Recommendation of Yelp's Reviewers: Who Should You Listen to?

Luciana B. Maroun[1], Izabela K. T. Maffra[1], Thiago H. Silva[1], Pedro O. S. Vaz de Melo[1], Renato M. Assunção[1]

[1]Universidade Federal de Minas Gerais (UFMG)

{lubm,karennina,thiagohs,olmo,assuncao}@dcc.ufmg.br

Abstract. *Online ratings are nowadays one of the most trusted sources of consumer confidence when they need to decide which business to choose. In a matter of fact, reviews have become so abundant that many times the task of deciding upon whether or not to elect a business is getting increasingly difficult – each person has their own values and intrinsic characteristics that contribute to forming an opinion. In this work, we discriminate relevant users from a review website considering the perspective of a second user. This way, we are able to infer what would be the individuals that think alike certain user and recommend their reviews amongst a plethora of diverse opinions.*

1. Introduction

Consumer reviews have revolutionized the way that people choose which businesses to attend. It is now very common to turn to the web in order to make everyday decisions, such as where to eat or where to get a haircut. Yelp[1] is an opinion and experience sharing system about businesses of several kinds. A user is able to write their impressions about a certain place and rate it with a score from one to five stars. Therefore, before choosing where to go, someone can investigate what others think about different places and make a decision with a wider knowledge basis.

However, users differ in values and taste. Thus, an opinion might be useful for someone and not so much for somebody else. Knowing the profile of an individual is helpful to understand the viewpoint behind a review and decide if its adequate considering the reader's angle.

The purpose of this work is to identify important reviewers on Yelp for determined user in order to recommend reviews aligned with the readers' preferences and style. Different features might be helpful in this process, including the friendship network of reviewers. In this work, we analyze the relevance of the friendship network in users profiling as well as of a hidden friendship network — defined by users who are similar but are not connected on the social network. By encountering individuals with similar opinions, validated by the homophilly on the network, we recommend the reviews of those to the reader and reduce the burden of manually searching for relevant viewpoints.

2. Dataset Specification

The data used consists of a set of reviews, business, users and related content of Yelp's website regarding the metropolitan area of Phoenix released for a challenge[2]. There are

[1]www.yelp.com

[2]www.yelp.com/dataset_challenge

$15,585$ businesses, $70,817$ users and $335,022$ reviews.

Yelp also contains an online social network (OSN), thus users are able to be friends of each other. This network, however, is not the main purpose of the platform and does not entirely represent the true interaction between individuals — sometimes the motivation of connection is not really a friendship, but similar businesses preferences and reviewing behavior. A great part of the users, however, does not even have any friends: only $30,255$ of them are socially connected. This proves that this network is not enough for discovering reference reviewers, demanding an extrapolation with unobserved edges.

3. Network Analysis

As previously mentioned, the friendship network of Yelp is rather incomplete: the majority of users does not have any connections. The users that participate on the social network, however, are highly connected, since 40.92% of them are in the large connected component, which means that only 1.80% of the OSN participants are absent. The average degree of the whole network is 4.68 and the clustering coefficient is 0.06.

Yelp provides an option for users to vote in reviews that they found useful. Figure 1 relates the closeness centrality to the total useful votes received by a user. We observe that there are roughly three situations: users with low centrality and low usefulness; users with high centrality and low usefulness; and users with high centrality and high usefulness. There is not such a configuration in which users with high usefulness have low centrality. This is a motivation to consider the network structure as an evidence for profile of reviewers.

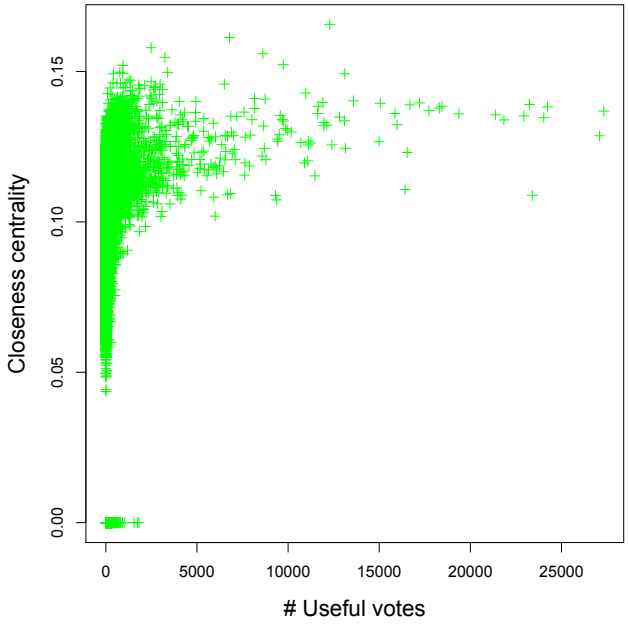


Figure 1. Correlation between closeness centrality and the number of useful votes for a reviewer.

The third set of users cited, a minority, corresponds to reviewers that influence a lot of people. However, few useful votes might indicate that those reviewers provide

important experiences for only certain kind of people. Thus, it is necessary to investigate further those reviewers in order to present the information they provide to the ones interested.

3.1. Homophilly Investigation

Aiming to validate the presence of homophilly in the network, it was conducted the following experiment: for each edge on the network, the business overlap was computed considering the jaccard similarity of reviewed establishments; a similar process was perfomed for a modified graph with the same nodes and number of edges, which were randomly assigned to pairs. The empirical cumulative distribution of the overlap values are depicted on figure 2. We observe a clear difference between the real and the random graphs — the second practically do not contain positive overlaps, while the first present some.

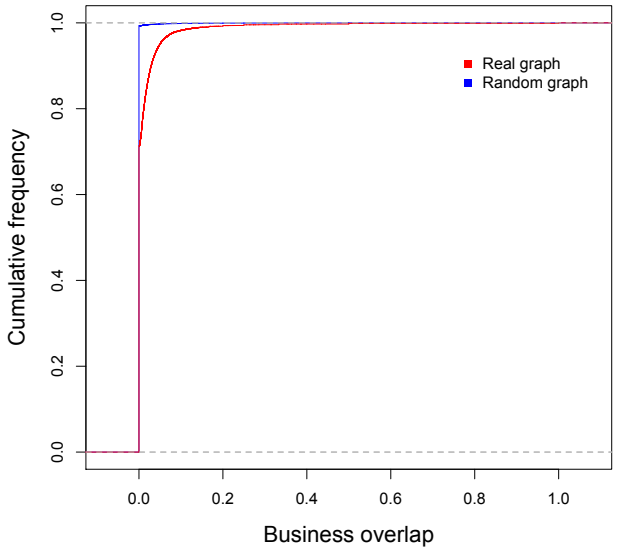
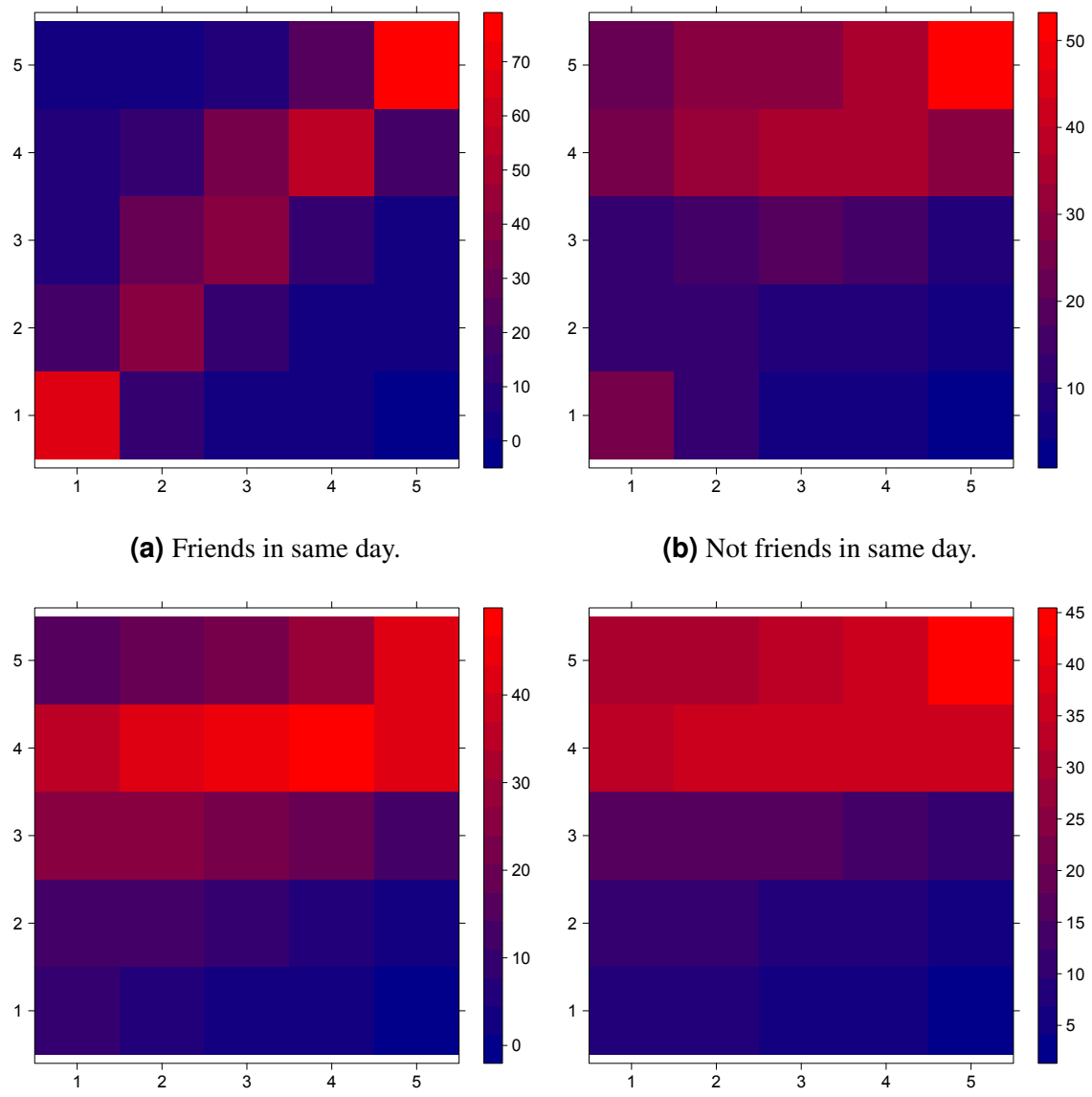


Figure 2. Empirial cumulative distribution of business overlap on real and random graphs.

3.2. Temporal Investigation

Another important aspect when considering similarity of opinions is the date of the review. People that visit a place in the same day are more likely to experience the same events and to have approximate opinions. The figure 3 contains the correlation reviews given to the same stablishment by friends in the same day 3a, not friends in the same day 3b, friends in different days 3c and not friends in different days 3d. The results are normalized — each cell represents the percentage of co-occurences for the given column. For example, in figure 3a, column 1 has around 40% in line 40%. That means that if one individual rates 1, there is a proportion of 70% of their friends who rated 1, nearly 30% who rated 2 and practically 0% who rated 3, 4 or 5.



(a) Friends in same day.

(b) Not friends in same day.

(c) Friends in different days.

(d) Not friends in different days.

Figure 3. Correlation of votes between pair of users percentual by column.

We can observe a gradative pattern from a strong correlation of ratings to almost no correlation, in which the proportion of votes on each value rules the intensity. Another interesting obervation is that there is a greater agreement between people in the same day than friends in alternative days. This may be related to the fact that people visiting certain place in the same date have something in common.

4. Reviewers Recommendation

5. Related Work

6. Conclusions and Future Work