaurants such as fast food centers, Italian restaurants, Chinese restaurants. But his friends may follow restaurants such as 5-star restaurants. Thus, his friends may give reviews about 5-star restaurants on Yelp. A user, thus, gets to read the reviews of his friends on 5-star restaurants and he may also end up visiting 5-star restaurants and reviewing them. A user, despite having his own liking, owing to his friends' influence, may tend to follow the businesses reviewed by his friends.

A particular user's friends, in any previous time frame, could have reviewed businesses of dissimilar interests to him and owing to such reviewing, the user may end up following those businesses (reviewed by his friends) and he may also review them. In fact, it was surprising to know that, friends' reviews on dissimilar stores have much more influence than reviews on similar stores.

Methodology

- For each user, his friends' reviews are collected until time frame t-1.
- Collected each user's reviews at time frame t.
- Compare these reviews based on the similarity of their business ids and the ratings given by the user's friends and user himself.

In the user dataset, a mapping of each user and his friends were obtained. The review dataset has a list of all users' reviews. The reviews of each user and his friends are obtained from the review dataset. For the reviews of each user and his friends, the business ids are matched to see if the user in time frame t has reviewed the same business ids as his friends (in time t-1). Matching business ids imply that a user has been influenced by his friends to review the same business ids as his friends. Further, the similarity of tastes between a user and his friends was analyzed. The review dataset has a field named "stars" that notify the extent to which a user likes that particular business. If users' friends have starred a business above 3 and the user has also starred it above 3, then it implies that the user and his friends have a similar liking and that his friends have strongly influenced him. If users' friends have starred a business above 3 and the user has starred it below 3, or vice-versa, then it implies that the user and his friends have a dissimilar liking but still his friends have influenced him to follow that business.

Hypothesis 3: Elite Status dampens the negative effect of redundancy

Yelp additionally recognizes a group of "elite" users each for their extraordinary contribution and invites them to local community events. An elite user' opinions are trusted and valued by many users. Even though there may already be reviews by other users, an elite user's opinion will still be desirable as an authoritative voice. Thus, we expect an elite user to be more likely to write reviews when receiving similar reviews by her friends. So we can conclude that an elite user is not affected by his friend's no matter they review similar/dissimilar stores.

Methodology:

In order to find if the hypothesis actually works we followed the following steps to determine the influence of a friend who reviews similar stores and also dissimilar stores on a user.

1) Find each instance of reviews from ‘review’ dataset and find the corresponding user ID of the user who has written the review
2) The date of the review can be got from the review dataset.
3) Find if the user is an elite user during when the review was given. This can be found out by checking if the user dataset has the year as the value of the “Elite” field.
4) If the user is elite during the time the review was written, get all his friends from the ‘user’ dataset
5) Find if his friends have reviewed any ‘similar’ business by checking the business ID of all friend’s reviews and if user has reviewed a similar business ID during his elite year, mark him as ‘Influenced’
6) Additional information that is collected is if the person positively or negatively influenced. If the friend and the user both have reviewed and have given 3 or more stars it means that the Elite user is really influenced. Otherwise do not mark him as influenced.
7) This process is carried out for all reviews by taking a review, finding user and checking if user’s friends have reviewed similar business.

One caveat to note here is that the friend’s reviews are taken in a time t-1 and elite user’s reviews are taken during the time t. This helps us analyse if the user is influenced by his social circle who reviews a similar business.

7. Results and Analysis

7.1 Hypothesis 1

Hypothesis 1 states that the users contribution in time period ‘t’ is positively related to the number of contribution made by his friends in time period ‘t-1’.The implementation to evaluate this is specified in Section 6.2. Since the Yelp dataset is very huge, we pruned it done to small sets of social networks of different sizes ranging from 100 users to 1000 users and collected the statistics for various sizes.

Pearson’s Correlation coefficient

In statistics,the Pearson’s correlation coefficient or the r-measure is defined as the measure of linear correlation(dependence) between two variables X and Y,giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation.We can obtain a formula for *r* by substituting estimates of the covariances and variances based on a sample. So if we have one dataset {x_1,...x_n} containing *n* values and another dataset {y_1,...y_n} containing *n* values then that formula for *r* is:

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The range of the positive correlation is given as follows:

(i) 0 - 0.25 - Weak positive correlation
(ii) 0.25 - 0.50 - Medium positive correlation
(iii) > 0.5 - High positive correlation

The number of reviews contributed by the user in T and those by the friends in T-1 form the values of X and Y for the correlation coefficient. Figure 1 is the graph showing the various values of correlation coefficient for the increasing sizes of the social network. Initially for the value of 100 and 200 nodes in the social network, there exists a medium positive correlation. It drops suddenly from 200 to 400 nodes to weak positive correlation. After the value of N = 400, the correlation coefficient becomes stagnant and remains at a value of +0.14.Let us look deep into this behavior. When the size of social network is relatively small (say 100 to 200) the number of distinct connected components is less i.e., the number of users who are completely disconnected are lesser. In this scenario the percentage of closely known people to a user is high. These friends form some of the strong ties for the user that we consider. Hence these are the people in his social circle who influence him more.

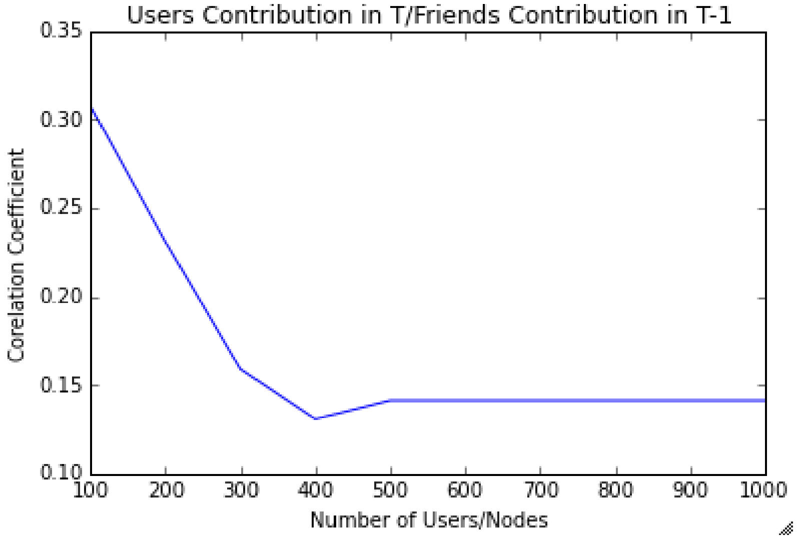


Figure 1 - Users contribution in T vs Friends Contribution in T-1

The number of reviews contributed by a user in this network size is moderately positively correlated with the number of contributions made by his friends in T-1.But when the size of a social network increases above N= 400, the correlation reduces from medium to weak positive correlation and retains the constant value thereafter. Because, in Yelp kind of social networks the upon a certain level of N, the number of strong ties to a user becomes relatively lesser than the weaker ties. These people might be mere acquaintances or friends of friends, people who met once at some restaurant/shop. They don’t have much influence on a user’s contribution.

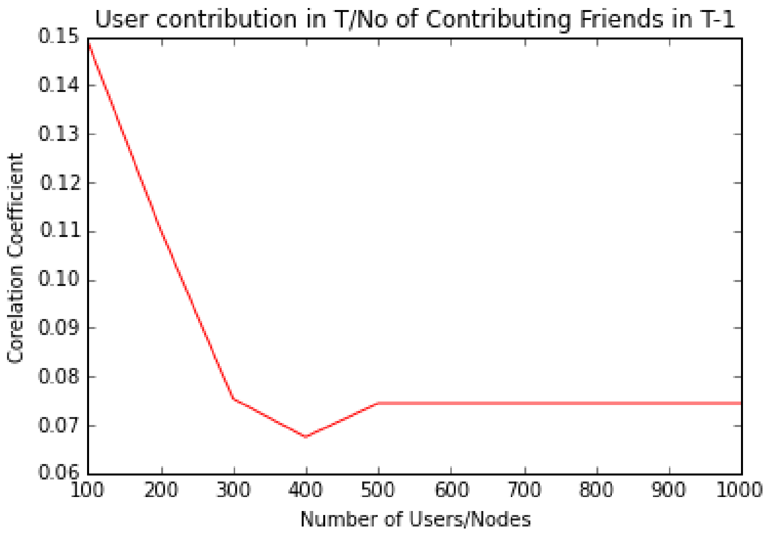


Figure 2 - User’s contribution in T vs No of contributing friends in T-1

This analysis is backed up by the observation of similar behavior of the correlation coefficient (as the N values increases) for the two other graphs. Figure 2 represents the correlation between the user’s contribution in T and no of contributing friends in T-1.Figure 3 represents the correlation between the number of review made by a user’s friends in T-1 and the number of contributing friends.

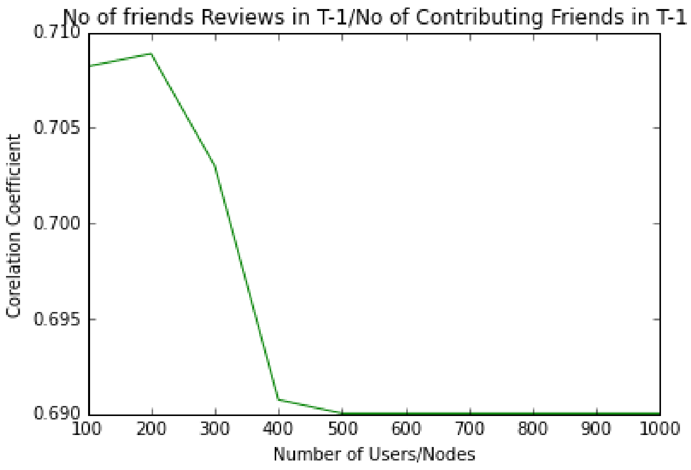


Figure 3 - No of friend’s reviews in T-1 vs No of contributing friends in T-1

7.2 Hypothesis 2

It is common for people to be attracted to novelty. Hence, if a user gets to know that his friends are checking out a new and different kind of store, then it is quite common for the user to get attracted to check out the store himself. Further by this argument, we can also state that, a friend who reviews a new or different kind of business is likely to trigger more reviews than a friend who reviews similar kind of businesses. Alternatively, we can state that, a user who reviews similar kind of businesses will often trigger less reviews compared to those who had reviewed dissimilar ones. Thus, a friend who reviews similar stores less likely adds additional value to the user and thus less likely triggers reciprocation. But whereas, owing to novelty, a friend who reviews dissimilar kind of businesses adds much additional value to the user triggering more reciprocation resulting in the user following new kind of businesses.

Thus it can be seen that it is quite common for people in a social network to be influenced by their friends. Analyzing the extent of impact of friends on a particular user is the second behavior that is to be analyzed. In Yelp, the influence of friends on a user is determined by comparison of businesses reviewed by a user and his friends. A yelp user has friends, just like users of Facebook. If a particular user's friends have reviewed a particular business, then there is a high probability that the user may also review it. Also, if the user and his friends have similar liking, they may even end up rating that particular business in a similar way. The comparison of reviews will be most efficient if comparison is based on the text in the reviews. But unfortunately, owing to the enormity of information in the dataset, it has been very difficult to compare reviews based on texts. Hence, the comparison of reviews is based on the 'stars' given to a business id in a review by the user and his friends. Comparison of reviews of a user and his friends that have the same business ids based on stars is performed for the analysis of this behavior.

The cosine similarity between i's and j's in period t along a particular dimension is calculated as follows:

$$SIM_{ijt} = \frac{\sum_{k=1}^{N}\left(R_{itk} \cdot R_{jtk}\right)}{\sqrt{\sum_{k=1}^{\mathrm{N}}(R_{itk})^2 \sum_{k=1}^{N}\left(R_{jtk}\right)^2}}$$

Where
R_{itk} - the number of reviews reviewer i writes in period t for k stores
R_{jtk} - the number of reviews reviewer j writes in period t for k stores
N - Number of dimensions of the focal vector

Based on this similarity index, the stores are classified to be similar or dissimilar stores. In the case of our project, matching business ids are considered as similar stores while non-matching business ids are considered to be dissimilar stores.

Some of the observations are represented below in the form of graphs.

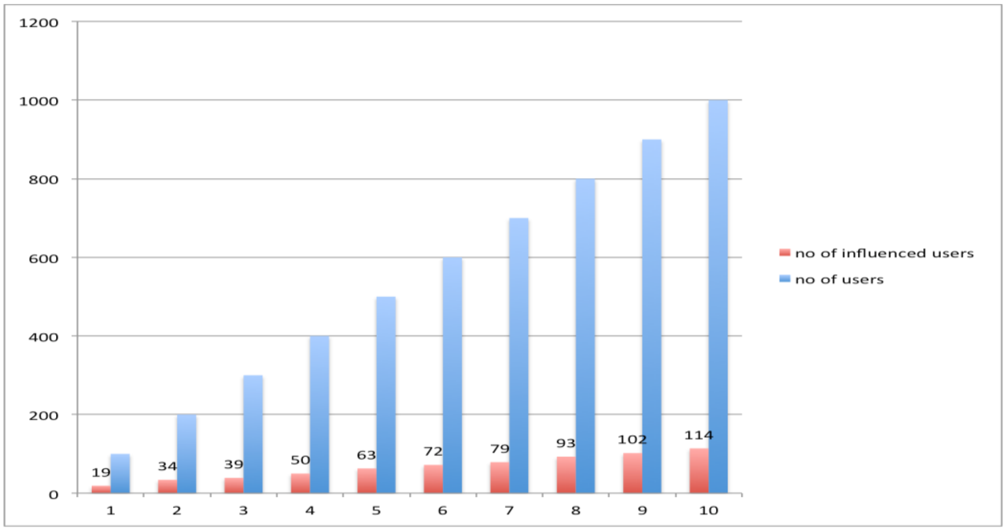


Figure 4 - No of users/No of influenced users

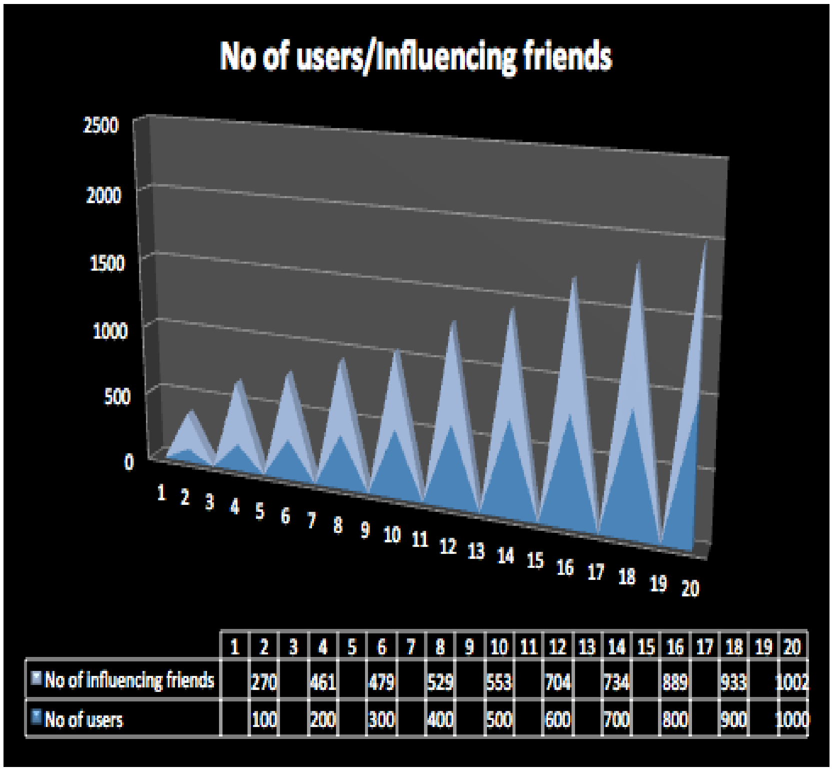
It can be seen from the graph that of the total number of users, approximately 10% of the users are influenced by their friends. We first retrieved the reviews of each user and his friends. If a friend had reviewed a business in time t-1 and the user had reviewed it in time t, then the user is considered to be influenced by his friend. The blue bar represents the number of users at time t, which increases with time, and the red bar represents the number of influenced users at time t. The reviews of users at time t (the blue bar) are compared with the reviews of their friends at time t-1. Comparing these information, the number of influenced users is calculated which is represented as the red bar.

The correlation coefficient between the total number of users and the number of influenced users was calculated and was found to be

```
[[ 1.          0.79818262]
 [ 0.79818262  1.        ]]
```

The Pearson coefficient was calculated to be 0.798.

The next comparison was made between the number of users and the number of influencing friends.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
▨ No of Influencing friends		270		461		479		529		553		704		734		889		933		1002
▨ No of users		100		200		300		400		500		600		700		800		900		1000

Figure 5 - No of users/No of influencing friends

The next observation was made on the number of influencing friends for a specified number of users. It was observed that, for 10% of users to get influenced, the number of influencing friends must be at least 10 times the number of users, i.e., to influence 10% of x users, there must be at least x influencing friends in all. When comparing the business ids of friends with the users', we maintained a count of number of friends who have an influence on the user, i.e., we stored the friends' ids of those who have reviewed a business id in time t-1 and that had had an impact on the user at time t. The count of this list gives the number of influencing friends.

The correlation coefficient between total number of users and the number of influencing friends was found to be:

```
[[ 1.          0.78518107]
 [ 0.78518107  1.        ]]
```

The pearson coefficient was found to be 0.785.

A final observation on the number of influencing friends needed to influence 10% of the total number of users was made, the results of which is presented below.

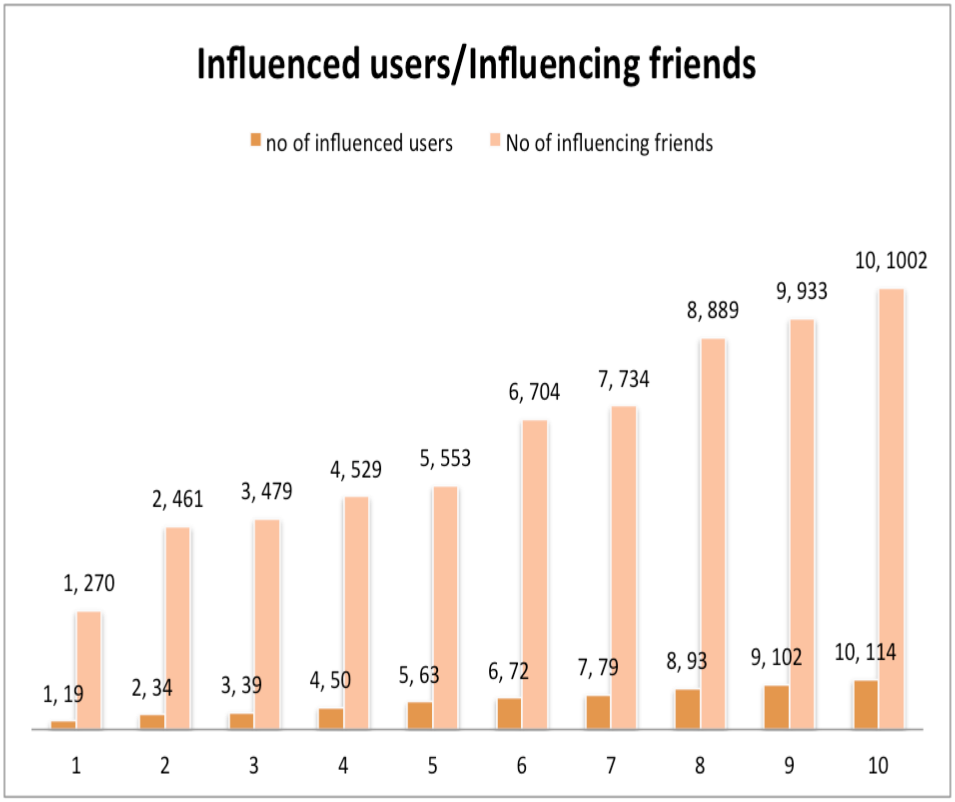


Figure 6 - No of influenced users/No of influencing friends

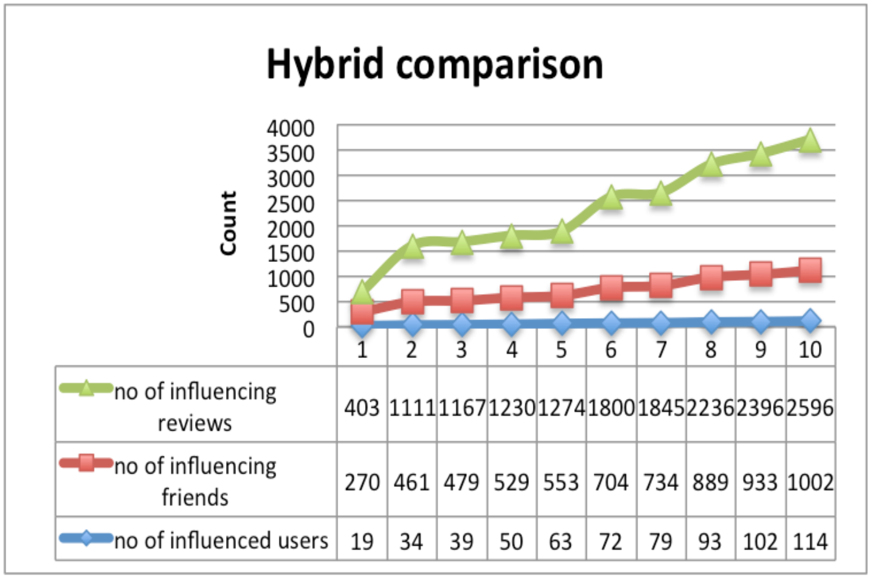
Analysis of number of influenced users vs number of influencing friends was performed next. This analysis is a combination of the analysis of the above two observations. This graph clearly depicts the number of influencing friends needed to influence 10% of the total number of users at any point of time. To perform this analysis, the count of number of influencing friends and influenced users was collected at each time frame in comparison with the reviews of the previous time frame. These counts are plotted as a graph.

The correlation coefficient between the number of influencing friends and the number of influenced users was found to be:

```
[[ 1.          0.78832749]
 [ 0.78832749  1.        ]]
```

The pearson coefficient was found to be 0.788.

Thus, we could see from all the 3 observations that, there is a very high positive correlation between the reviews of friends on dissimilar stores (dissimilar from the perspective of the user) in time t-1 and the reviews of the user himself at time t. Thus, we can safely conclude that the reviews of a user's friends at time t-1 will have a great impact on the user's reviews at time t.
As a consolidation, a hybrid comparison is now presented here which depicts the number of influenced users, the number of influencing friends and the total number of influencing reviews.



To retrieve the total number of influencing reviews, for each influencing friend of each user, we kept a track of all the reviews the influencing friend had posted with regard to dissimilar stores and retrieved its count which gives the count of total number of influencing reviews for each user. A graph was plotted which combines these 3 factors to retrieve the overall influence of friends' reviews on a user which has been presented in the graph above. It is thus clear from the graph that to influence 10% of total number of users, 10 times the number of influencing friends and enormously high number of reviews are needed.

7.4 Hypothesis 3: Elite user

The observations from the analysis conducted on elite users had showed that the distribution of users that are influenced by his social circle and the set of elite users who are not influenced by his social circle are almost same.

Influence:

We were trying to calculate how the elite users are affected by his friends who review similar business. So we took a smaller dataset of around 400 users and did the analysis. We divided the group into two classes depending on the number of friends they have. Users with friends less than or equal to 5 are grouped together and other users are grouped together.

The minimized set of ~400 users were analysed for any influence by their circles. And the results were plotted in a histogram as shown below.

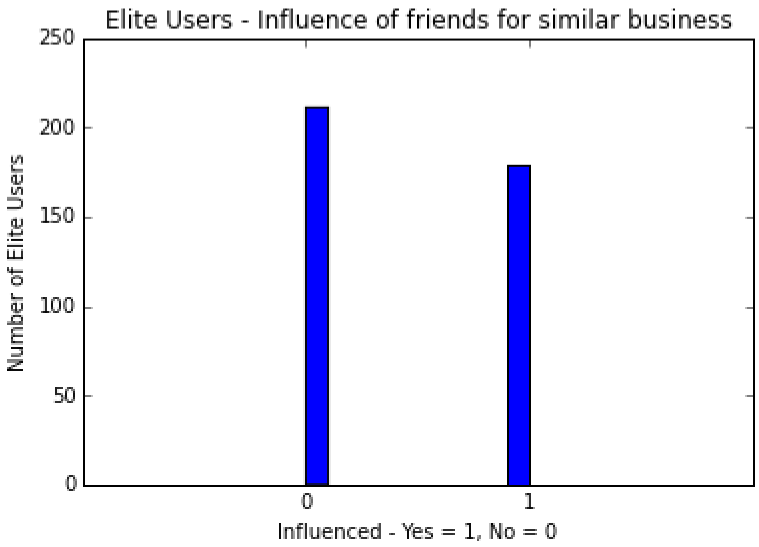


Fig: Number of users who are influenced by friends reviewing similar stores

The above figure shows that the number of elite users who are influenced by his friends who reviewed similar business at a time t-1 is almost equal to the number of elite users who are not influenced by his social circle. But we saw in hypothesis 2 that a normal user is influenced by his friends who review a dissimilar store/business rather than a similar store/business. This is a difference than what we are observing here which means that the 'elite' users are more positively affected by his friends who review similar stores.

Awareness is when a user gets aware of a business from the reviews made by his friends. Hence the user is likely to visit or review that business. But elite users are already aware of the businesses. Similarly the 'social norm' (which says that a user may contribute in order to abide to the social norm of his friends) factor is not actually true for elite users.

Correlation:

Before proceeding on to how was this established we thought of having a look at what the data has to say about the correlation between various attributes in the user dataset. Correlation between these attributes were calculated with both Pearson and Spearman techniques. The pearson method correlation is plotted as shown below.

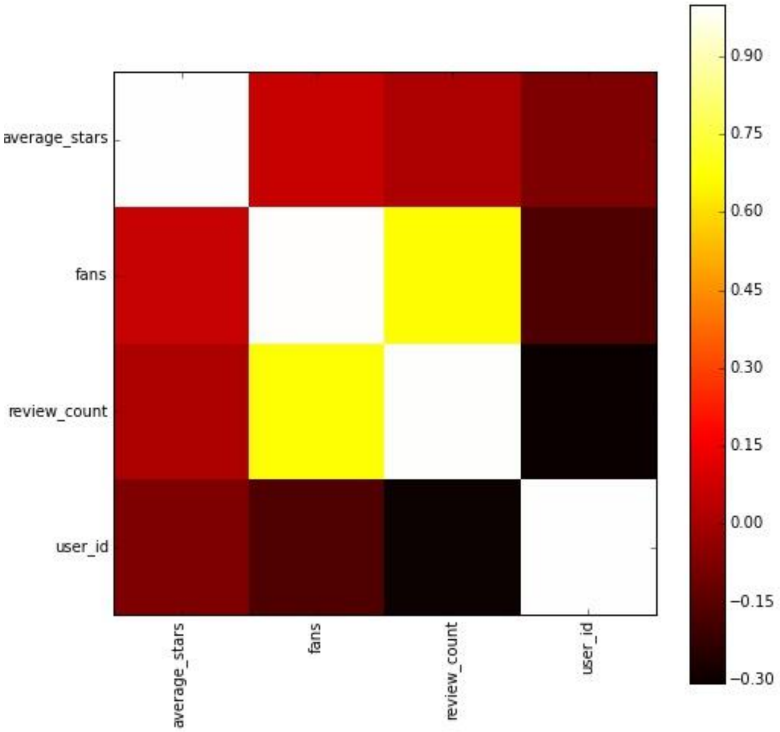


Fig: Correlation between attributes

The above figure shows about how an attribute has a correlation with other attributes. In the image, we can see that 'white' box is marked for high positive correlation and it is followed by 'yellow' for a little lesser positive correlation. So from the values above we can see that obviously each of these attributes are strongly correlated with themselves. Apart from this we can see high correlation for 'fans' and 'review_count' and a lesser correlation for 'fans' and 'average_stars'. This can be intuitively guessed to be right because having a high number of reviews means it attracts a lot of fans. But this still does not explain how the elite users' influence by his friends can be visualized. So we decided to plot a multi-dimensional graph with the information we have to get an understanding of the influence phenomena.

Multi-Dimensional Plot:

The following plot was done using python library 'matplotlib' and explains how the elite users' influence depends on other attributes.

The X-axis has the 'USERS' which is nothing but the index of our dataframe. The Y-axis has the 'REVIEW_COUNT' which is the number of reviews that a particular user has written. Two sizes are used to represent a user (Size = 45 pixels represents that the user has '>5 friends' and Size = 8 pixels represent that the user has '<= 5 friends'). Furthermore, we used two colors to represent the influence factor (GRAY - Elite user is Influenced, BROWN - Not influenced). We can see that the ‘Influenced’ elite users are those with most number of reviews and those users with lesser reviews are those who are not influenced.

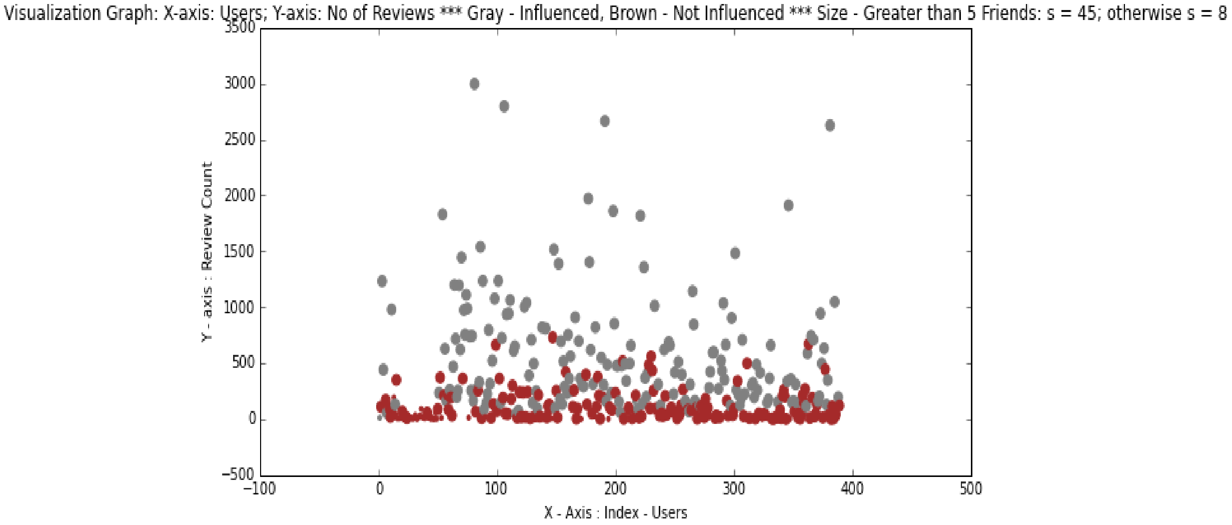


Fig: Influence on ELITE Users: No of Reviews, No of friends and Are they influenced?

We can see from the above figure that all the influenced elite users are those with most number of reviews. And all the non-influenced elite users are those at the bottom of the Y-axis with fewer reviews. Interpreting the visualization tells us that because of the large number of reviews given by certain elite users these elite users have some matching reviews with his friends, who have reviewed a similar business. Also users with 5 or lesser friends are those who are mostly not influenced by social circle that reviews similar stores. These users are identified differently from other users by having a smaller size for the representation. We can separately view these users with lesser friends alone for better visualization.

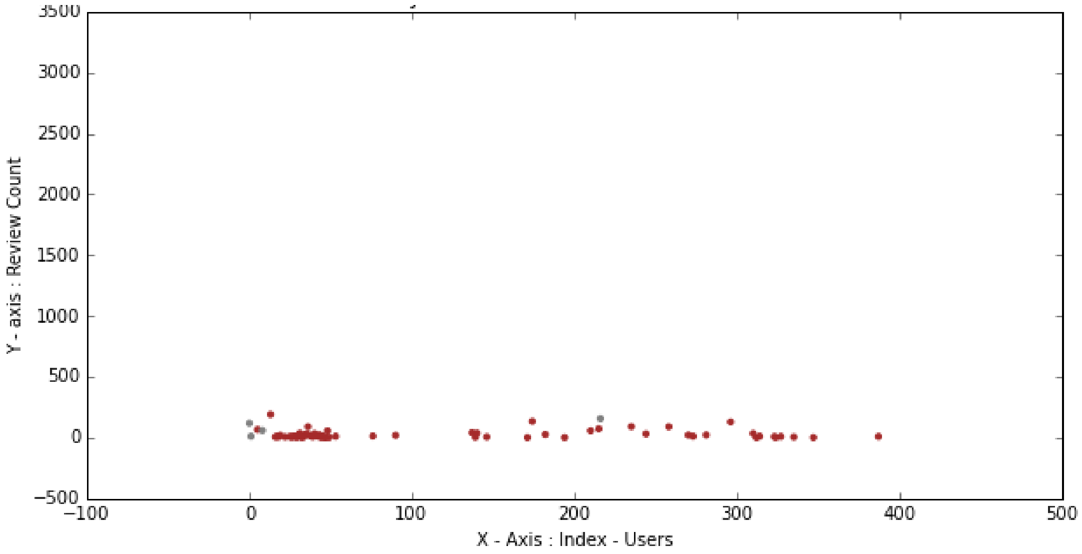


Fig: Influence on Elite Users with 5 or lesser friends

The above plot is similar to the previous plot except that we have just plotted those users with 5 or lesser friends. We can clearly see that these are the elite users that are not influenced by their social circle, possibly because of the too small social circle. So based on the 'review_count' and 'number of friends' we can derive to a conclusion that the elite users are actually independent of their opinions no matter what their social circle is or how this social circle behaves with regards to a similar or a dissimilar store.

8. Future Scope

The other aspect to the review contribution by a user can be seen as an impact of the information spreading. The information spreading probability or contagion in a social network can be studied by applying various models. Studying the strength if the ties within a social circle which may be contributing to a particular business like a restaurant/mall can be useful in understanding the scenarios under various sub-categories. It can be observed that of the total number of users at any instant of time t, 10% are influenced by their friends. Now the question arises if these 10% of the people belong to the same social circle or different social circles. The dataset given by Yelp doesn't contain information regarding users' social circles. If there could be a way to incorporate this information, it can be found if these 10% of the users are from the same social circle or not.

9. Conclusion

We have implemented few of the hypotheses mentioned in the base paper with the dataset provided by Yelp Social network. We analyzed factors like awareness, social norm and reciprocity and studied its influence on users. We identified the transition of positive correlation from medium range to weak range as the size of the social network increase beyond 400 nodes. Further analysis of reviews of users and his friends revealed that a specified percentage of users are influenced by their friends at some point of time. This project analyzes how a user is actually influenced by his friends. A user is seen to remain elite for a period of time and eventually becomes a normal user.