Yelp: An Analysis of User Behaviour Over Time

Brandon Dixon
University of Chicago

Jack Liang
University of Chicago

This paper attempts to understand the changes in behaviour of Yelp users over time. In particular, this paper investigates the distribution of one-time users' reviews and compares it to the typical distribution of reviews. Furthermore, this paper studies the effects of number of reviews given by a user prior to a fixed particular year has on their reviews in that fixed year. We find that the distribution of one-time reviews is significantly different from the distribution of all reviews, suggesting that initial user behaviour on Yelp differs from typical user behaviour. We also find that the having more prior reviews causes users to review businesses more favourably than they did earlier, suggesting that additional experience on Yelp makes users nicer.

Introduction

In the age of social media and widespread mobile internet, consumers have both unprecedented access to quantities of information and the ability to share their thoughts on a mass scale. This has led to increasingly informed consumer decision making, which has real effects in consumer-driven markets. While there exists a plethora of rating services, some more sophisticated or complex than others, none is more widespread in both the number of consumers that use it and the number of businesses it reviews than Yelp. Yelp has, as of December 2015, about 160 million unique visitors monthly, as well as 95 million cumulative reviews since its inception in 2004[1], something that print media would never have been able to accomplish. Yelp operates primary on a star-rating system, where businesses are awarded somewhere between 1 and 5 stars (with 5 stars obviously being the best) based on a simple average of user reviews[2]. These stars have non-trivial meaning; a study by Luca (2011) found that a one star in Yelp rating corresponded to a 5 to 9 percent increase in revenue.

As such, understanding the meaning behind these reviews, as well as any patterns in user reviews can provide insight both into consumer decision making and the effect of such vast quantities of information have on spending. However, such data is often difficult to analyze due not only to its subjectivity and the variance between individuals (why does person A's 5 star rating and person B's 1 star rating differ so much? why did person A give a 5 star rating rather than a 4 star rating?) but also to its sheer quantity.

We first should standardize some Yelp-terminology. First, it should be noted that not everything on Yelp is a business; for example, churches are up for review on Yelp, as are historical landmarks. For the sake of brevity, we will refer [perhaps slightly improperly] to everything on Yelp as a business. Each business is sorted into one or more categories, which are classifications defined by Yelp that are some reflection of what

[1] See **this Yelp factsheet**

[2] There are some subtleties, such as Yelp's review filtering algorithm that are not particularly important to us.

the business does (i.e., restaurant or hotel). Some categories are broader than others (for example 'professional services' compared to 'spa') and some have clear overlap (for example 'restaurant' and 'small food'), but rather than attempt to create a more precise sorting system we shall stick to Yelp's predefined categories.

As previously mentioned, Yelp's stars reflect the average of the user ratings, rounded to the nearest integer multiple of .5. We call the average of the user ratings for a particular business, before the rounding, the average rating. Similarly define a user's average rating and their rating in a year Y. We call the normalized mean given a set of categories the actual result minus the predicted result when we regress average rating on a set of indicators for each category. All other terminology should be obvious.

This paper studies the effects of experience with Yelp on user behaviour. In particular, we explore the observation by Booth Professor Matt Taddy that regular Yelp users tend to become more generous with their reviews. We do this by first 'normalizing' business ratings across different business functions. That is to say, because we cannot assume that the Yelp rating for restaurants carries the same meaning as the rating on shopping malls, we standardize ratings by using Yelp's predefined categories as a proxy for business function. These results in itself are interesting if perhaps not particularly surprising; Yelp overall seemed to have an average business rating of about 3.6 with real-estate and hotels scoring particularly below this while businesses in the 'local flavours' category (i.e., local monuments or landmarks) scored particularly highly.

Next, we looked at the distribution of users with only one review as a proxy for first reviews. We hypothesize that due to the fixed cost of joining Yelp (the time to sign up for an account, verify one's email, etc.), initial reviews would be particularly polar, as people would need some strong reason to sign up. We expect this bimodal distribution to lessen at least somewhat after 2010, as in 2010 Yelp eased the difficulty of making an account by allowing users to link with Facebook[3]. We find sufficient evidence to reject the null hypothesis that the reviews from users with only one review and the reviews from all users were sampled from the same distribution, suggesting that at some point after a user makes their first review their behaviour on Yelp changes in a significant way, or, equivalently, a Yelp user's motivation for signing up differs from their motivation for posting a review. However, we do not find a significant difference after the change in Yelp's account creation policy, suggesting that initial reviews are polar regardless of how high the fixed cost is to sign up.

Finally, we then analyzed the effect of the number of ratings a user had written prior to a fixed year had on how much 'nicer' that user was in that fixed year. To put this precisely, we studied the effects of additional reviews prior to a fixed year (as a proxy for experience) on the difference between the average normalized review in that fixed year and average normalized review in all prior years. We hypothesize, in accordance with Taddy's claim, that the number of reviews prior to a fixed year should be positively related to the difference between the average review in that fixed year and that user's previous average review. We do find statistically significant evidence to support this claim, suggesting that additional reviews beyond the first in earlier years cause users to be more generous with reviews in that year, that is to say, users tend to get nicer with experience.

[3]See **this Yelp announcement**

Related Literature

There is quite a bit of literature concerning Yelp data, but we would like to highlight a few of the more interesting or novel results. In particular, we look at a pair of working papers by Michael Luca, with the second being co-authored by Georgios Zervas, entitled "Reviews, Reputation, and Revenue: The Case of Yelp.com" and "Fake it Till You Make It: Reputation, Competition, and Yelp Review Fraud". The first paper examines the real effects of consumer reviews on demand, in particular, for restaurants. In this paper, Luca makes causal claims about the impact of Yelp's ratings using a regression discontinuity model.[4] To summarize, Luca finds that different Yelp star ratings have very real effects; as previously mentioned, a one star increase in Yelp leads to a 5-9 percent increase in revenue. Luca also claims that the mere existence of Yelp's service has real effects; he suggests that as restaurants with chain affiliations are less impacted by Yelp ratings and have declined in market share since the inception of Yelp, suggesting that online consumer reviews are in some way replacing previous forms of reputation (for example, mass advertising that small restaurants do not have access to).

In the second paper authored by Luca and Zervas, the authors explore the credibility of Yelp's reviews by looking at Yelp's filtering algorithm. [This paper was what motivated us to look into Yelp data at all and prompted our first inquiry, which attempted to look into the prevalence of filtered reviews around rounding points as evidence that Yelp was, as articles[5] sometimes claim, extorting businesses for advertising revenue. Unfortunately, that endeavor was perhaps beyond our abilities at the present moment.] Luca and Zervas found that review fraud committed by restaurants was fairly frequent, and that restaurants with a 'weak reputation' (i.e., very few reviews or primarily poor reviews) were more likely to commit review fraud. Finally, they round that restaurants with increased competition were more likely to face unfavourable fake reviews.

In addition to these papers, there is a wealth of interesting papers written by individuals participating in Yelp's dataset challenge[6], with topics ranging from using keywords in written reviews to attempt to predict star ratings to the bias in user ratings when reviewing newer companies. We would encourage anyone interested to survey the abstracts of these papers; the enormous volume of data Yelp has made available allows for quite interesting analysis and interpretation.

Data

This paper makes use of the Academic Dataset published by Yelp. Those affiliated with a University or in possession of a Yelp API key can obtain Yelp's permission to download the data. The original format of this data is a .json file. In order to make this data more usable, we applied a publicly available Python script[7] to convert it to a .csv file. The data spans 2004 to 2012.

The data itself is structured in three object types; user, review, and business. User objects contain information about individual Yelp accounts including their number of reviews, and their average review. There were 130,874 user objects. Review objects contain information about particular reviews including the date, the user who posted the review, and the business that was being reviewed. There were 330,072 review objects. Finally, business objects describe particular

[4] Note that Yelp rounds to the nearest half-star, so by looking at discontinuous jumps in revenue for restaurants with similar but differently rounded Yelp averages, Luca isolates the effect of the shown Yelp rating.

[5] Such as **this article by the Huffington Post**

[6] See **the Yelp dataset challenge**

[7] See **this script**

American businesses with information on the number of reviews, their star rating, location, and a series of dummies representing their categories. There were 13,491 business objects.

Note that this list of category dummies is not exhaustive, in particular, the exact categories a business is listed under can be found listed in each business object. However, if we chose to get incredibly specific [for example, {restaurant, small food[8], Chinese, dim sun} could be the categorization for a dim sun restaurant] our sample size for each particular set of categories would be tiny, and the number of unique category sets which has cardinality $2^{\text{number of categories}}$ would be far too large for us to work with. We created dummies to represent the largest categories (i.e., the ones with the most number of businesses) in such a way that each business was represented by at least one dummy.

We also should note that the number of review objects is not the total number of reviews associated with each user object. That is to say, while each review object is associated with exactly one user object, not every review posted by a particular user can be found in the set of user objects. We believe that the purpose of this is to make the file size workable [if the data set included all such reviews, there would be nearly 7 million reviews, a more than 20-fold increase from the number currently included], however, it inconveniences us because it renders a lot of the information associated with the user object useless, as for example we do not know when the 'missing' reviews were posted, what category or categories they were associated with, and what the actual ratings were [we can compute their average of missing reviews, but not the individual breakdown]. So, we need to add an assumption about the dataset, namely that the reviews included in Yelp's dataset were in some way a random sample from the collection of all reviews from this set of users.[9]

Methodology

Normalization

We first attempted to normalize the data, that is to say, generate predictions for the ratings of a particular business given its categories. By doing this, we can then look at the residuals to determine both whether a business as a whole was doing better than other businesses in its categories and whether a particular user's rating for a given business in some fixed set of categories was above or below what a typical rating might be.

It is first important to note that businesses could be in more than one category (sometimes businesses were in no category; these data points were very sparse and were thus dropped). Because a simple regression of average score on our set of categories would lead to biased estimators, we attempted to isolate interaction effects by constructing 'double categories', i.e., if a business was both a restaurant and in the small food category, we would remove them from the restaurant category and from the small food category and place them in the new restaurant and small food category. We did this for the largest 'double categories', throwing away double categories with less than 30 data points.

The explicit regression model we look at is

$$Y_i = \alpha_0 + \sum_{k=1}^{n} \alpha_k \text{category}_{i,k} + U_i$$

where Y_i is the set of businesses' star rating and $\text{category}_{i,k}$ were indicators that the i^{th} business was in

[8]Small food is a business such as a cafe or food stand

[9]If the set of reviews included was in anyway not random, the dataset would be rather useless, so this seems to be a safe assumption to add.

the k^{th} category (in total, there were 31 categories including the 'double categories').

The coefficients α_k from this regression along with their standard errors allow us to test the hypothesis that Yelp ratings vary across different categories. That is to say, if we can reject the null hypothesis $H_0 : \alpha_k = 0$ at a fixed confidence level of 5 percent, then we have evidence to say that businesses in the k^{th} category have an expected mean rating in some way different than the typical Yelp business.

We then use these normalized reviews to generate residuals for both user ratings and business ratings. Note that we include all coefficients from the previous regression (including the ones that were not statistically significant) as dropping only the statistically insignificant dummy variables makes model interpretation difficult.

It is important to note that while the business object includes information about the location of the business [in particular, the nearest major city], we do not attempt to control for this for a few reasons. The first is that including dummies for 119 major cities included would make our regression large and unwieldy. That is to say, if we include dummies on each city, then perhaps we should also in some way control for interaction effects (what if restaurants in Chicago are simply better than those in Albany?), which simultaneously makes the number of variables we are concerned about explode while sending the number of data points in each sample to be very small. The second is we are primarily concerned with user behaviour, and tying a user to a particular location is difficult. For example, many of the local attractions are reviewed by tourists, whose (original) location would vary greatly, but presumably real estate agents are only used by locals. As such, we make an assumption that user behaviour is relatively homogenous across different locations.

Distribution of First Reviews

We now consider the hypothesis that the first reviews of Yelp users differs from the user's typical behaviour. We proceed by taking all users with exactly one review and comparing the distribution of their reviews to the overall distribution. The distribution comparison is done through a two-sample Kolmogorov-Smirnov test, which uses a test statistic to compare the empirical cumulative distribution function of the normalized first ratings with the empirical cumulative distribution function of all ratings. Note that Kolmogorov-Smirnov performs poorly when there are frequent ties (which lowers the power of the test).[10] However, our normalization process should [greatly, but not completely] reduce the probability of ties, and in either case the p-value generated should be conservative.

In particular, we run a two-sample Kolmogorov-Smirnov test on the residuals $Y_i - \hat{Y}_i$, where as previously mentioned the Y_i is a businesses' star rating and $\hat{Y}_i$ is its predicted star ratings given the set of categories it belongs to. These residuals reflect how far from 'normal' or 'typical' a particular review it; a particularly positive or negative residual thus reflects a particularly polar review.[11]

We then run a two-sample Kolmogorov-Smirnov test on the normalized one-time reviews prior to April 2010 and after April 2010 to test if Yelp's new policy had any

[10] See **this paper** for a more thorough discussion, including notes on why the p-value is conservative.

[11] Note that we will discuss later in the paper that our normalizing process is time sensitive- that is to say, our normalizing constants from the span 2004-2012 may not be the same as the normalizing constant for a subset of that span. This is not an issue here, as the dates of the reviews for the one-time reviewers follows the distribution of all reviews, so using the most general normalizing procedure [not correcting for category *and* year] is justified.

effect on users. Note that accounts in the post period are not guaranteed to have been created after the policy change, however, we suspect for the most part that they are, as there is little benefit to having a Yelp account but not post reviews.

It is important to note that we take all users with exactly one review, *not* the first review of each user. The reason for this is because user objects do not have the date the account was created [and even if it did, we cannot guarantee that they would have created a review on that first day], so with our limited subset of all review objects simply picking the earliest review would not be a guarantee of picking the first review.[12] However, in the set of users with exactly one review we know that this one review, if it is listed, must be their first. Using this set of reviews thus allows for a comparison of first reviews with the overall data set.

However, such methodology is problematic as it is not necessarily true that people who reviewed exactly once are somehow indicative of a general user. In this way, perhaps it would be reasonable to expect that users with only one review would be in some way biased away from the overall sample of Yelp users, suggesting that any result we attain would be meaningless. This problem would be easily overcome if (1) the academic data set contained all reviews for each user object or (2) it was legal for us to run a data scraper to extract all the reviews for a particular user, however, unfortunately neither of these things are true.

Thus, we need to add the assumption that people who only reviewed once posted reviews that are representative of the overall first review. This seems like a fairly strong assumption; if this review was fairly late into the data set [say late 2012], then it seems very reasonable to think that they may or may not have come back to post additional reviews, however, if this single review was posted earlier in the data set then the fact that they never came back suggests that they either felt so strongly that they needed to post one very polar review and never reached that same 'emotion' threshold again.[13] In this way, our data may be more extreme [bimodal] than the actual distribution of first reviews; again, without full access to all user data we cannot explore this further and must accept this assumption.

Effect of Prior Reviews

Finally, to analyze the effects of the number of prior reviews which can be thought of as a proxy for gained experience on a user's reviewing behaviour, we propose two methods: for the first, fix a particular year and find users with a non-zero number of reviews in that year. Normalize these reviews and take the mean of the normalized reviews. Now subtract the normalized mean of reviews made by this user made prior to 2012, and call this quantity M_i. Regress M_i on the number of reviews prior to our fixed year, which we call X_i. That is to say, regress on the following model:

$$M_i = \beta_0 + \beta_1 X_i + U_i$$

Note that we require X_i to be at least one so that we can compute the difference between the mean normalized review in the fixed and the mean normalized review for prior years. In this way, then, we test not the effect of having experience versus not having experience but rather the effect of additional experience.

Under the null model that user behaviour does not change over time, we expect that both β_0 and β_1 to be

[12] For example, for a user with 15 reviews but only 10 listed, it is possible that the first review listed was in fact their third review.

[13] Alternatively, a mundane explanation like they lost their password or created multiple accounts is possible.

zero. Note that $\beta_0 \neq 0$ suggests that user behaviour in that year differed, on average, from user behaviour in previous years [for example, $\beta_0 > 0$ suggests that users in this year are nicer than users in previous years]. Furthermore, $\beta_1 \neq 0$ suggests that user behaviour changes as a function of the number of reviews they had posted prior to the fixed year [for example, $\beta_1 > 0$ would support the observation that users get 'nicer' the more reviews they post].[14]

Note that this methodology is not perfect because the use of normalization assumes that Yelp's categories have been stable across the 8 year span, which is not necessarily the case. This change should be reflected in the constant term β_0; if the categories had equal proportion over time we would expect β_0 to be close to 0, however, this is not the case for some of our regressions. Furthermore, this methodology relies on sampling from user reviews- the dataset we use does not have all user reviews, so we assume that the subset of reviews that were included are a random sample from the set of all reviews [in particular, the distribution of dates is the same].

Our alternative model is to regress the difference between the average rating of restaurants for a given user in a fixed year and their average rating prior to that year on their number restaurant reviews prior to that year. Note that we no longer normalize, as we only consider restaurants. Our model is thus

$$D_i = \gamma_0 + \gamma_1 R_i + U_i$$

where D_i is the difference of the average ratings, R_i is the number of restaurant reviews prior to the fixed year. We again require R_i to be at least 1. We propose this model because it is unaffected by changing category proportion, that is to say, because it does not require normalization, we can safely interpret γ_0 as some baseline difference between restaurant reviews in a our fixed year and prior years and γ_1 as the effect of one additional restaurant review prior to the fixed year (which can serve as a representative of experience on Yelp).

This methodology is also not perfect, as it limits our sample to restaurants. We argue that because Yelp's function, at least early on, was primarily to review restaurants, and our dataset is limited to primarily Yelp's earlier years, that any sort of observable trend in user behaviour is generalizable.

We now include a brief discussion of why OLS may or may not be a valid model for the data in both these regressions. A fundamental question that must be addressed is why we expect our data to be linear; that is to say, why we expect additional early reviews to increase the difference between the mean rating in a fixed year [say 2012] and the mean rating from 2004-2011 in some linear fashion. In general, this seems false; certainly the difference between the mean rating in 2012 and the mean rating from 2004-2011 is bounded above by 5, so if our hypothesis that there is some positive correlation is correct, sending the number of reviews X_i or R_i to infinity would, in our model, send the mean difference to infinity. In that sense, then, perhaps a modified logistic model would perhaps be more suitable.[15] However, a more complex model leads to more difficult interpretation of results. As the number of reviews prior to a fixed year tend to be reasonably bounded [certainly less than 1000, in general, less than 100], we can perhaps apply

[14] A similar model would be to regress M_i on an indicator if the user had any reviews at all prior to our fixed year to see if the having any experience at all affected a user's behaviour. However, as previously mentioned, the academic dataset does not include all associated reviews with a particular user, so for many users there is no way to know if they did or did not have reviews prior to our fixed year.

[15] That is to say, a model bounded above by 5 and below by -5 that sends $D_i \to 5$ as $n \to \infty$.

some local linearity argument to the logistic model to allow our data to fit a linear pattern.[16]

We also should address the fact that when we add an additional rating before 2012, there will be some change in the mean of reviews before 2012, that is to say, our quantity M_i or D_i will change both because [if our hypothesis is correct] we have additional experience and because the mean review prior to 2012 has changed. This is seemingly some sort of an endogeneity problem; however, this problem is mitigated when we note that the expected change in mean review prior to 2012 with an additional review prior to 2012 is zero, given the assumption that all reviews before 2012 are 'the same'. Note that this assumption is flawed under our current model as, for example, the 5^{th} review in the set of reviews before 2012 will be different from the 15^{th} review in the same set because the 15^{th} review would be effected by the additional experience from the 10 more reviews that occurred between the two (i.e., reviews 5-14). We cannot fix this problem without some time-series considerations, so we simply add the assumption that, given a fixed year, all of a user's experience from prior to that year is applied just before the start of the fixed year and reviews during that fixed year do not add to this experience. However, it should be noted that the effect of additional reviews is small for a small number of additional reviews [see the results section], so this should not greatly hurt our model.

As previously noted, the Yelp academic dataset does not provide entire user histories, so despite the fact that we have their overall average rating and their total number of ratings, we cannot use this as (1) we do not know the breakdown of each rating and thus cannot normalize this average and (2) we do not know the dates of each of the individual ratings. However, with the added assumption that our data is a random sample of all reviews associated with the listed set of users, our regression results are valid for the sample (and thus for the population of all Yelp users).

Results and Discussion

Normalization

The results from the regression used to normalize the ratings across different categories are in Figure 1.

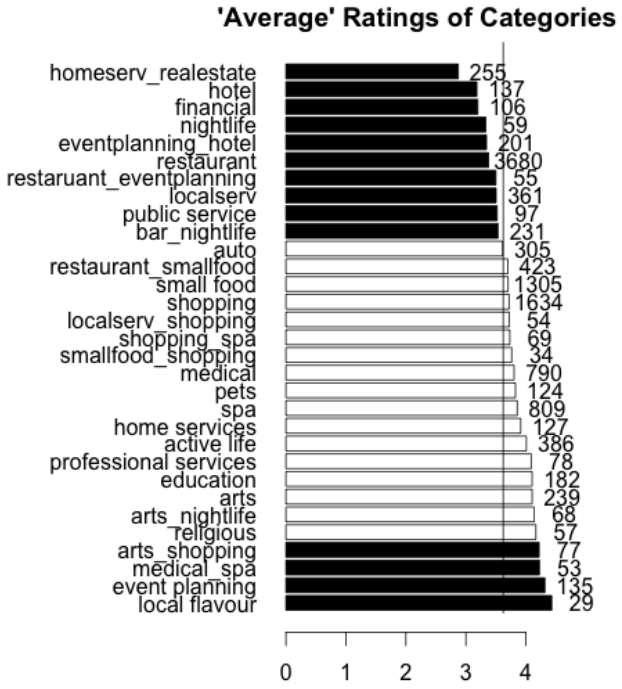
We note that for many of these categories which we call category$_k$, the coefficient α_k was not statistically significant at the 5 percent level. In particular, the coefficients that were significant were the coefficients on restaurant, local services, hotel, event planning, nightlife, financial, public services, local flavour, home service and real estate, bar and nightlife, event planning and hotel, arts and shopping, medical and spa, restaurant and event planning, and the intercept.

To highlight some of the more interesting points, we first notice that our largest category [restaurants] has a statistically significant [negative] effect on our predicted rating. This is surprising because we would expect the largest category to in some way have the largest effect in determining the mean, so this suggests that it is very likely that being a restaurant has a non-zero effect on rating.

Another interesting, if perhaps confusing, point is that event planning alone has a very positive effect [significant at even the 1 percent level], while it has a very negative effect when combined with restaurants or hotels, which both have a negative effect. This seems to be because businesses categorized solely in event planning are primarily photographers and caterers, whereas businesses in both event planning and another category

[16] Such an argument deserves a justification much stronger than this hand-wave, but unfortunately a rigourous argument is beyond the scope of this paper.

Figure 1: Results from Normalizing



(a) Note: black bars indicate that the result was statistically significant at the 5 percent level. The vertical bar is the mean for all the data, which was 3.62 stars. The number to the right of the bars is the number of data points we had for that category.

Variable	Coefficient	(Std. Err.)
restaurant	-0.487**	(0.135)
spa	-0.006	(0.138)
auto	-0.252†	(0.144)
shopping	-0.143	(0.136)
smallfood	-0.160	(0.137)
medical	-0.060	(0.138)
localserv	-0.359*	(0.142)
hotel	-0.681**	(0.155)
arts	0.241†	(0.146)
education	0.241	(0.150)
eventplanning	0.459**	(0.155)
nightlife	-0.533**	(0.178)
financial	-0.666**	(0.160)
homeservices	0.050	(0.156)
pets	-0.037	(0.156)
activelife	0.144	(0.142)
publicserv	-0.343*	(0.162)
religious	0.303†	(0.179)
profserv	0.226	(0.168)
localflavour	0.567**	(0.213)
restaurant_smallfood	-0.164	(0.141)
homeserv_realestate	-0.991**	(0.146)
bar_nightlife	-0.329*	(0.147)
smallfood_shopping	-0.099	(0.204)
eventplanning_hotel	-0.520**	(0.148)
localserv_shopping	-0.141	(0.181)
arts_nightlife	0.276	(0.173)
arts_shopping	0.357*	(0.169)
medical_spa	0.363*	(0.182)
restaurant_eventplanning	-0.364*	(0.180)
shopping_spa	-0.132	(0.172)
Intercept	3.864**	(0.134)

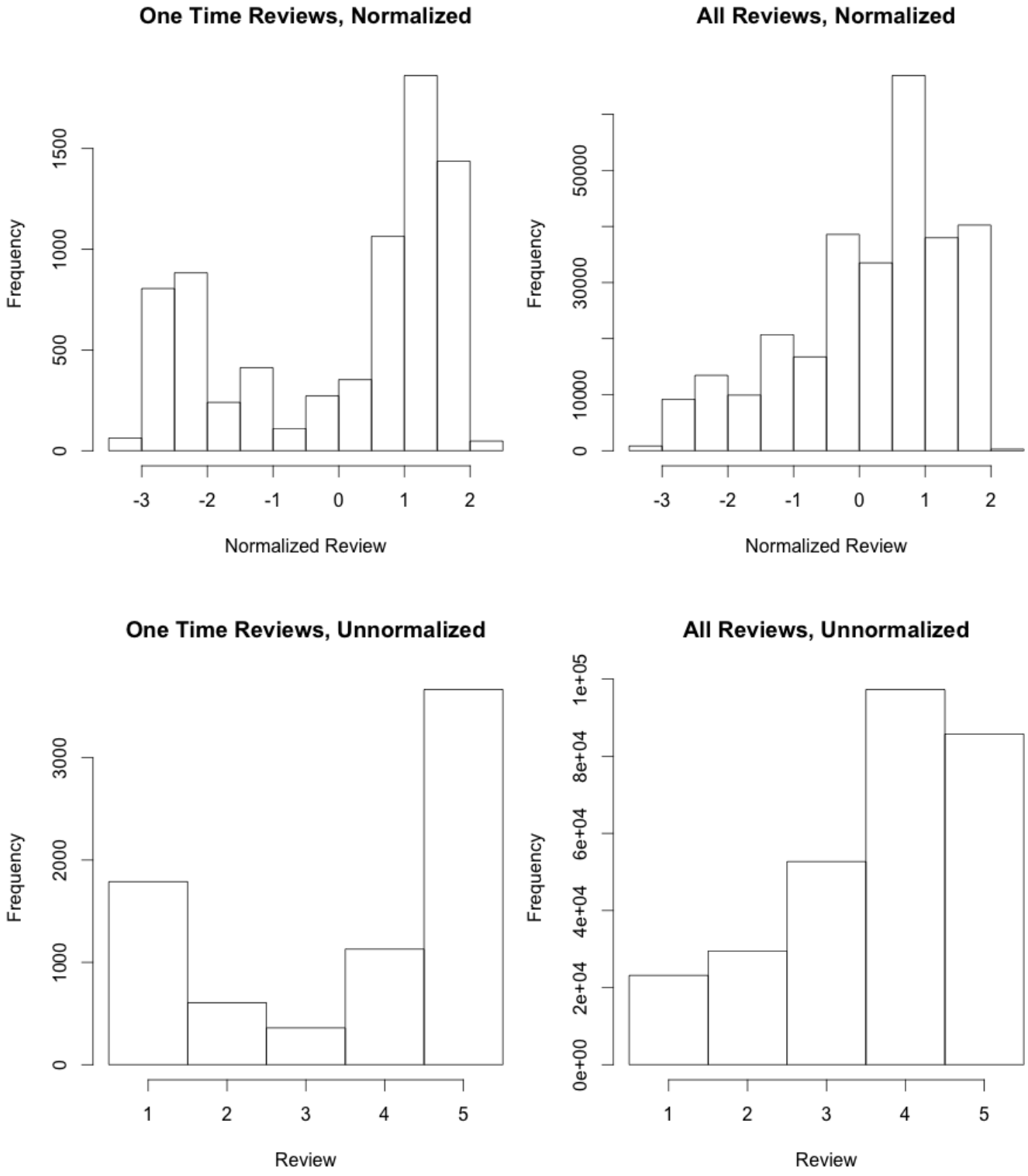
N	12204
R^2	0.085
$F_{(31,12172)}$	36.417

(hotel or restaurant) are almost always businesses of the other type that decided to list event planning as one of their functions.

Distribution of One Time Reviewers

We have included several histograms in Figure 2; the first depicting the distribution of normalized one time reviews and the second depicting the distribution of all normalized reviews, while the third and fourth depict the equivalent unnormalized reviews. In comparing the first two histograms, we see that the histograms themselves very strongly support our hypothesis that first time reviews are polarized; note that there is a much larger peak at around -2.5 as well as a far lower valley at around 0 in the distribution of one time reviews than compared to all reviews.

Figure 2: A Comparison of the Distributions of One Time Reviewers and All Reviewers



This result was further validated by the Kolmogorov-Smirnov test, which yielded a p-value less than $2.2 \cdot 10^{-16}$. As previously noted, this p-value is in fact approximate because of the presence of ties [in fact, there was a non-trivial number of ties: for example all restaurants reviews that had the same star ratings would have tied]; however, this makes our p-value conservative. So seemingly if we had more powerful tools available to us, our p-value would be even smaller.[17]

This result thus suggests that a typical Yelp user's behaviour changes at some time after they make their first review; in particular, they give out less very poor ratings and more 'middling' ratings (3's and 4's), while the number of very good ratings only drops slightly. While the Kolmogorov-Smirnov test does not tell us exactly how or where the distributions are different, we can see from the histograms that there is a fairly reasonable peak at -2.5 on the one time reviews histogram but no peak of equivalent size on the left tail of the histogram of all reviews. This suggests that much more abnormally negative reviews are given out by one time reviewers. Similarly, there is a valley at around -.5 in the one time reviews, but no such valley in the histogram of all reviews, suggesting that one time reviewers are much less likely to give reviews around the mean of 3.6. Finally, we see that there is a sharp peak at around 1.5 for the one time reviews, while there is a corresponding peak at around 1 for all reviews. This suggests that one time users are perhaps slightly more likely to give a very favourable review than a typical reviewer, but without some method of comparing portions of distributions[18] it is difficult to quantify how much more likely this is.

We also note that in the un-normalized case, the histograms look fairly similar, at least in the sense that the one-time reviewers are much more likely to give very good or very poor ratings than the typical reviewer. A Kolmorogov-Smirnov test on these distributions gives us a very similar p-value to above; however, we are hesitant to strongly interpret this p-value due to the immense number of ties [all 5 star reviews are tied with each other, as are all 4 star reviews, etc.].

If we reject the null hypothesis that these two samples were drawn from different distributions [which seems valid, given the stark differences in the histograms], then this corroborates our results from above; namely, it suggests that the behaviour of Yelp users changes at some point after their first review.

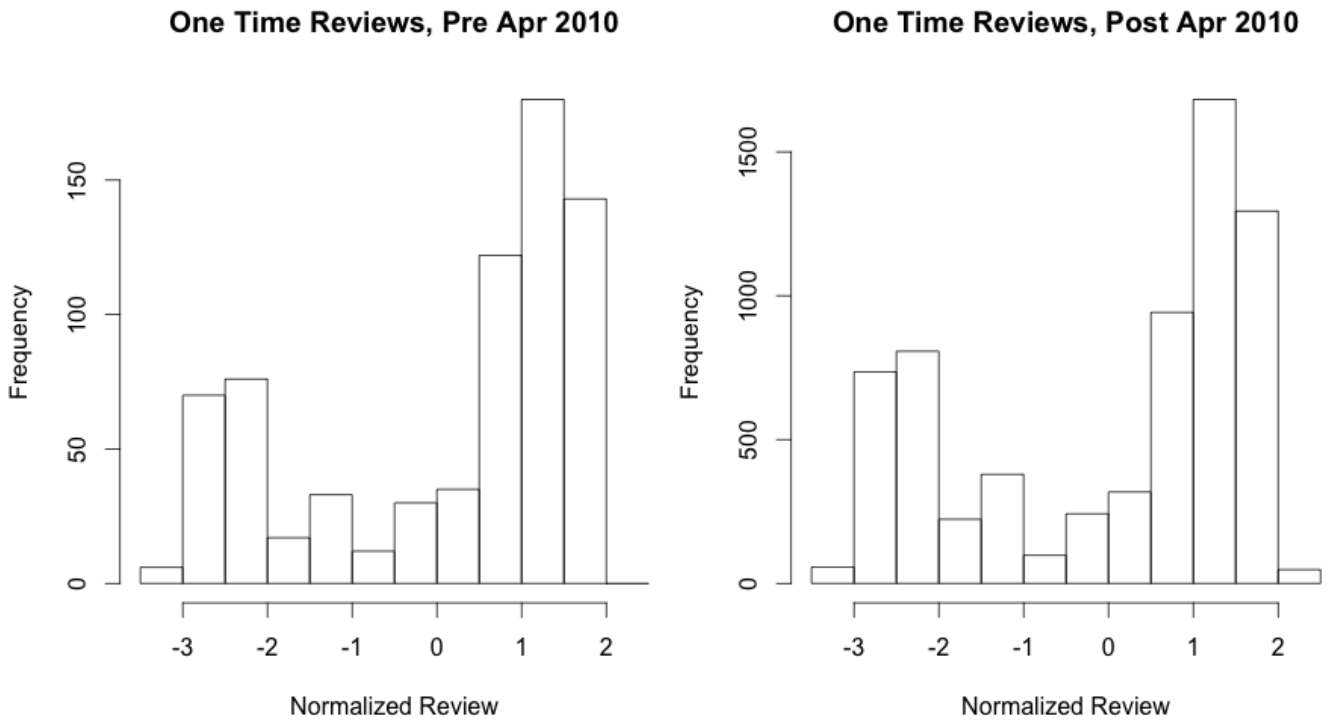
To now discuss the effect of Yelp's account creation policy change in April 2010, we present two histograms in Figure 3 of the distribution of one time reviewers before and after the policy change. These histograms are nearly identical; a Kolmogorov-Smirnov test does not find sufficient evidence to reject the null hypothesis that the two were sampled from different samples with a p-value equal to 0.1126. This result is conservative due to the presence of ties, but rejecting the null hypothesis seems unreasonable regardless.

This presents interesting interpretation of our original hypothesis. Originally, we posited that the increased ease in which users could sign up would cause the distribution of reviewers who signed up after April 2010 to be closer to that of the typical Yelp user, as the lower fixed cost of creating an account would lead to a lower 'feeling' threshold for such account creation [that is to

[17] The same paper referenced previously suggests a test called the Cramer-von Mises test. This test is theoretically more powerful than the Kolmogorov-Smirnov test, however, as we have [seemingly correctly] rejected the null hypothesis using the Kolmogorov-Smirnov test, further application of Cramer-von Mises should not change our result.

[18] It perhaps is possible to re-run Kolmogorov-Smirnov on the two data sets restricted to certain subsamples, however, this seems artificial [for example, arbitrary cutoff points must be chosen] and unnecessary.

Figure 3: A Comparison of the Distributions of One Time Reviewers Before and After Yelp's Policy Change



say, an individual would not have to have as strong an opinion on a business to go through the effort of making an account and reviewing that business after April 2010]. However, this seems to not be the case; suggesting one of the following: (1) the original 'feeling' threshold was not particularly high, so easing the account creation process slightly did not attract significantly different users,[19] (2) the announcement was either not heard or not well received, i.e., people did not want to connect their Facebook accounts to Yelp.

We again reiterate the caveat that it is possible that our sample is biased, as these one time reviews were made by people who choose not to return to Yelp (or at the very least choose not to review again). If we can make an assumption of homogeneous Yelp users, this problem is mitigated slightly; alternatively, if we had access to complete user information rather than just a sample of their reviews we could more concretely make statements about the distribution of first reviews compared with later reviews.

Effect of Prior Reviews

The results from our first OLS regression of normalized reviews in 2010, 2011, and 2012 as a function of number of prior reviews are found in Figure 4.

We see the coefficient on number of reviews [either general reviews or specifically restaurant reviews] was positive and significant, that is to say, the additional 'experience' from having more previous reviews was reflected in more positive reviews in the later years. This effect persists through all three years with relatively

[19]Certainly such a threshold must be non-zero, or we would expect the distribution of one-time reviewers to be the same as all users

Figure 4: Regression Results for Mean Review in a Fixed Year as a Function of Prior Reviews

(a) Effect of all prior reviews on difference between mean normalized review in 2012 and prior mean normalized review

	Dependent variable:
	diff_total12
num_rev_pre2012	0.003**
	(0.001)
Constant	−0.003
	(0.014)
Observations	11,375
R^2	0.0005
Adjusted R^2	0.0004
Residual Std. Error	1.311 (df = 11373)
F Statistic	5.483** (df = 1; 11373)
Note:	*p<0.1; **p<0.05; ***p<0.01

(b) Effect of prior restaurant reviews on difference between unnormalized mean restaurant review in 2012 and unnormalized mean restaurant review prior

	Dependent variable:
	diff_rest12
num_rev_rest_pre2012	0.008**
	(0.003)
Constant	−0.009
	(0.021)
Observations	5,560
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	1.292 (df = 5558)
F Statistic	5.565** (df = 1; 5558)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Effect of all prior reviews on difference between mean normalized review in 2011 and prior mean normalized review

	Dependent variable:
	diff_total11
num_rev_pre2011	0.003*
	(0.001)
Constant	−0.019
	(0.015)
Observations	10,067
R^2	0.0004
Adjusted R^2	0.0003
Residual Std. Error	1.296 (df = 10065)
F Statistic	3.747* (df = 1; 10065)
Note:	*p<0.1; **p<0.05; ***p<0.01

(d) Effect of prior restaurant reviews on difference between unnormalized mean restaurant review in 2011 and unnormalized mean restaurant review prior

	Dependent variable:
	diff_rest11
num_rev_rest_pre2011	0.008**
	(0.004)
Constant	−0.021
	(0.023)
Observations	4,903
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	1.306 (df = 4901)
F Statistic	4.816** (df = 1; 4901)
Note:	*p<0.1; **p<0.05; ***p<0.01

(e) Effect of all prior reviews on difference between mean normalized review in 2010 and prior mean normalized review

	Dependent variable:
	diff_total10
num_rev_pre2010	0.003*
	(0.002)
Constant	−0.039**
	(0.018)
Observations	6,855
R^2	0.0005
Adjusted R^2	0.0003
Residual Std. Error	1.285 (df = 6853)
F Statistic	3.129* (df = 1; 6853)
Note:	*p<0.1; **p<0.05; ***p<0.01

(f) Effect of prior restaurant reviews on difference between unnormalized mean restaurant review in 2010 and unnormalized mean restaurant review prior

	Dependent variable:
	diff_rest10
num_rev_rest_pre2010	0.010**
	(0.004)
Constant	−0.059**
	(0.027)
Observations	3,314
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	1.256 (df = 3312)
F Statistic	4.602** (df = 1; 3312)
Note:	*p<0.1; **p<0.05; ***p<0.01

comparable effects throughout; in general, it seems that one additional (general) review caused a 0.003 increase in normalized rating, and one additional restaurant review caused somewhere between a 0.008 and 0.01 increase in restaurant rating.

We should perhaps be hesitant to strongly reject the null hypothesis of no effect in the general case; as previously noted, there are issues with normalization due to the changing proportions of categories, and in the years 2011 and 2010 our coefficient on the number of reviews prior to that year was only significant at the 10 percent [but not the 5 percent] level. For restaurants, however, because we do not have a normalization problem and our results were consistently significant at the 5 percent level, we should reject the null hypothesis that there is no effect of additional previous ratings.

We also would like to point out that in 2010, the constant term for both regressions was negative and significant at the 5 percent level. This suggests that there was some significant difference between how users rated businesses [both in general and for restaurants specifically] between the years 2004-2009 and in 2010. However, this constant term goes away in our later regressions, suggesting that user behaviour stabilized.

Conclusion

Limitations of this Paper

As mentioned throughout this paper, much of our analysis is limited by the data that Yelp provided in its academic data set. Theoretically, all of Yelp's data is publicly available online, however, using a scraper to collect the millions of reviews is against Yelp's terms of service.[20] As such, this paper could be improved by a few additions to the dataset, namely: (1) the inclusion of *all* reviews review objects for each associated user object, as well as their dates and times, so that the first review for each user could be determined, (2) further information about users, including but not limited to when the account was created, rough demographics, etc., (3) the inclusion of more recent data.

In addition to the expansion of data, this paper could be expanded with a time-series consideration of the normalizing process. That is to say, as different categories reflect different percentages of reviews in each particular year, it would be more rigourous to generate a larger set of normalizing coefficients for reviews in a particular category *and* a particular year. While this seems relatively straightforward [by repeating the methodology of the normalizing section, but with only considering reviews in one fixed year], using Yelp's academic dataset would lead to sample size problems, particularly in the earlier years. Also, a time-series consideration of the effect of additional reviews prior to a fixed year has also been touched on and would be very useful to fixing some potential endogeneity issues in our OLS model.

Furthermore, we would also suggest more rigourous methodology when comparing distributions. While the two-sample Kolmogorov-Smirnov is sufficient for a rough comparison of two distributions, it certainly does run into issues when ties exist, as they frequently do in our normalized data set. More powerful tests suggested include but are not limited to Cramer-von Mises or clever applications of Mann-Whitney-Wilcoxon, however, the formalization of such tests are beyond the scope of this class.

Finally, throughout this paper, we have ignored the effect of region on our reviews. It is feasible, but not easy, to include indicators on the rough geographic location of each business and include those in the normalization process, which in theory should give more accurate results about user behaviour. Similarly, users

[20] see section 6.B.iii of **Yelp's terms of service.**

are assumed to be 'the same'; Yelp does not require users to provide demographics, but perhaps controlling for straightforward characteristics such as age, sex, and race would be appropriate.

Possible Causal Explanations of Results

Because much of this investigation relies on observations of user behaviour, it is difficult to formally attribute such behavioural patterns to consequences of some mathematical model. As such, we propose a few rational explanations for the patterns we observed in the user behaviour that seem to line up with our results. It should be noted that these explanations are by no means exhaustive, and it is difficult to isolate the effects of each result to test experimentally.

We will omit a thorough discussion of the normalization results; for the most part, they seem to reflect that people tend to enjoy art or local attracts but poorly review businesses where they could have had a particularly poor experience [for example, it is very possible to have a bad experience at a hotel or with a real estate agent].

Our results from the remaining two sections seem to indicate (under the assumption that one time reviewers are in some way analogous to first reviews) that Yelp users face a significant fixed cost when signing up. As such, they in general need to feel particularly strongly when initially signing up, leading to their first reviews being particularly polar. However, we notice that when this fixed cost is decreased, there is very little significant change in the distribution of their first reviews, suggesting that this fixed cost decrease was relatively insignificant (the April 2010 announcement was either poorly received or potential users were hesitant to link their Facebook accounts).

Furthermore, our results suggest that Yelp users either learn over time or get nicer. These are two separate but not necessarily disjoint possible causes; the first reflects the idea that Yelp users may, after posting their initial few reviews, notice that they are reviewing in some way that is significantly different from the typical Yelp user and adjust their future reviews, while the second reflects the idea that long time users are, quite simply, more likely to give generous reviews. We should note that our data suggests that Yelp as a whole was giving out less generous reviews in later years [namely, the constant term in our 2010 regressions was negative, suggesting that the typical review was lower in 2010 than in previous years], so it is possible to modify our 'nicer' hypothesis to perhaps reflect that such users are either getting nicer or not getting as critical as the Yelp userbase was.

Acknowledgements

We would like to thank Winnie van Dijk and Hyunmin Park , without whom this project would not have been possible. We would also like to thank Jonathan Dingel and Matt Taddy for pointing us in the right direction.

References

[1] Luca, Michael. "Reviews, Reputation, and Revenue: The Case of Yelp.com." Harvard Business School Working Paper, No. 12-016, September 2011.

[2] Luca, Michael, and Georgios Zervas. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." Working Paper. May 2015.

[3] Arnold, Taylor and Emerson, John. "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions". The R Journal, Vol 3/2, December 2011.