Impact of Social Networks on Buying Behavior: Predicting the Success of Pittsburgh Businesses Through Analysis of Yelp Social Networks

Submitted by: Apoorva Havanur

Undergraduate Economics Program
Tepper School of Business
Carnegie Mellon University

In fulfillment of the requirement for the
Tepper School of Business Senior Honors Thesis in Economics

Advisor:

Professor Maryam Saeedi
Assistant Professor of Economics
Tepper School of Business
Carnegie Mellon University

May 2018

Impact of Social Networks on Buying Behavior

Predicting the Success of Pittsburgh Businesses Through Analysis of Yelp Social Networks

Apoorva Havanur
B.S Statistics and Machine Learning, additional major in Economics
Carnegie Mellon University
Pittsburgh, PA
E-mail: ahavanur@andrew.cmu.edu

Abstract

In this paper, we analyze data collected from Yelp to understand the importance of the social networks created between Yelp reviewers, and the impact it has on the businesses they patronize. We then look at the shape of the social network generated by reviewers and identify differences between the behavior of 'elite' users vs. non-elite users. We then view the trends in network formation for particular businesses over time, and use features of the network of the business early in its development in order to predict its future success. We construct linear regression and random forest models that solely use features derived from review data, as well as models that are built using a combination of review and social network features. We see that the additional network features are statistically significant, and help reduce the root mean squared error of our models by a significant percentage. Ultimately, using our network features of reviewers from the first three months of business, we can predict the number of reviewers for a business within its first year with an error of less than 2.5 reviewers, an error of 8.5 over two years, and 13 over three.

Keywords: Yelp, Reviews, Social Networks, Business

Introduction

Historically, word of mouth marketing has been the lifeblood of new and upcoming businesses. A customer who has experienced good service and a high quality product at an establishment is likely to share that experience with their friends, family, and coworkers who then in turn go visit the same place. Today, the sharing of these experiences largely happens online via social media websites. The information related to the experience then disseminates through the person's social network. Yelp.com is a popular crowd-sourced local business review and social networking site that allows users to review local businesses and connect with other reviewers on the website. Using a combination of data collected on reviews and friendships between users made available by the Yelp Open Dataset Challenge, this paper looks in depth into the effect that the Yelp social networks have into Pittsburgh businesses' long term success.

This paper first gives an overview of the functionality of Yelp and terminology associated with its use, as well as an overview of social networks and their application. We then explore related works to this topic to establish known trends in social network structures, as well as quantitative measures of influence between two nodes in a social network given a series of actions.

We then move into exploring the data in more detail, and provide a series of summary statistics describing the user network consisting of over 16,000 nodes and over 120,000 friendships. We see that the attributes of the Yelp review network are similar to findings made about other social networks in terms of composition, with a vast majority of nodes in a single large connected component, and the rest divided into singleton components of 1 or 2 nodes and a middle region of clusters of 5 or 6 nodes. We explore the implications of various distributions of user attributes, such as average number of reviews made, the average rating given, the average number of friends, and the average number of businesses each pair of reviewers have mutually reviewed, and see that most follow a power law distribution, with the exception of ratings, which are distributed normally and centered around 4.0. We also look at distinctions between the subgraph comprised of “Yelp elite“ users and non-elite users, and see that the graph comprised of elite users is much more centralized, with a larger proportion within the giant component and the clustering coefficient of the elite graph (0.33) is significantly higher than for non-elites (0.06). However, despite the elite status of these users, the average influence they exert on their network is almost identical to the influence of non-elites (0.02 vs 0.01).

We also look at the formation of networks of reviewers of a specific business over time. Over time, as the number of reviewers of a business increases, the graph grows in both size and complexity. We show that over time proportional values such as proportion of Yelp elite reviewers and proportion of nodes in the largest component start out quite high and then start to decline sharply until flattening out.

We finish the paper by exploring different modelling techniques for predicting the future success of a business. Using number of reviews as a proxy for overall number of guests, we take data for each business at one month and three months after their opening and attempt to predict the number of reviews they’ll get after one, two, and three years. In our first series of linear regression models, we only use features that could be derived from the reviews themselves, like percentage of elite users, average rating, number of reviews thus far, and the average difference between each user’s rating given to this particular business and the average rating they normally give. We then compare the root mean squared error of these models to models that use a combination of these features and network based features, such as graph density, clustering coefficient, influence, etc. We see that the network feature models performed better the further into the future we attempted to predict, and that an ANOVA test validates the significance of the features. We also find that using 3 months as a starting point as compared to one month gives significantly better results for almost all models except for random forest models, in which the one month trained models performed better.

Background

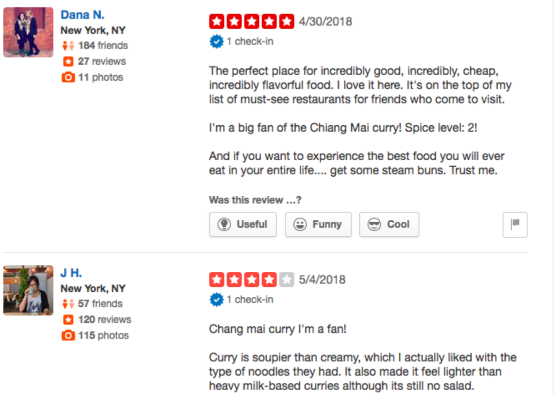
Yelp Overview

Yelp is a social media website that provides crowd sourced reviews of local businesses. Founded in 2004, the online review service now gets more than 135 million monthly visitors and more than 95 million reviews.

Users create accounts on Yelp in order to post reviews, “check in“ at places they’ve been, upload photos, and share/save particular businesses. Users who have contributed often are invited to become “Yelp Elite“ members, which gets displayed with their account information and provides them access to special perks and rewards through Yelp.

Each review on Yelp comes with a rating (out of 5 stars), and a short written description. Other users can then react to a rating, calling the review “Funny“, “Useful“, or “Cool“. Reviews are generally displayed with the most recent reviews coming first, and primarily in English. Figure 1 shows the format that these reviews come in as of 2018, in which site viewers see a feed of reviews for a particular business whose page they are on.

A lot of other information about a business is displayed on Yelp. In addition to reviews, users can look at store hours, menus, locations/directions, store policies (such as BYOB) and any special deals that the business may have at the time. In addition, reviewers can provide their input on a variety of other criteria, such as “good for kids“ or “good for large groups“, etc.



Dana N.
New York, NY
184 friends
27 reviews
11 photos

4/30/2018
1 check-in

The perfect place for incredibly good, incredibly, cheap, incredibly flavorful food. I love it here. It's on the top of my list of must-see restaurants for friends who come to visit.

I'm a big fan of the Chiang Mai curry! Spice level: 2!

And if you want to experience the best food you will ever eat in your entire life.... get some steam buns. Trust me.

J H.
New York, NY
57 friends
120 reviews
115 photos

5/4/2018
1 check-in

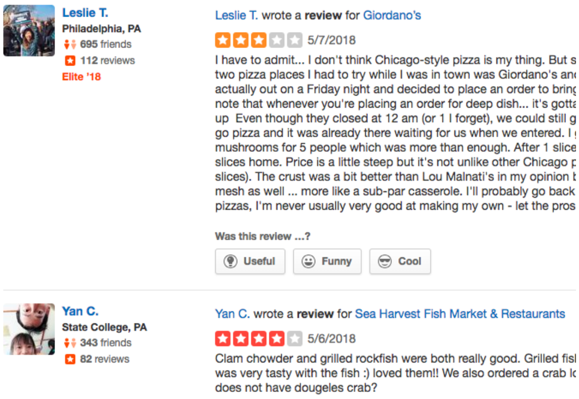
Chang mai curry I'm a fan!

Curry is soupier than creamy, which I actually liked with the type of noodles they had. It also made it feel lighter than heavy milk-based curries although its still no salad.

Fig. 1. Example of a Yelp Review Feed

A visitor to the site does not need have an account in order to view the reviews of a business; however, to contribute, they must make an account. Users can also log in using credentials from other websites, such as Facebook or Google.

In addition to their own account, users also have the option of adding friends to their profile. Being friends with a user on Yelp allows one to see recent reviews made by people in their network, see the reviews their friends made when viewing a particular business and message them directly. Figure 2 shows a friends-only feed of reviews on Yelp, in which the user can see all the posts made by their friends in chronological order.



Leslie T.
Philadelphia, PA
695 friends
112 reviews
Elite '18

Leslie T. wrote a **review** for Giordano's
★★★☆☆ 5/7/2018
I have to admit... I don't think Chicago-style pizza is my thing. But s
two pizza places I had to try while I was in town was Giordano's an
actually out on a Friday night and decided to place an order to brin
note that whenever you're placing an order for deep dish... it's gott
up Even though they closed at 12 am (or 1 I forget), we could still g
go pizza and it was already there waiting for us when we entered. I
mushrooms for 5 people which was more than enough. After 1 slice
slices home. Price is a little steep but it's not unlike other Chicago
slices). The crust was a bit better than Lou Malnati's in my opinion
mesh as well ... more like a sub-par casserole. I'll probably go back
pizzas, I'm never usually very good at making my own - let the pros

Yan C.
State College, PA
343 friends
82 reviews

Yan C. wrote a **review** for Sea Harvest Fish Market & Restaurants
★★★★☆ 5/6/2018
Clam chowder and grilled rockfish were both really good. Grilled fis
was very tasty with the fish :) loved them!! We also ordered a crab l
does not have dougeles crab?

Fig. 2. A Yelp feed showing reviews made by friends, in chronological order

Since it's inception, Yelp has had a significant impact on the livelihood of local businesses. Analysis done by the Harvard Business Review has suggested that each "star" in a Yelp rating affected the business owner's sales by 59 percent [1]. However, the company is not without its share of controversy. Yelp has continuously had to deal with fake reviews, and estimated that anywhere between 5-20% percent of their reviews are inaccurate [2]. Businesses also have complained about the fact that reviews cannot be verified, which opens them up to negative reviews that significantly affect their business. Nevertheless, Yelp continues to be a major driving force for many local businesses, and as Michael Luca of Harvard Business School describes it, has "replaced traditional forms of reputation" [3].

Social Networks Overview

The study of social networks has been ongoing since the 1890s as an interdisciplinary field between economics, psychology, sociology, statistics, computer science, and more. In short, social networks attempt to model social interactions between individuals. The graphical representation of a social network involves displaying individuals as nodes, and a friendship between them as an edge connecting those two nodes. Applications of social networks have been seen in fields as diverse as biology, linguistics, organizational behavior, and more.

The rise of social media such as Twitter, Facebook, Instagram, etc has given increasing importance to the study and understanding of social networks and their effect on behavior. The aim of studying a social network is to understand how the relations between agents affect their behavior, more so than the agents themselves.

An important field of study within social network research is the study of network formation. Over time, the addition and removal of agents in a network affects its shape and structure. Different network structures yield different properties, such as the rate at which information spreads through the network. In our case, we are particularly interested in the idea of the *influence* that a person has on their friends. Being able to identify influential users would be a key asset to businesses, who can target those users specifically in the hopes that their review drives new customers to their store. By attracting these users earlier then, one would expect to see a change in the shape of the social network that forms from the reviewers of that business versus a business without such influential reviewers.

Related work

Ample literature exists regarding both social networks and Yelp reviews.

The impact of Yelp on businesses has been studied extensively by scholars since its inception, such as the extensive analyses done by Jamie Doward and Michael Luca [1,3]. These papers however, tend to look at the contents of the Yelp reviews themselves, and mostly in isolation, but does establish the importance of ratings and reviews on the overall.

Social network research is a relatively old field but recent advances have helped to shape the type of questions one can ask. Subramani and Rajagopalan [4] provided a framework of how to think about viral marketing in the contexts of online social networks, arguing for two types of influence: normative influence and informational influence, as well as arguing that for viral marketing to be successful ”success hinges upon the recognition of the strong need for influence to be viewed as knowledgeable helpers in the social network rather than as agents of a marketer”.

Kumar et al [5] go in depth into online social networks on Flickr and Yahoo! and describe the shape and structure of these online networks, including the composition of these networks as a combination of numerous singletons, one large giant component, and a middle region consisting primarily of star shaped components.

Goyal et al [6] provide a quantitative measurement of this influence, providing an efficient algorithm to measure the influence one user has on another within an action-based network that depends on the time of a common action taken by each user as well as the total number of actions taken by each. Moreover, they introduce the idea of *partial credit*, defined as follows: Suppose that user u performs an action a at time $t_u(a)$ and S is the set of it’s neighbors that have performed that same action a before $t_u(a)$. Thus, the credit assigned $\forall v \in S$ is:

$$credit_{v,u}(a) = \frac{1}{\sum_{w \in S} I(t_w(a) < t_u(a))}$$

This paper computes Goyal et al’s *Bernoulli model with partial credit* for the influence user v has on u, defined as:

$$p_{v,u} = \frac{\sum_a credit_{v,u}(a)}{A_v}$$

Where A_v is the total number of actions v takes, as a feature of each edge in the user network and aggregated for all reviewers of a business in regression models, and I is the identity function.

Data

Data Source

This paper utilizes data collected from *Yelp* made public through their Yelp Open Dataset challenge [7]. In order to narrow our analysis, we focus only on users and businesses in the Pittsburgh area.

The uniqueness of this dataset is that it provides insight into individual user behavior and their network simultaneously. Much of the literature in related works draws from datasets that have one or the other - either a complete set of user actions, but no data on the relationships between users, or a complete view of a user’s networks, but no information about actions taken by individuals within the network. With our dataset, we will be able to look at both simultaneously in order to best understand network effects on decisions.

Data Attributes

There were two primary types of data available in the dataset: reviews and users. The review data includes the name and id of the business, the rating given, the id of the user who left the review and the timestamp of when the review was left. The user data included whether or not the user was a Yelp Elite member, their friends (limited to those also in Pittsburgh), and how long they have been a Yelp member for.

Overall, our dataset includes more than 165,000 reviews of approximately 5600 businesses, and more than 120,000 friendships between 16,000 reviewers, with over 10 years worth of reviews.

Limitations

As good as this dataset is, it does come with caveats. The first is that the dataset is incomplete, as Yelp only released a part of its collected information (Pittsburgh being in the dataset was a happy coincidence). However, since the data released was selected randomly, the overall insights drawn from it should still be valid.

Furthermore, we are restricted into only knowing about *Yelp reviewers*, and so are not able to see those that aren’t on Yelp or did not leave a review for a particular business. Yelp’s own internal estimates suggest only about 10% of their users ever post on the site, and only 1% are “active“ users [8]. This is a valid flaw, however, in the context of our analysis, is not insurmountable. As we will show in the next section, the network formed by reviewers aligns with what we we would expect from a more comprehensive online social network. Second, according to Yelp's own internal statistics as shown in Figure 3, the makeup of their user-base closely follows that of traditional social networks, with the exception of income, which is significantly higher. Furthermore, we also have no reason to believe that the behavior of reviewers differs significantly from that of non-reviewers in terms of which businesses they support. So a business that attracts a large number of reviewers would most likely also be attracting a large number of non-reviewers. Thus, by measuring the effect of social networks on reviewers, we will largely be seeing the same effect in non-reviewer networks as well.

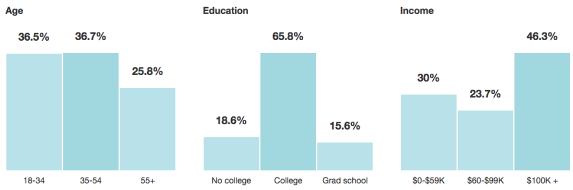


Fig. 3. Data on Yelp demographics. Although the income level of the typical Yelp user is higher than the U.S average, this aligns with the demographic of those most active on social media.

Networks

Using our review and user data, we were able to construct a network of approximately 16,000 Yelp reviewers in the Pittsburgh area, and looked at some of the properties of the network.

User Network

There were a total of 16698 users in our generated network. Figure 4 shows a visual representation of this graph, where users are represented as green dots and friendships between them are represented as lines connecting nodes.

The mean number of reviews each user makes is 5.98, but the median is only 2. This distribution is skewed by a small proportion of users who have reviewed more than 100 businesses (approximately .5%). Removing these power users, we see in Figure 5 that the number of businesses that users review follows a power distribution, with a majority of users reviewing only a handful of businesses and a sharp decline following that.

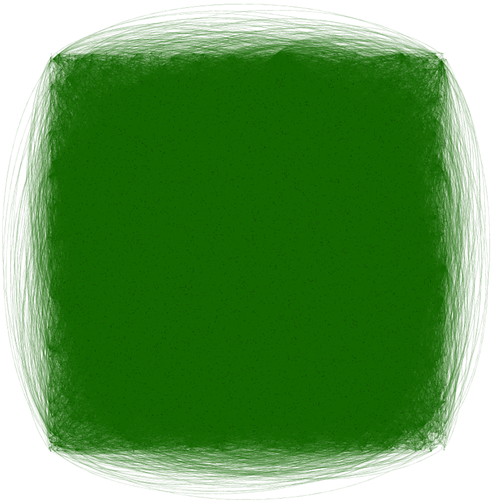


Fig. 4. Visual Representation of Network of Users

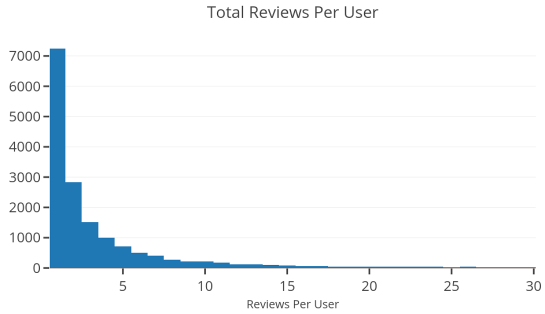


Fig. 5. Distribution of Number of Reviews

The average rating users give to businesses follows an interesting distribution. Nominally, the average rating given is a 3.823 with a median of 4. Since Yelp only allows for whole number stars, there are spikes at each of these values between 1 and 5 for the users who only provide one review. However, for the users that provide more than one review, Figure 6 shows how the distribution looks much more uniform, while the mean and median tend to stay the same.

In addition to the individual attributes of each user, we also look at the structure of the network formed by these users. Overall, the graph was comprised of 16698 nodes (users) and 60513 edges (friends), giving the graph an overall density of 0.000434 - as with most large social networks, the graph appears to be very sparse. Figure 7 shows the distribution of friends

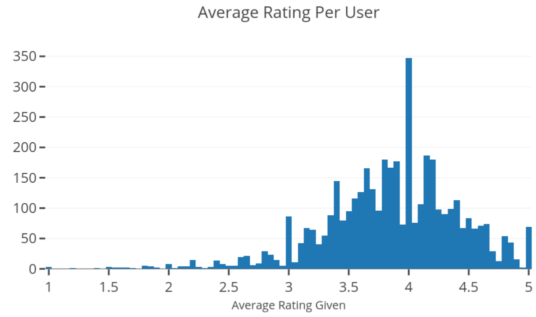


Fig. 6. Distribution of Average Rating Given For Users with More than One Review

per user in the graph. On average, each user had an average of only 3 friends, with only 10% of users having more than 10 friends.

We also look at the individual ties themselves. We weight each edge by the number of businesses the two users commonly reviewed. As stated previously, the average number of businesses reviewed by a user is approximately 6. Figure 8 shows how this distribution is heavily skewed right - the average edge has a weight of 1.89 between two friends, while the median weight is a 0. For each user, their average edge weight is small, only 0.475. Furthermore, the average edge weight is strongly correlated with the number of reviews given - suggesting that the more active users tend to overlap with their friends more often.

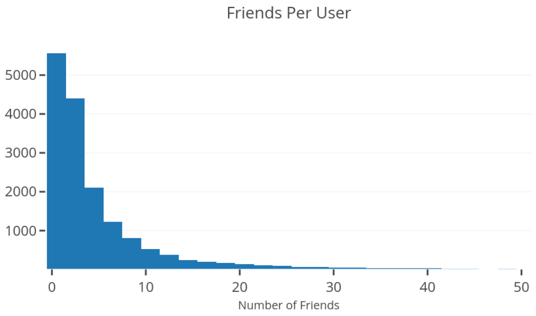


Fig. 7. Distribution of Friends

As observed by R. Kumar et al., online social networks can be partitioned into three regions - singletons made up of singular/dual nodes that do not participate with the rest of the network, isolated communities that are overwhelmingly in some kind of star structure, and one giant component that has a tightly connected core region. Our analysis of the Yelp network reveals the same pattern: of the 455 connected components in the graph, 398 are singleton communities (87.4% of all communities, 4.7% of nodes). The largest component includes 15726 nodes (94% of all nodes), and has a diameter (the greatest distance between any pair of vertices) of only 15. The average clustering coefficient (number of triangles in the graph) is around 13.7% in the overall graph, and jumps to 14.4% in the largest connected component.

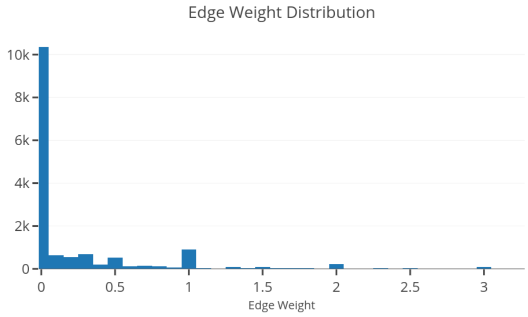


Fig. 8. Distribution of Common Businesses

In addition to the number of connected components, the graph also contains more than 60,000 maximal cliques (fully connected subgraphs), with the largest clique containing only 21 nodes - so while the graph remains connected, it is not very interconnected, thus following the results found by Kumar et. al.

Of the 16,000 reviewers in the network, about 22% of the reviewers are “elite users“. Compared to non-elite reviews, Yelp elites average about 14 reviews per user compared to 4 for non-elites, and their graph is more more centralized. Borrowing terminology from Subramani and Rajagopalan, the elite members exhibit all the characteristics of active, involved members of the network who wish to share information organically, and not at the behest of any marketing party. Table 1 compares the graphs formed by elite users to the one of non-elites. Overall, elite users are far more centralized, with a much higher average clustering coefficient, graph density, and a smaller graph diameter. However, elite users only make up a fraction of the overall number of reviewers in the Yelp network, and despite their increased activity and interconnectedness, their influence on average is on par with the non-elite members, suggesting that merely being elite doesn’t have an effect on how influential a user is to their friends. Visually, we can further see this in Figure 9, which shows the

	Attribute	Elites	Non-Elites	Overall
1	nodes	3762	12936	16698
2	edges	27005	19103	60513
3	avg. reviews made	13.95	3.67	5.98
4	connected components	139	3015	455
5	singleton components	131	2275	103
6	size of giant component	3617	9014	15726
7	clustering coefficient	0.33	0.06	0.137
8	diameter	7	18	15
9	avg. edge weight	3.58	0.08	1.88
10	average influence	0.02	0.01	0.01
11	cliques	33994	17854	623
12	density	0.0038	0.000228	0.00043
13	median component size	1.00	1.00	2.0
14	number of friends	18.19	4.07	7.25

Table 1

Attributes of the graph of elite Yelp reviewers, non-elites, and overall

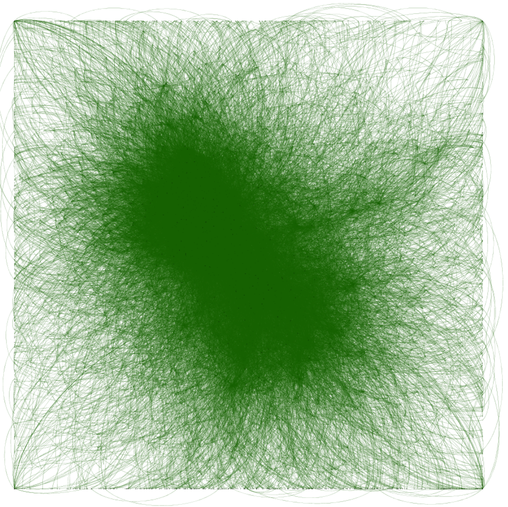


Fig. 9. Graph comprised solely of Elite Yelp users. Compared to non-elites, the graph is much more interconnected and centralized

Longitudinal Business Reviewer Networks

While the overall network of users does give us some insight into the nature of a reviewers relationship with a business, we also care about reviewers of a specific business. Moreover, we are interested in seeing how the network formation changes over time for each business.

To do this, we look at each individual business and the network formed by their users after each subsequent review. Using the timestamp of when the review was posted, we then get a longitudinal view of the networks formed over time, which gives us a better insight into how they change.

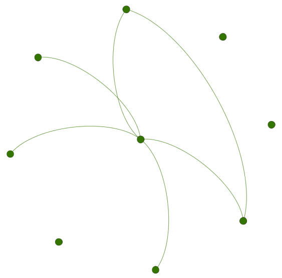


Fig. 10. Example Business Reviewer Graph After 1 Month of Business

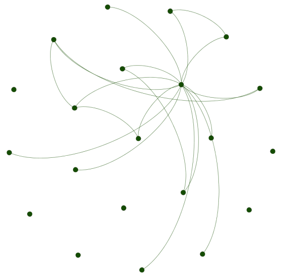


Fig. 11. Example Business Reviewer Graph After 3 Months of Business

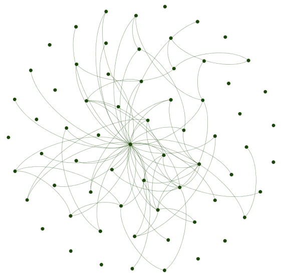


Fig. 12. Example Business Reviewer Graph After 1 Year of Business

Table 2 shows the changes in certain graph properties aggregated for all businesses at the the start, one month later, three months, 1 year (12 months), 2 years, and 3 years. Figures 10 - 14 represent the networks formed by the reviewers of Noodlehead, a local Pittsburgh restaurant, and it's change over time visually.

month	nodes	edges	reviews	pct elite	components	cliques	singletons	density	largest	avg. clustering	diameter	avg. rating
0	5.00	0.54	5.00	0.35	4.52	4.67	4.23	0.05	1.47	0.02	0.40	3.74
1	5.66	0.73	5.66	0.35	5.04	5.26	4.71	0.05	1.61	0.03	0.47	3.73
3	7.20	1.32	7.20	0.35	6.22	6.69	5.82	0.05	1.95	0.03	0.60	3.72
12	11.79	3.99	11.79	0.34	9.52	11.11	8.95	0.04	3.20	0.05	0.92	3.69
24	16.83	7.34	16.83	0.34	13.11	16.18	12.42	0.04	4.58	0.06	1.19	3.67
36	21.18	10.77	21.18	0.33	16.12	20.73	15.35	0.04	5.86	0.06	1.36	3.67

Table 2

Change in Graph Properties Over Time, Longitudinal Business Review Networks

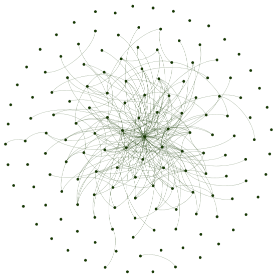


Fig. 13. Example Business Reviewer Graph After 2 Years of Business

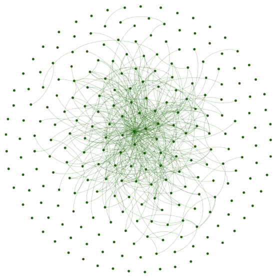


Fig. 14. Example Business Reviewer Graph After 3 Years of Business

As time increases, the number of nodes in the graph increases, along with the complexity of the graph in terms of edges and shape.

It is logical to see that aggregate values like number of nodes, reviews, etc, will all increase over time. However, we also notice the trends present in the proportional and graph measure values.

Over time, the proportion of reviewers of a business that are Yelp elites decreases steadily over time. This fits with common marketing knowledge about the early adoption of a new business/product before the eventual acceptance by the mainstream population.

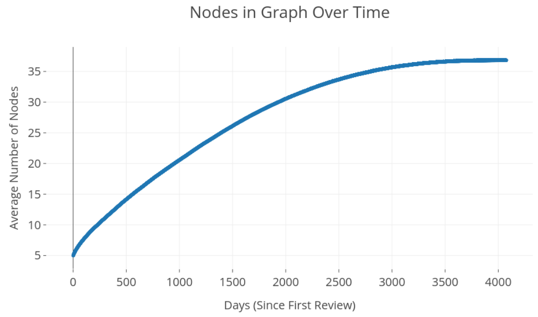


Fig. 15. Aggregated Values, like number of nodes in a graph, increase steadily over time before flattening out

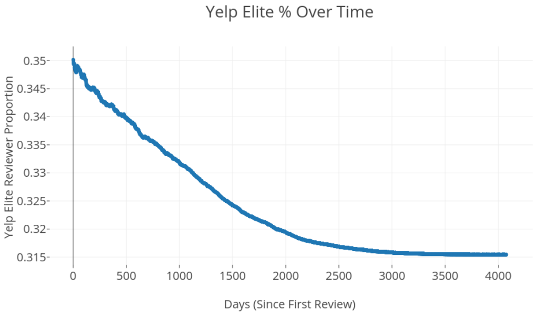


Fig. 16. Steady decline of Yelp elite reviewers in the network fits what we know about the characteristics of elite reviewers as trend-setters / new experience seekers

What the results of this analysis tells us is that our initial hesitation about looking at reviewer networks specifically rather than a completely organic social network is unfounded; Since the properties of our reviewer networks align well with prior literature on larger, less specific forms of social networks, we can comfortably use these networks for our analytic purposes and draw conclusions based on these findings about the impact of networks in general on buying behavior.

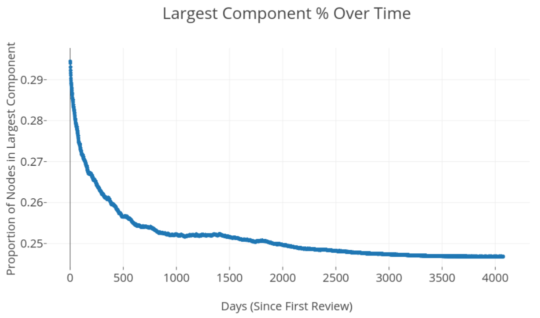


Fig. 17. The proportion of the largest component of the network also shows this behavior, as the proportion starts off high and then rapidly declines before evening out and begin gradually declining instead.

It is important to note here that we face a survival bias in looking at this data. As the amount of days increases, the only businesses that remain are the ones that have remained active for that duration of time. However, the features of these surviving businesses are important to us - the key to their long term success is likely linked to the properties of their reviewer network. But, in order for us to predict business success, it is necessary to come up with a more robust means of modelling business success.

Regressions

While understanding network structure has yielded interesting results thus far, we are ultimately interested in seeing if this information is useful for predicting the future success of a business.

We first make two assumptions about our review data. The first is that the earliest review of a business comes close to the day it opened - we make this assumption out of necessity, because we are not given access to the actual opening date of a restaurant, just the date of its earliest review. The second assumption we make is that a business with a large number of reviewers is more successful than one with fewer number of reviewers. It stands to reason that a larger number of reviews implies a larger number of customers, which we are attempting to approximate.

With these assumptions in mind, we first looked at the snapshot of each business at one month and three months after it's first review (our proxy for the business being open for one and three months respectively). Next, we looked at the number of reviews that each of these businesses accumulated after one, two, and three years, as ultimately we use the number of reviews a business has gotten as a measure of the amount of traffic, and therefore revenue, the business has gotten thus far. Using the review data collected at each interval as well as the properties of the networks at these early moments in the business's life, we will model the future success of each business.

Predicting Future Reviews

We explored the use of cross validated OLS regression and lasso regression models in order to predict the number of reviews for each business, and assessed the success of our model using root MSE.

To establish a baseline, we first modeled our year-end review numbers using only features that could be collected from the reviews themselves. This included the number of reviews/reviewers, the average rating given, the difference between the rating that each reviewer gave and their mean rating, and the percentage of reviewers that were Yelp elite. Table 3 summarizes the model created using these features, in which very few features are statistically significant, and the adjusted R^2 value is relatively low at 0.617.

We then compared the results of these models with the results of a more comprehensive model that included network based features, such as component size, density, clustering coefficient, etc., and summarized the difference in their prediction power (in terms of RMSE) in Table 4).

Overall we found that our network-based models were significantly more accurate in predicting the overall number of reviewers at each year-end time that we chose.

Moreover, with the LASSO model, we see that the variables of highest importance were a combination of the review based features as well as network based features. In particular, within the network features, it seems that more centralization is detrimental to future success - attributes like number of edges, proportion of nodes in the largest component, and proportion from neighborhood (a measure of how many reviewers were friends with at least one prior reviewer) have negative coefficients associated with them. This makes sense when considering the low amount of influence

	Dependent variable:
	num_reviews_1yr
pct_elite	0.163
	(0.582)
avg_rating	0.304
	(0.298)
avg_rating_diff	0.355
	(0.456)
num_reviews	4.646***
	(0.061)
age	-0.016
	(0.031)
age_in_months	0.363
	(0.843)
Constant	-17.172^{***}
	(1.170)
Observations	4,278
R^2	0.617
Adjusted R^2	0.617
Residual Std. Error	9.609 (df = 4271)
F Statistic	1,148.944*** (df = 6; 4271)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3

Cross validated OLS regression using review only features

users exert on each other - a more diverse graph with users that have little in common with one another extends the possible number of new customers that can be reached, and therefore, having a less centralized network is more beneficial - moreover, a less centralized network could be indicative of a business with more universal appeal, rather a business that only appeals to a niche crowd that might be more insular. Thus, it appears that a key to a businesses success is in the size of its middle region.

	Time Frames	Net.	Non-Net.	Diff
1	One Month, One Year	8.22	9.68	1.46
2	One Month, Two Year	15.25	20.48	5.23
3	One Month, Three Year	21.07	31.53	10.46
4	Three Month, One Year	3.08	6.79	3.71
5	Three Month, Two Year	12.57	14.02	1.45
6	Three Month, Three Year	21.58	26.66	5.08

Table 4

Comparison of the root mean squared error (RMSE) between models containing network features (along with review based features) and models without. Trained on earlier time frame data to predict later time frame number of reviews.

	Coefficient
(Intercept)	11.30
number of singletons	1.75
median component size	-0.95
largest component size	6.43
density	18.47
average clustering	-0.04
network diameter	0.42
% elite	0.34
average rating	0.24
average rating diff	0.39
average rating diff in large comp.	-0.11
edges	-1.05
average influence	7.93
proportion of singletons	-13.32
proportion of nodes in largest comp	-37.08
proportion from neighborhood	-6.93

Table 5

Lasso Coefficients for model predicting the number of reviews after one year using 3 month data. Note the combination of network based features and non-network features

An ANOVA test between the two types of models also shows that the network features were significant with an F-value of about 28 and a p-value ≤ 0.005. Thus, for the purposes of prediction, we are confident that understanding the network of the reviewers does significantly improve the results of any model.

	Res.Df	RSS	Df	F	Pr(>F)
non-network	4271	394369.92			
network	4252	350430.55	19	28.06	0.0***

Table 6

Comparison of Network and Non-Network OLS models for predicting one year reviews after one month

Random Forest

In addition to OLS and lasso models, we attempted to predict future reviews using a CART random forest. After training the model with 1000 trees and providing the same training and testing data for both, we once

	%IncMSE
nodes	16.06
edges	5.92
cliques	14.07
components	12.20
singletons	10.33
median comp. size	-0.48
largest comp. size	6.98
density	8.06
average clustering	1.91
diameter	5.60
pct elite	9.03
avg rating	7.43
avg rating diff	7.75
lc avg rating	7.85
lc avg rating diff	3.94
avg_edge_weight	11.75
avg_influence	5.29
num_reviews	21.42
prop_singletons	7.92
prop_largest_component	9.03
age	17.61

Table 7

Random Forest

again saw significant improvement by using network based features compared to just review based ones. The variable importance plots below highlight how each of the different features were utilized, as well as the relative importance of each one. While review-based features alone are important, performance significantly improves once the non-review features are incorporated into the model.

Comparison Across Starting Points

Across our different models, we can also compare how well each one did when trained using data collected from one month out vs three months out for predicting the number of reviews after 1, 2, and 3 years.

	Prediction Year	M-1	M-3	diff
1	One Year	6.59	2.41	-4.18
2	Two Year	11.67	8.40	-3.27
3	Three Year	15.06	12.84	-2.22

Table 8

Lasso RMSE Model Comparisons

	Prediction Year	M-1	M-3	diff
1	One Year	8.87	3.25	-5.62
2	Two Year	16.52	14.04	-2.48
3	Three Year	23.96	22.35	-1.61

Table 9

Random Forest RMSE Model Comparisons

	Prediction Year	M-1	M-3	diff
1	One Year	6.59	9.32	2.73
2	Two Year	12.23	14.11	1.88
3	Three Year	16.88	15.58	-1.30

Table 10

OLS RMSE Model Comparisons. Note that this is the only model in whch the one month data gave a better RMSE, however, it performed worse than lasso and random forest overall.

Conclusion and Future Work

Long term success for businesses depend on a variety of factors, and the early reviews are an essential catalyst for future outcomes. We find that the social networks formed by reviewers on Yelp in the Pittsburgh area are similar to other online social networks in terms of connectivity, shape, degree distribution and centrality. Using the social network formed by a business in early stages, we can construct linear regression models that can predict the number of reviews that the business will receive in the future, with relatively low root mean squared error. One of the most important features in predicting the long term success of a business is the proportion of its review network comprised of star-shaped components, most commonly found as part of Kumar et al.'s "middle region" of online social network components.

Future work on this topic can expand upon the data sources used in order to get a more fully fleshed out view of relationships between reviewers and their buying decisions - in addition to reviews, Yelp users can also "check-in" to a particular business which also gives an insight into where they are going. Furthermore, Yelp is not usually thought of as a primarily social networking website, and so while users can add their friends, they may not necessarily be as unconnected as their graph might indicate them to be.

More work can also be done in terms of the modeling by expanding on the range of possible features. It's reasonable to believe that different types of businesses, such as restaurants, mechanics, barbers, etc, may have different network effects associated with them. Furthermore, reviewers based in cities might differ from those in smaller communities as well as larger metropolitan areas.

This work has a lot of intriguing possibilities for future expansion. As we move further into developing analytic tools that capture human interactions in more quantitative detail, recognizing and making decisions based on social interactions will become more and more feasible, and ultimately have long lasting impacts to a variety of business applications.

Acknowledgements

Thank you to Yelp for providing their open dataset for academic research purposes.

Thank you to Professor Maryam Saeedi for advising this paper and providing guidance each step of the way, as well as the Carnegie Mellon Economics department for the opportunity.

Thank you to my parents, for pretty much everything else.

All code and data for this project can be found here: https://github.com/ahavanur/yelp-networks. Graph visualizations made with Gephi.

References

(1) Jamie Doward (September 1, 2012). ”How online reviews are crucial to a restaurant’s takings”. The Guardian. Archived from the original on November 13, 2013. Retrieved November 27, 2013.
(2) Tom Gara (September 24, 2013). ”Fake Reviews Are Everywhere. How Can We Catch Them?”. Wall Street Journal. Archived from the original on February 14, 2017.
(3) Luca M. Reviews, reputation, and revenue: The case of Yelp.com. (2011) . Harvard Business School NOM Unit Working Paper 12-016
(4) Mani R. Subramani , Balaji Rajagopalan, Knowledge-sharing and influence in online social networks via viral marketing, Communications of the ACM, v.46 n.12, December 2003
(5) Ravi Kumar , Jasmine Novak , Andrew Tomkins, Structure and evolution of online social networks, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, August 20-23, 2006, Philadelphia, PA, USA
(6) Amit Goyal , Francesco Bonchi , Laks V.S. Lakshmanan, Learning influence probabilities in social networks, Proceedings of the third ACM international conference on Web search and data mining, February 04-06, 2010, New York, New York, USA
(7) The Yelp Open Dataset. https://www.yelp.com/dataset
(8) Yelp and the 1/9/90 Rule:
https://www.yelpblog.com/2011/06/yelp-and-the-1990-rule