Discovering Alpha Yelpers

Rojin Aliehyaei[§], Angkul Kongmunvattana[†], and Alexander Little[‡]

[§]School of Computational Science and Engineering, Georgia Institute of Technology
Atlanta, Georgia, 30332, USA
Rojin@gatech.edu

[†]Department of Computer Science and Information Technology, Clayton State University
Morrow, Georgia, 30260 USA
AngkulKongmunvattana@clayton.edu

[‡]School of Electrical and Computer Engineering, Georgia Institute of Technology
Atlanta, Georgia, 30332, USA
AlxLit@gmail.com

Abstract

This paper presents a novel method for discovering influential reviewers on online review websites. The proposed method relies on temporal correlations of reviews from a cluster of social network groups. These correlations are represented as graphs and the analyses are carried out through queries on graph database. For evaluation, the proposed method was applied to the dataset from Yelp Dataset Challenge, which includes 500K+ reviewers (a.k.a. yelpers), 75K+ businesses, and 2M+ reviews, yielding 12K+ trendsetting reviewers with drastically different levels of influences on Yelp. The PageRank algorithm then ranks these trendsetters with top ten ranks chosen as alphas. The evaluation results from assessing the accuracy of business rating prediction by these alphas are shown to be superior to that of the elite yelpers. The graphical user interface for utilizing the alpha's capabilities also demonstrated that rating and review from alpha yelpers can be used to reduce the length of time for locating good businesses.

1. Introduction

Contemporary consumers rely on reviews and ratings of businesses to select service providers. A large amount of reviews on popular businesses and a lack of robust reviews on new businesses can equally baffle these consumers. In this work, we define trendsetters as reviewers with influences in their social network groups. These trendsetters are different from elites. Yelp designated elites using the numbers of written reviews, friends, and compliments, among other factors, whereas the trendsetters are discovered through their level of influences on social network groups.

Identified trendsetting reviewers can be used to pinpoint ratings of popular businesses without the need of sifting through all reviews and to predict ratings of new businesses that have garnered only a few reviews. In a nutshell, this work aims to answer the following question: Can we identify trendsetters effectively and use them to predict ratings of businesses accurately?

To identify trendsetters, we captured the level of influences among reviewers in their social network groups by temporally correlating reviews from yelpers and their friends, depicting them as graphs. We then identified the top trendsetters (i.e., the alphas) by applying PageRank algorithm [3, 8, 10] on these graphs. Experimental results not only showed significant data points supporting the existence of influences in social network groups, but also yielded suitable graphs for the PageRank algorithm to identify the alphas effectively. Specifically, our experiments on a dataset from the Yelp Dataset Challenge yielded 12,815 trendsetters with drastically different levels of influences. The top-ranked trendsetters based on PageRank's results are chosen as alphas. The ratings posted by these alphas are then used for predicting business ratings.

To evaluate the accuracy of business rating prediction by the alphas, we compared a rating posted by the highest rank alpha for a particular business against the Yelp's stars rating for that business, which is an average of business ratings posted by all reviewers. The results showed the ratings by alphas are within half-a-star of the Yelp's stars ratings for businesses on average. When comparing ratings posted by Yelp's elites against the Yelp's stars rating, the results showed the ratings by elites are with in three-quarter of a star of the Yelp's stars ratings for businesses on average. It is evident from these results that the alphas predicted business ratings with higher accuracy than the elites.

To utilize the alpha's capability, an interactive user interface providing recommendations for service providers based on business ratings given by top-ranked alphas was developed and a user survey was carried out. The results from the user survey demonstrated that alpha's rating and review can be used to reduce the length of time for finding good restaurants. The user interface design also received a

high mark for ease of use in the survey. In summary, the key innovations in this work are listed as follows:

• An introduction of influential relationships based on temporally correlated reviews between yelpers as a feature for identifying trendsetting characteristic and social network clout.

• An application of PageRank algorithm to identify trendsetters on Yelp.

• An exploitation of trendsetters for determining business rating on Yelp instead of relying on the aggregate rating from a large group of reviewers.

The rest of this paper is organized as follows. Section 2 provides a survey of related work. Section 3 describes the proposed methods. Section 4 presents and discusses the results. Finally, Section 5 summarizes the contributions made in this work.

2. Related Work

A literature survey yielded prior work related to identifying: (a) the characteristics of extraordinary yelpers, such as elites and local experts, (b) the influences of reviews on consumers and businesses, and (c) the review contents and their correlations with ratings. Specifically, Crain et al. [5] and Tucker [13] performed graph and data analyses with the results indicating that elites have accumulated more friends, wrote more reviews, and received more compliments on their reviews as being useful, cool, or funny. Jindal's master thesis adopted classification algorithms to identify local experts, which constitute a subset of Elites [9]. None of these studies went beyond using the existing features in the given dataset.

In this study, we created a new feature by temporally correlating the reviews written by friends to identify the influences of an individual yelper on other yelpers, which is a one-to-many relationship. The proposed feature is a better indicator on the influences of an individual yelper than the number of friends. This approach is well-supported by previous studies on twitter users, where they found the number of followers is not as good an indicator of influences as the number of retweets [2, 4].

In other works, Alluri's master thesis used regression-based models for predicting the effect of peer pressures (i.e., influences) from friends on an individual yelper, which is a many-to-one relationship [1]. Our work studies the opposite side of the influences, which is more beneficial to other users because it harvests peer knowledge to identify trendsetters and then uses their expertise to assist the crowds in making decisions.

Luca [12] and Hood et al. [7] studied the influences of posted reviews on future businesses using regression-based models. Zhao et al. used matrix factorization algorithm to predict future ratings using past rating behaviors [16], which is limited to individual users. These studies correlated posted reviews with future reviews or future revenues whereas our study focused on correlations of posted reviews within the social network groups, directly benefiting all users instead of businesses.

On review contents and ratings correlation, Lim and Van Der Heide studied a correlation between positive tones in posted reviews and their credibility through statistical analyses [11]. Vinson and Dale found that reviews with high entropies often led to extreme ratings of very good or very bad [14]. In this work, we captured the extremely good ratings in reviews posted by the discovered alphas to recommend restaurants to users.

Finally, Cui performed data analyses on the existing features of dataset from previous round of Yelp Dataset Challenge (YDC) using a graph database called Neo4j with results presenting various insights about that dataset [6]. We adopted the same database framework (Neo4j) to store the dataset from the current round of YDC, but our new implementation for dataset imports into Neo4j significantly decreased the execution time for populating database from several days to just a few minutes. We also developed new algorithms for different graph analyses needed for feature engineering and ranking algorithms in this work.

3. Proposed Methods

The intuition of this work is based on the notion of influence. The current online review websites attracted users to service providers with good average ratings from all reviews. To make an informed decision, users often went further and spent significant amount of time to read or skim through many posted reviews (or disappointedly found that there are very few posted reviews to make an informed decision). This work aims to find influential reviewers so called alphas and utilizes them to provide recommendation for good service providers, obviating the need for reading multiple reviews. The proposed approach ranks alphas based on their level of influences in their social network groups. The level of influences is discovered through temporal correlation of reviews written by reviewers in the same social network group. This is a better approach for measuring influences than the current approach used in designating elites that simply counts the number of friends and other attributes of the reviewers. The proposed method is as follows.

The first step of our work is creating a database from the dataset made available for the 7th Round of Yelp Dataset Challenge. The given dataset was in the JSON format consisting of five different files, namely business, check-in, review, tip, and user. A python script was developed for converting these JSON files into CSV files, namely business, check-in, friends, review, tip, and user. The "friends" attribute in a JSON file called user was extracted to create a CSV file called friends. The "date" attribute in a JSON file called review was also translated into the UNIX time when a CSV file called review was created. This date-

and-time conversion aims to facilitate the temporal correlation extraction in the feature engineering phase. Once the conversion was completed, we developed a shell script using import and shell tools to import these CSV files into a graph database called Neo4j, populating it within a few minutes. We chose Neo4j for this work because most of the proposed work is related to graph analyses.

Once the graph database of yelp dataset is ready, the next step is feature engineering to extract reviewers that have written influential reviews. The influence is measured based on the number of reviews for each business that were posted by friends within a two-week timeframe starting when a review was first written by a reviewer, who is connected with them in the social network groups. A pseudo code and a cypher (Neo4j's SQL-inspired query language) for extracting a pair of influencer and follower based on review's timestamps and friendship using MATCH clause are shown in Figures 1 and 2, respectively.

This cypher is embedded into a python script that also performs the graph analyses using the PageRank algorithm. The intuition and past work on twitter analyses suggested that measurable follow-up actions (such as re-tweets) within social network groups is a better indicator for level of influences than a static headcount of followers. Our proposed approach is based on this intuition, measuring follow-up actions though reviews posted by friends on the same business within the two-week timeframe, which is a better indicator of influences than a static headcount of friends. The discovery of trendsetters is carried out through graph analyses using the PageRank algorithm. Our implementation for this step also used python with a call to PageRank method from NetworkX package. The top-ranked trendsetters from the results are chosen as alphas. The last analysis step used the ratings posted by these alphas to predict business ratings.

The accuracy of business rating prediction by alphas is evaluated by comparing business rating given by the highest rank alpha against the Yelp's stars rating for that business, which is an average rating of all posted reviews. The accuracy of alpha's ratings is compared against the accuracy of ratings by Yelp's elites for the same business. The basis for this evaluation lies in the goal of this work, which aims to replace the needs for reading all reviews (or many reviews by elites) in order to judge a business with only one rating and review given by the highest rank alpha. This evaluation was also implemented in python.

To benefit from alpha's capability in identifying businesses with good ratings, an interactive user interface providing information about restaurants that have received five-star rating by the top ranking alphas (dubbed Alpha Selects) was developed. Figure 3 depicts a screenshot of an initial webpage for the city of Las Vegas in Nevada with a default selection of American restaurants. When a particular restaurant is selected, a rating and review from alpha is displayed on the left panel and pertinent information, such as street address, business hours, and price ranges, are provided on the right panel. Figure 4 shows a screenshot when Egg & I restaurant is chosen. This interactive user interface aims to enhance user experiences when searching good restaurants by providing recommendations from influential reviewers (i.e., the alphas) and simplifying selection process. The interface was developed in JavaScript with d3, jquery, js-image-slider, gantt-chart-d3-customized, and gmaps libraries. This user interface was evaluated using a within-subjects design with two conditions: (1) using Alpha Selects and (2) using Yelp where users attempting to select a restaurant in a particular city. A questionnaire for user survey is given in the Appendix.

4. Evaluation

The evaluation was carried out using a dataset from the 7th Round of the Yelp Dataset Challenge [15]. This dataset includes 552,339 reviewers, 77,445 businesses, and 2,225,213 reviews in JSON format. After converting the given JSON files into CSV files using our python script, there are two common approaches for creating a graph database on Neo4j from a dataset. The first approach is directly populating records into a graph database in neo4j through a python script with py2neo using CREATE clause. The second approach uses Neo4j's import and shell tools. Both approaches were evaluated on a laptop with eight 2.8 GHz Intel Core i7 and 16 GB of RAM. The first approach took a much longer time to complete (i.e., two weeks) whereas the later took about five minutes.

Once a graph database is created, a few simple cyphers shown in Figure 6 were used to query the database to verify that the dataset import was carried out correctly. Apart from verifying the number of reviewers, businesses, and reviews, the results also revealed that social network groups of these reviewers represented 3,563,817 friendships. Figure 6 depicted a subset of these friendships. Figure 7 depicted a subset of reviews, which are relationships between reviewers and businesses.

To detect influences between reviewers in their social network groups, a cypher shown in Figure 8 was executed to create a new set of relationships called *influenced-by*. It is similar to a cypher in Figure 2 where the output was a set of tuples representing influential relationships. Figure 9 showed a subset of this influential relationships. A cypher shown in Figure 10 was used to count the total number of influential relationships in the dataset, which was 37,489.

To measure and rank the level of influences in social network groups of reviewers, a set of tuples representing influential relationships (generated by a cypher shown earlier in Figure 2) was fed as an input to the PageRank algorithm. The results in Table 1 showed the top ten ranked influential reviewers (dubbed alphas) from this dataset. These alphas were resided primarily in the city of Las Vegas (NV) and most of them have a high number of influential relationships.

FormInfluenceGraph

(

Inputs:

n = Number of users,
b = Number of businesses,
uid_i = i^{th} user id ,
bid_y = y^{th} business id,
Friend-List = List of uids that are friend with i^{th} user,
Review-List = List of bids that are reviewed with i^{th} user

Outputs:

Graph-of- Influence=$[g_{ij}]_{nxn}$

)

{

Friendship_Matrix=$[f_{ij}]_{nxn}$
IF $uid_i \in$ Friend-List$_j$: $f_{ij=1}$
IF $uid_i \notin$ Friend-List$_j$: $f_{ij=0}$

Review_Matrix=$[r_{iy}]_{nxb}$
IF $bid_y \in$ Review-List$_i$: $r_{iy=1}$
IF $bid_y \notin$ Review-List$_i$: $r_{iy=0}$

Date_of_Review =$[d_{iy}]_{nxb}$
IF $r_{iy==1}$: d_{iy} = Date of i^{th} user's review for business y

For all (uid_i, uid_j) , i≠j:
IF ((f_{ij} ==1) AND ($\exists\ y$ s.t.: ((r_{iy} ==1) AND ((r_{jy} ==1) AND ($d_{jy} - d_{iy} \in$ (0,2weeks]):
$g_{ij=1}$
else:
$g_{ij=0}$

}

Figure 1: A cypher for extracting influential relationships from a graph database

```
MATCH (u1)-[:friends_with]->(u2) WITH u1, u2 ORDER BY id(u1) DESC
MATCH (u1)-[r1:reviewed]->(b:Business)<-[r2:reviewed]-(u2)
WHERE (r1.date < r2.date) AND (r2.date < (r1.date + 14*24*60*60))
RETURN u2.user_id AS uid2, u1.user_id AS uid1;
```

Figure 2: A cypher for extracting influential relationships from a graph database

To evaluate alpha's capability in predicting business's rating, all average business ratings were compared with the ratings given by the top ranked alphas. Table 2 depicted the inaccuracy of rating predictions. In term of Yelp's five-star rating, the results translated to about half-a-star of inaccuracy. In comparison, elites produced a higher rate of inaccuracy when predicting business ratings at about three-quarter of a star on average.

An interactive user interface was designed and developed to exploit of alpha's capability in term of pinpointing business ratings, aiming to reduce the time needed for users to find businesses with good ratings. A user survey was carried out with nine participants using the questionnaire shown in the Appendix. Every participant was told to find one restaurant that he/she would like to visit using our interface (called Alpha Selects) and then Yelp. Neither tutorial nor guidance on Alpha Selects was provided to the participants.

The results are shown in Figures 11 and 12. Specifically, subjective ratings of Alpha Selects in Figure 11 demonstrated that participants rated our interface favorably especially on ease of use and in reducing the length of time for finding good restaurants. On impressions of Alpha Selects, participants also rated our interface favorably when comparing with Yelp (see Figure 12). Specifically, they feel Alpha Selects is easier to learn, easier to find good reviews, and more helpful in finding new restaurants.

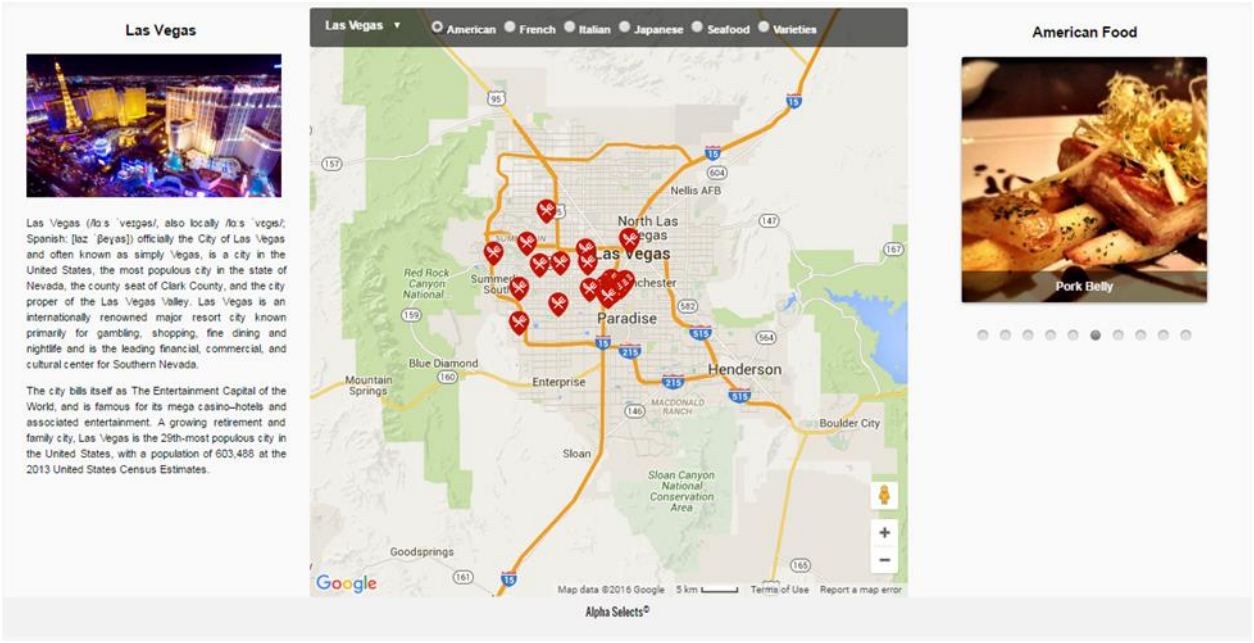

Figure 3: A screenshot of Alpha Selects showing an initial webpage for the City of Las Vegas

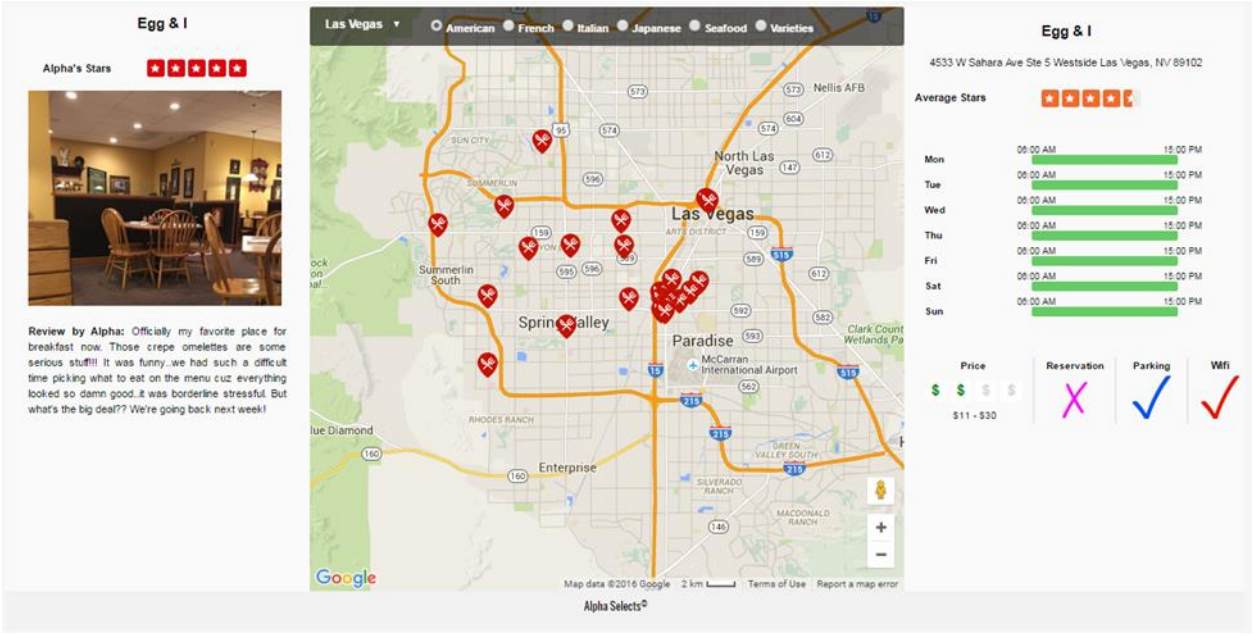

Figure 4: A screenshot displaying information of a selected restaurant called Egg & I

```
MATCH (n:User) RETURN count(n);
MATCH (n:Business) RETURN count(n);
MATCH ()-[r:reviewed]->() RETURN count(r);
MATCH ()-[r:friends_with]->() RETURN count(r);
```

Figure 5: Cyphers for extracting the number of reviewers, businesses, reviews, and friendships

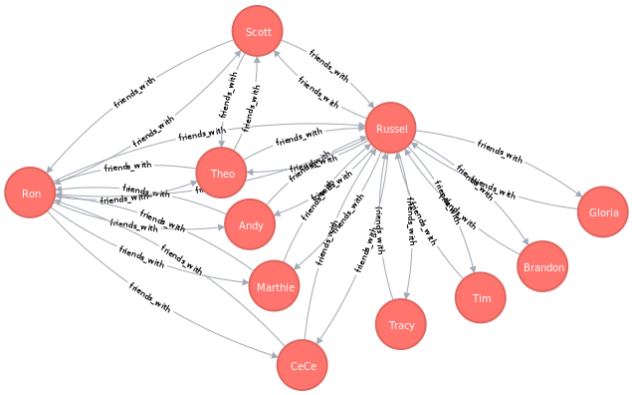

Figure 6: A subset of friendships in social network groups from Yelp's dataset

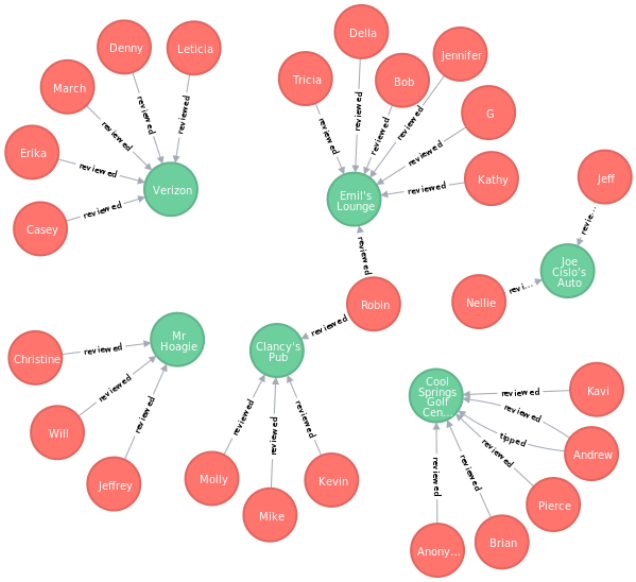

Figure 7: A subset of relationships between reviwers and businesses from Yelp's dataset

```
MATCH (u1)-[:friends_with]->(u2) WITH u1, u2 ORDER BY id(u1) DESC
MATCH (u1)-[r1:reviewed]->(b:Business)<-[r2:reviewed]-(u2)
WHERE (r1.date < r2.date) AND (r2.date < (r1.date + 14*24*60*60))
CREATE UNIQUE (u2)-[:influencedby]->(u1);
```

Figure 8: A cypher for creating a set of directed graphs representing influential relationships

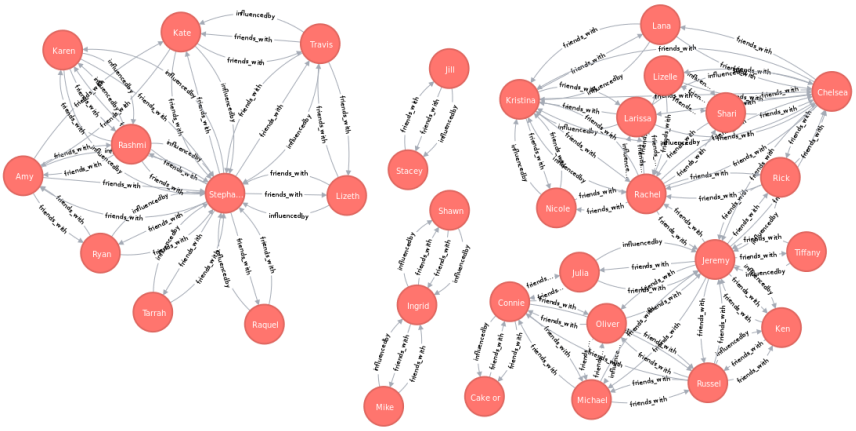

Figure 9: A subset of influential relationships from our feature engineering

```
MATCH ()-[r:influenceby]->() RETURN count(r);
```

Figure 10: A cypher for counting the total number of influential relationships

5. Conclusions

The proposed method for detecting influences in the social network groups through temporally correlated reviews successfully identified influential yelpers in the Yelp Challenge Dataset. The PageRank algorithm was also successfully used to pinpoint level of influences, generating a ranking of trendsetting yelpers so called alphas. The evaluation demonstrated that the top ranked alphas provided highly accurate ratings for businesses (i.e., within half-a-star in Yelp's five stars rating system). A user survey on the interactive user interface called Alpha Selects also confirmed that ratings and reviews from top ranked alphas helped reducing the time needed for selecting a service provider.

References

[1] A. V. Alluri. Empirical study on key attributes of yelp dataset which account for susceptibility of a user to social influence. MS Thesis, University of Cincinnati, Ohio, 2015.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*, pages 65-74, 2011.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertext Web search engine. In *Proceedings of the seventh International Conference on World Wide Web (WWW7)*, pages 107-117, 1998.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: the million follower fallacy. In *Proceedings of the fourth AAAI International Conference on Weblogs and Social Media*, pages 10-17, 2010.

[5] K. Crain, K. Heh, and J. Winston. An analysis of the "elite" users on yelp.com. CS224W Term Project Report, Stanford University, 2014.

[6] Y. Cui. An evaluation of yelp dataset. arXiv preprint arXiv:1512.06915, 2015.

[7] B. Hood, V. Hwang, and J. King. Inferring future business attention. Retrieved from https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_InferringFuture.pdf

[8] F. Jiang, Y. Yang, S. Jin, and J. Xu. Fast search to detect communities by truncated inverse PageRank in social networks. In *Proceedings of the fourth IEEE International Conference on Mobile Services (MS)*, pages 239-246, 2015.

[9] T. Jindal. Finding local experts from yelp dataset. MS Thesis, University of Illinois at Urbana-Champaign, 2015.

[10] F. Lamberti, A. Sanna, and C. Demartini. A relation-based Page Rank algorithm for semantic Web search engines. *IEEE Transactions on Knowledge and Data Engineering*, 21(1):123-136, 2009.

[11] Y.-S. Lim and B. Van Der Heide. Evaluating the wisdom of strangers: the perceived credibility of online consumer reviews on yelp. *Journal of Computer-Mediated Communication*, 20:67-82, 2015.

[12] M. Luca. Reviews, reputation, and revenue: the case of yelp.com. Working Paper 12-016, Harvard Business School, 2011.

[13] T. Tucker. Online word of mouth: characteristics of yelp.com reviews. *The Elon Journal of Undergraduate Research in Communications*, 2(1):37-42, 2011.

[14] D. W. Vinson and R. Dale. Valence constrains the information density of messages. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1682-1687, 2014.

[15] Yelp Dataset Challenge. https://www.yelp.com/dataset_challenge

[16] G. Zhao, X. Qian, and X. Xie. User-service rating prediction by exploring social users' rating behaviors. *IEEE Transactions on Multimedia*, 18(3):496-506, 2016.

Table 1: Top ten ranked alphas produced by PageRank algorithm

Alpha Rank	Yelp's User ID	City	PageRank Score	Influential Relationships
0	WmAyExqSWoiYZ5XEqpk_Uw	Las Vegas	0.003883813	363
1	nEYPahVwXGD2Pjvgkm7QqQ	Pittsburgh	0.003500185	277
2	fczQCSmaWF78toLEmb0Zsw	Scottsdale	0.003186067	379
3	4ozupHULqGyO42s3zNUzOQ	Scottsdale	0.002938267	310
4	ia1nTRAQEaFWv0cwADeK7g	Las Vegas	0.002869387	466
5	9A2-wSoBUxlMd3LwmlGrrQ	Las Vegas	0.002461253	292
6	CvMVd31cnTfzMUsHDXm4zQ	Charlotte	0.002051691	188
7	OaFcpi3W4AwxrD8W2pgC_A	Las Vegas	0.001752581	250
8	9OZH1Ecw-qUkCW5MS0NefA	Pittsburgh	0.001718194	90
9	Iu3Jo9ROp2IWC9FwtWOaUQ	Las Vegas	0.001696339	358

Table 2: Rating Prediction Inaccuracies of Alphas and Elites

Alpha Rank	Yelp's User ID	Alpha Rating Inaccuracy	Elites Rating Inaccuracy
0	WmAyExqSWoiYZ5XEqpk_Uw	0.136920530	0.152127151
1	nEYPahVwXGD2Pjvgkm7QqQ	0.124032258	0.135011023
2	fczQCSmaWF78toLEmb0Zsw	0.110380267	0.140248510
3	4ozupHULqGyO42s3zNUzOQ	0.114994097	0.141271246
4	ia1nTRAQEaFWv0cwADeK7g	0.132208029	0.147611482
5	9A2-wSoBUxlMd3LwmlGrrQ	0.160756677	0.150253862
6	CvMVd31cnTfzMUsHDXm4zQ	0.107537688	0.125438700
7	OaFcpi3W4AwxrD8W2pgC_A	0.121869783	0.144971569
8	9OZH1Ecw-qUkCW5MS0NefA	0.113230769	0.133794873
9	Iu3Jo9ROp2IWC9FwtWOaUQ	0.112790698	0.146845623

Figure 11: Subjective ratings of Alpha Selects

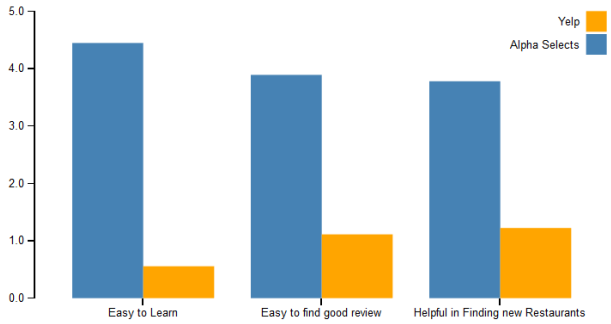

Figure 12: Impressions about Alpha Selects

APPENDIX - Questionnaires for User Survey of Alpha Selects

Questions	Strongly Agree (5)	4	3	2	Strongly Disagree (1)
Alpha Selects was easy to use.					
I was confused when I searched for restaurants using Alpha Selects.					
The results returned by Alpha Selects were satisfying.					
Descriptions and details given for each restaurant were useful.					
Finding a good review using Alpha Selects seemed easier than Yelp.					
Learning to use Alpha Selects seemed easier than Yelp.					
A review of individual restaurant by Alpha Yelper was helpful.					
Alpha Selects seemed more helpful in finding new restaurants than Yelp.					
Alpha Selects reduces the length of time for finding good restaurants.					
Alpha Selects improved my effectiveness in finding restaurants.					
Alpha Selects reduces the length of time for finding good reviews.					
I would like to use Alpha Selects in the future.					