Uncovering Differences in Ratings and Reviews between Yelp and Google

Hanlin Li
lihanlin@u.northwestern.edu

Introduction

Online review platforms are ubiquitous and powerful tools that shape people's consumption behaviors. These systems reflect how a crowd of consumers evaluate a business and subsequently affect how their in-platform location recommender systems direct future consumer foot traffic. For example, if a restaurant has a low rating on Yelp, it is unlikely to be recommended as its rating carries great weight when the platform is making a recommendation for consumers. In other words, the impact of ratings is most salient in terms of the locations being recommended.

Despite this growing impact of online ratings and location recommender systems, the complexity of cross-platform differences is often overlooked by both application developers and researchers. Evaluations of cross-platform differences have solely considered design features and their influences on community norms and content but miss the external factors that shape peer production. For instance, certain categories of restaurants or restaurants located in certain areas may be rated highly on Yelp but poorly on Google Maps due to population differences on the two platforms. In other words, cross-platform differences may reflect and perpetuate population bias that stems from user generated reviews.

Thus, in this paper, we aim to address this important gap in the literature. To do so, we investigate two prominent review platforms, Yelp and Google Maps. We first compare the differences in ratings and reviews with respect to endogenous and exogenous factors. Endogenous factors include restaurant properties, i.e. category, price level, and whether the business is claimed on Yelp by its owner. Exogenous factors include demographics that characterize a restaurant's neighborhood, including education level, poverty, and racial diversity. We then examine the implications of such rating and review differences in location recommendation.

Specifically, we ask the following research questions:

1) Rating differences
    a. Do rating differences vary with respect to endogenous variables?
    b. Do rating differences vary with respect to exogenous variables?
2) Review differences
    a. Do differences in the number of reviews vary with respect to endogenous variables?
    b. Do differences in the number of reviews vary with respect to exogenous variables?

In uncovering the geographic and topical nature of cross-platform differences, we highlight the importance of considering population differences in research and applications of location recommender systems. Furthermore, we demonstrate the importance of considering cross-platform differences in studying the social and economic implications of rating system

Dataset and Methods

We constructed our dataset from two popular review platforms - Yelp and Google Maps. The Yelp data is made publicly available by Yelp. It contains restaurants located in Phoenix (AZ), Las Vegas (NV), Cleveland (OH), Urbana-Champaign (IL), Madison (WI), Pittsburgh (PA), and Charlotte (NC). We filtered out restaurants that receive fewer than ten reviews as these restaurants are often newly-opened with less reliable average ratings. In the end, our Yelp data consisted of 25316 restaurants and 2881871 reviews.

We then constructed a matched Google dataset based on these restaurants. We used a python script to search the combination of each restaurant's Yelp name and address on Google Maps. We were able to retrieve 69% restaurants from Google Maps. Through manual inspection, we determined that most of the missing restaurants are either permanently closed or have addresses that do not lead to a result on Google Maps.

With respect to analysis methods, we first provide an overview of our dataset including summary statistics and distributions of three types of variables. We then move on to specific statistical tests and regression models to answer our research questions.

Results

RQ1: Rating Differences.

We calculate rating difference as a restaurant's Google rating being subtracted from its Yelp rating. Using this definition, we find that Google ratings are inflated compared with Yelp ratings, with the majority (93%) of restaurants having a positive rating difference value.

Endogenous Variables

H1.1$_0$: Rating differences do not vary with respect to category.

As a restaurant may belong to multiple categories, we first create an expanded dataset in which each restaurant appears as many times as the number of categories as it is labelled with. This dataset is only used for our analysis of restaurant categories. We first conducted ANOVA to test this hypothesis. Focusing on the most popular restaurant categories, our F-statistics(33, 39469)=108.55 ($p<0.01$) rejects the null hypothesis.

Table 1: ANOVA on restaurant category

	SUM_SQ	DF	F	PR(>F)
CATEGORY	832.20	33.00	108.55	0.00
RESIDUAL	9,168.96	39,469.00		

The plot below further demonstrates the categories whose rating differences are statistically different from the overall distribution using t-test. Here, we see that Fast Food, burgers, chicken wings, and Tex-Mex face greater rating differences across platforms. Conversely, bakeries, desserts, Mediterranean, and Specialty Food exhibit smaller rating differences than the overall dataset. Overall, the extent of disagreement that Google reviewers and Yelp reviewers have varies across restaurant categories.

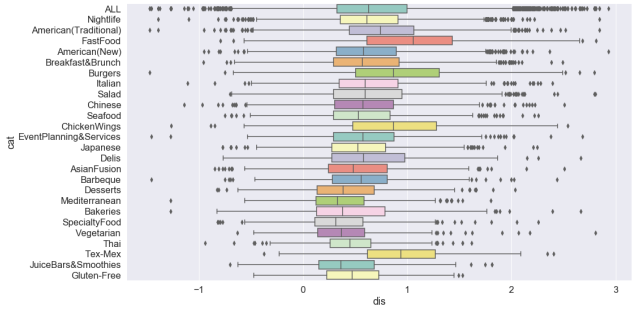


Figure 1: Rating differences across categories

H1.2$_0$: Rating differences do not vary with respect to price level.

Similarly, we conduct ANOVA on restaurant price level and fail to reject the null hypothesis.

Table 2: ANOVA on price level

	sum_sq	df	F	PR(>F)
C(price)	25.81	3.00	33.69	0.00
Residual	4,585.73	17,959.00		

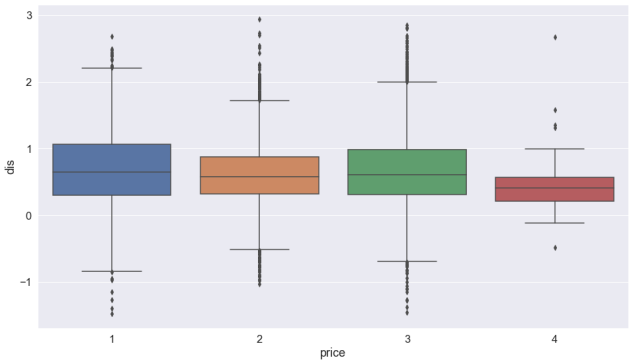


Figure 2: Rating differences across price level

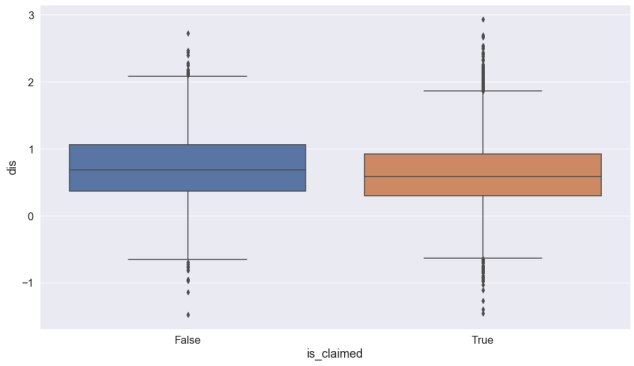
H1.3$_0$: Rating differences do not vary with respect to claim status.

We observe a small effect of being claimed on Yelp on rating differences (cohen's d=-0.20), meaning that a claimed business on Yelp faces a smaller rating difference.

Table 3: ANOVA on claim status

	sum_sq	df	F	PR(>F)
is_claimed	13.18	1.00	52.99	0.00
Residual	2,477.65	9,964.00		

Table 4: Rating differences across claim status



H1.4$_0$: Rating differences do not vary with respect to exogenous variables, including education level, racial diversity, and poverty rate.

We include endogenous variables that we have tested above in a regression model and compare whether exogenous factors matter in predicting a restaurant's rating difference. Overall, we find a trivia effect of education level – with a neighborhood's education level increases 50%, its restaurants will face 0.04 decrease in their cross-platform rating differences. There seems to be no significant relationship between rating difference and racial diversity and no strong effect between poverty rating and rating difference.

	coef	std err	t	P>|t|	[0.025	0.975]
is_claimed[T.True]	-0.0854	0.011	-8.129	0.000	-0.106	-0.065
edu_bachelor	-0.0849	0.021	-4.070	0.000	-0.126	-0.044
non_white	0.0083	0.021	0.403	0.687	-0.032	0.049
PovertyRate	-0.0021	0.000	-8.924	0.000	-0.003	-0.002

RQ2: Differences in the number of reviews

Due to the long tail distribution of numbers of ratings (Yelp: Mean=114, Median=48, Min=10, Max=7,968; Google Maps: Mean=254, Median=165, Min=10, Max=5639, we log transoformed all restaurants' number of reviews and used z-scores for the following comparison. We calculate differences in the number of reviews per restaurant as its z-score of its log-transformed number of reviews on Google being subtracted by its counterpart on Yelp.

H2.1$_0$: Differences in the number of reviews do not vary with respect to category.

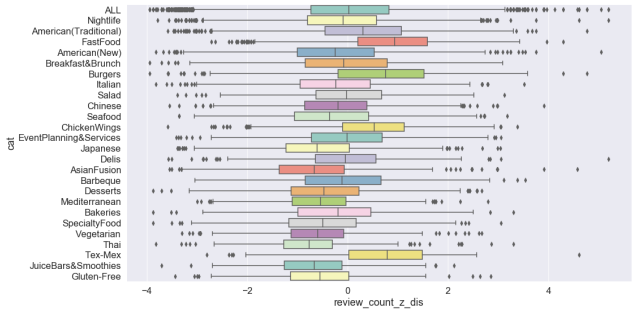
Similar to rating differences, we observe that differences in the number of reviews vary across categories, with our F-statistics failing to reject the null hypothesis.

Table 5: ANOVA on restaurant category

	SUM_SQ	DF	F	PR(>F)
CATEGORY	5,624.83	33.00	149.51	0.00
RESIDUAL	44,996.91	39,469.00		

Interestingly, we see that, again, Fast Food, Burgers, Chicken Wings, and Tex-Mex exhibit the greatest differences in z-scores, meaning that Google reviewers review these restaurants more than Yelp reviewers. Conversely, Yelp reviewers review Asian Fusion, Thai, and Juice Bars more than Google reviewers.

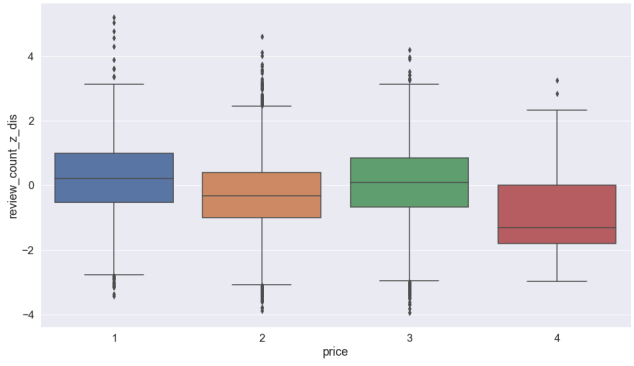
Table 6: Review differences across restaurant categories



H2.2$_0$: Differences in the number of reviews do not vary with respect to price level.

Again, we fail to reject this null hypothesis and find statistically significant differences among restaurants with varied price levels. Specifically, restaurants that are priced as 1 and four exhibit the greatest differences (cohen's d = -0.19 and 0.76, respectively). Yelp reviewers give the highest priced restaurants more reviews than Google reviewers, whereas Google reviews give the lowest priced restaurants more reviews than Yelp reviewers.

	sum_sq	df	F	PR(>F)
C(price)	702.49	3.00	187.02	0.00
Residual	22,485.68	17,959.00		



H1.3$_0$: Differences in the number of reviews do not vary with respect to claim status.

We observe statistical significant differences in numbers of reviews across the restaurants that are claimed on Yelp vs. the ones that are not claimed. More importantly, restaurants that are claimed on Yelp have a negative mean value of review differences, meaning having a relatively large number reviews on Yelp in comparison with Google. In contrast, restaurants that are not claimed on Yelp have a positive mean value of review differences, indicating that these restaurants have a relatively small number of reviews on Yelp.

Table 7: ANOVA on claim status

	sum_sq	df	F	PR(>F)
is_claimed	143.95	1.00	108.93	0.00
Residual	13,166.99	9,964.00	nan	nan

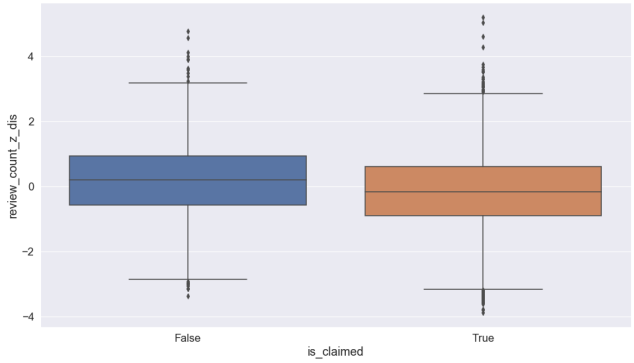


Figure 3: Review differences across claim status

H1.4$_0$: Differences in the number of reviews do not vary with respect to exogenous variables, including education level, racial diversity, and poverty rate.

Building upon previously mentioned endogenous variables, we further construct a linear regression model to investigate whether exogenous variables covary with differences in the number of reviews.

Surprisingly, we see that with the neighborhood's education level increase, the difference in the numbers of reviews decrease, meaning restaurants have more reviews on Yelp. Racial diversity presents a similar effect in this model as well. Both relationships are consistent with prior work that shows Yelp reviews concentrate on neighborhoods with higher education backgrounds and a dominant percentage of white residents.

	coef	std err	t	P>|t|	[0.025	0.975]
edu_bachelor	-0.7466	0.047	-16.030	0.000	-0.838	-0.655
non_white	-0.8009	0.046	-17.310	0.000	-0.892	-0.710
PovertyRate	0.0033	0.001	6.350	0.000	0.002	0.004

Conclusion

Using over seventeen thousands restaurants in seven metropolitan areas in the U.S., this study presents alarming differences in ratings and numbers of reviews across Yelp and Google. Moreover, these differences vary with respect to both endogenous factors such as restaurant category and exogenous factors such as education level. Our study underscores the importance of considering population differences in research and applications of rating systems.