3 Stars on Yelp, 4 Stars on Google Maps: A Cross-Platform Examination of Restaurant Ratings

Hanlin Li, Northwestern University, Evanston IL, USA
Brent Hecht, Northwestern University, Evanston IL, USA

Even though a restaurant may receive different ratings across review platforms, people often see only one rating during a local search (e.g. "best burgers near me"). In this paper, we examine the differences in ratings between two commonly used review platforms—Google Maps and Yelp. We found that restaurant ratings on Google Maps are, on average, 0.7 stars higher than those on Yelp, with the increase being driven in large part by higher ratings for chain restaurants on Google Maps. We also found extensive diversity in top-ranked restaurants by geographic region across platforms. For example, for a given metropolitan area, there exists little overlap in its top ten lists of restaurants on Google Maps and Yelp. Our results problematize the use of a single review platform in local search and have implications for end users of ratings and local search technologies. We outline concrete design recommendations to improve communication of restaurant evaluation and discuss the potential causes for the divergence we observed.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**

KEYWORDS

Restaurant reviews; local search; multi-site research; user-generated content.

ACM Reference format:
Hanlin Li and Brent Hecht. 2020. 3 Stars on Yelp, 4 Stars on Google Maps: A Cross-Platform Examination of Restaurant Ratings. In Proceedings of the ACM on Human-Computer Interaction, CSCW, Article XX, November 2020. ACM, New York, NY, USA. 23 pages.

1 INTRODUCTION

Restaurant ratings are both popular and influential. A number of restaurant review platforms, e.g., Yelp, Google Maps, TripAdvisor, Facebook, and Dianping, have become a critical source of information for restaurant patrons, surpassing newspapers and word-of-mouth in importance [16,25]. The user-provided ratings on these platforms also power local search technologies such as Google Search, where "restaurants near me" is the most popular local search query over the past five years [39]. Moreover, a restaurant's average rating has been shown to have a significant effect on the restaurant's revenue [17].

Despite of the existence of multiple prominent review platforms, many local search technologies often highlight average ratings from only one review platform. Google Search elevates its own review platform, Google Maps, whereas Google's competitors often use ratings from Yelp. In doing so, local search technologies inherently adopt a **universal assessment assumption**, i.e. that one platform's average rating can suitably represent reviews on all platforms

In reality, however, the same restaurant may receive different ratings across review platforms. An example can be seen in Figure 1, which depicts a Panera Bread restaurant shown in Google

https://doi.org/124564

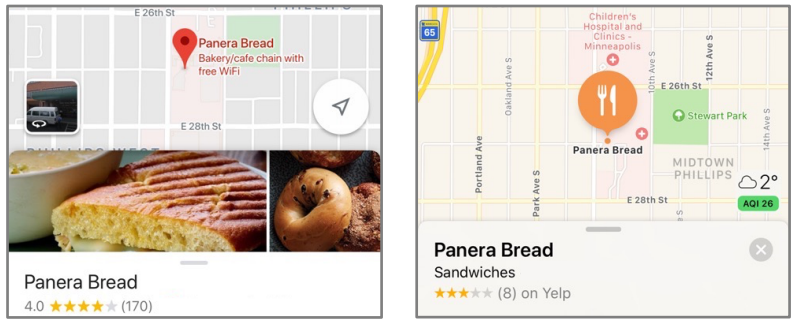

Figure 1: The same restaurant's information on Google Maps (on the left with its average Google Maps rating) and Apple Maps (on the right with its average Yelp rating).

Maps (left) and Apple Maps (right). A Google Maps user will see this restaurant with four stars (out of five) , while an Apple Maps user will see the same restaurant with three stars (out of five).

By displaying one review platform's rating, local search technologies exhibit the universal assessment assumption and thereby may inadvertently prevent users from seeing the full picture about a restaurant. In the example above, users of Google Maps and Apple Maps miss out very different restaurant evaluations and may make different decisions on whether to patronize this Panera Bread. That being said, it is possible that the restaurant's ranking on Google Maps is identical with its ranking on Apple Maps, and this may reduce the degree to which the absolute value of average rating makes a difference for real-world decisions.

To understand the implications of the universal assessment assumption, we investigated the magnitude of absolute and relative cross-platform differences in restaurant evaluation between two popular restaurant review platforms: Google Maps and Yelp. These two platforms not only have millions of users [40,41], but also provide ratings to widely-used local search technologies such as Google Search, Bing, and Apple Maps [42,43]. We gathered and analyzed parallel Google Maps and Yelp ratings for 21841 restaurants across seven metropolitan areas in the U.S.

Overall, we found evidence that problematizes the universal assessment assumption. 93% of the restaurants in our dataset have average Google Maps ratings that are higher than their corresponding average Yelp ratings. The mean difference in average rating between Google Maps and Yelp is 0.7 stars. For 24% of the restaurants in our dataset, their average Google Maps ratings are at least one star higher than their Yelp counterparts. We observed that these cross-platform differences are driven in large part by chain restaurants: on average, chain restaurants are rated 1.1 stars higher on Google Maps than on Yelp, whereas the equivalent figure for independent restaurants is 0.6 stars.

As a step toward understanding the implications of the flaws in the universal assessment assumption on local search results, we also investigated to what extent top-ranked restaurants differ across platforms. We saw little agreement in the top-ranked restaurants. For instance, the top-ten lists for the majority of the metropolitan areas in our dataset have only one restaurant in common across the two platforms.

We close the paper by discussing our results' implications for end users and local search technologies, as well as providing concrete design recommendations to better inform users of potentially diverse restaurant assessments. Additionally, building upon prior work, we discuss the potential causes for the cross-platform differences we observed. Finally, we highlight how our findings support social computing researchers' growing calls for multi-site studies [4].

2 RELATED WORK

In this section, we first discuss two lines of literature that provide important foundation for our work: the literature on restaurant reviews and the literature on hotel reviews. Importantly,

both bodies of literature provide us with methodological guidance and potential factors that may influence rating production to help us interpret our findings.

Additionally, we specifically highlight another factor that is well investigated by the research community and has an impact on average ratings: the handling of fraudulent reviews.

2.1 Restaurant Reviews

Research on restaurant reviews has identified review platforms as a critical source of information for restaurant patrons [16,25]. Restaurant reviews strongly influence people's restaurant choices, and therefore have a real-world impact [1,17]. For example, Luca and colleagues found that a 0.5-star increase in average Yelp rating causes an independent restaurant to have a 9% increase in revenue [17].

The vast majority of the literature on restaurant reviews has focused on characterizing ratings within a single review platform. Bakhshi and colleagues found restaurant ratings on CitySearch are correlated with price tier, geographic features, and weather [2]. Similarly, Jurafsky found that star ratings are correlated with restaurant categories and price tier on Yelp [11]. These studies led to our use of restaurant category and price tier in our characterization of cross-platform differences in average rating.

Fewer papers have taken a cross-platform lens to restaurant reviews. Wang examined reviews from Yelp, CitySearch, and Yahoo Local and found that Yelp reviewers are more prolific than reviewers on other platforms [31]. The paper provided the first hint that challenges the universal assessment assumption. Given the same set of restaurants, CitySearch and Yahoo Local have larger shares of one- and five-star ratings than Yelp in aggregate. Our study picks up where Wang left off by measuring and characterizing differences in average rating for a set of restaurants across platforms. Moreover, we consider the introduction of newer, more popular review platforms, such as Google Maps, and include a comparison of restaurant rankings.

Previous work has also highlighted several design features that may have influenced people's rating behaviors. For example, Wang suggested that reputation features such as special recognition of high-quality reviews create more productive reviewers, who subsequently give lower ratings than novice reviewers [29,31]. Similarly, Kang and colleagues found that anonymity allows reviewers to provide more honest, negative feedback, as it helps them avoid reactions from the reviewed business and people with opposing views [12]. These studies helped us interpret our findings and identify potential causes for cross-platform differences in average rating.

2.2 Hotel Reviews

In contrast with the lack of cross-platform studies of restaurant reviews, a considerable amount of research on hotel reviews has taken a cross-platform approach. Many of these studies have focused on the linguistic features of hotel reviews [7,15,24,34,35]. For example, Xiang et al. compared hotel reviews from Yelp, TripAdvisor, and Expedia and found that sentiments on TripAdvisor and Expedia tend to be more positive than those on Yelp [32].

A few studies of hotel reviews found that the universal assessment assumption may not be true for hotels cross-listed on multiple review platforms and suggested various potential causes. For example, Zervas and colleagues observed that hotels rarely have consistent average ratings between Airbnb and TripAdvisor, and suggested that the two platforms' different reviewer populations with potentially divergent aggregate preferences may be a reason [34]. Similarly, Eslami and colleagues cautioned that a hotel might be rated differently across Booking.com, Hotels.com, and Expedia.com due to differences in platform design [7]. These studies provided us with important context to interpret the observed cross-platform differences in our study.

The studies on hotel reviews also provide methodological guidance to our paper [7,34]. For instance, Zervas and colleagues examined the rankings of 1959 hotels that are cross-listed on

Airbnb and TripAdvisor and found them to be only weakly correlated [34]. In our work, we used a similar method by calculating the rank correlation between Google Maps and Yelp as part of a larger suite of analyses. Our work also takes inspiration from Eslami et al. [7], which compared 803 hotels' absolute average ratings on Booking.com, Hotels.com, and Expedia.com and found different minimum values in the distribution of average ratings across platforms. We examined whether this phenomenon—or an analogous one—occurs in Google Maps and Yelp for restaurant ratings.

2.3 Fake Reviews

Both Google Maps and Yelp have a zero-tolerance policy for "fake reviews" [17]—reviews that are not written by real customers. While both platforms allow anyone to flag and report potential fake reviews [13,44], the two platforms have published different amounts of information about how they handle fake reviews. Google Maps states "(Google Maps) may take down reviews that are flagged as fake", but no substantial information about the platform's approach to the removal of fake reviews has been made public.

Yelp, on the other hand, has published multiple official announcements regarding its review filtering algorithm [45], which has been well studied by researchers [33]. The algorithm aims to automatically detect fake reviews, resulting in 16% of all the reviews submitted to Yelp being flagged [18]. These flagged reviews are not used to calculate a restaurant's average rating and are displayed as "not recommended" in a separate section of the Yelp interface. Although the effectiveness of this filtering algorithm has been questioned by both business owners and Yelp reviewers [8], Luca and colleagues provided preliminary evidence supporting the algorithm's capability to flag fake reviews [18]. Using a set of restaurants that were caught for soliciting fake reviews in a sting operation by Yelp, they found that the algorithm indeed flagged a larger share of reviews as fake from these restaurants than those that were not known for review fraud [18]. More recently, Mukherjee and colleagues found that the algorithm targets abnormal behaviors such as writing multiple reviews in a day and argued that the algorithm is "at least doing a reasonable job at filtering (fake reviews)" [22].

As fake positive reviews (i.e. fake reviews that are in favor of a business) are shown to be more widespread in online review platforms than fake negative reviews [18] , Yelp's filtering algorithm may potentially lead to its ratings being lower than those on Google Maps. However, given that both algorithms are proprietary (i.e. "black box"), it is impossible to make accurate external assessments of the role of these algorithms in our results. It could be that Google's filtering algorithm is equally or more effective. We discuss this issue further in Discussion.

3 METHODS

This section first details how we collected and processed Google Maps and Yelp ratings for our 21841-restaurant dataset. We then introduce the key metric used to evaluate and describe pairwise difference in average rating. Finally, we unpack how we compared the top-ranked restaurants across platforms. Overall, this paper focuses on average rating and ranking, and we provide additional descriptive statistics about cross-platform differences in number of ratings in the Appendix.

3.1 Data Collection

We studied the two most influential review platforms in the U.S.—Google Maps and Yelp. A 2017 survey showed 81% and 59% of the U.S. population get information about local businesses from Google Maps and Yelp, respectively [43]. Moreover, both review platforms license their ratings to many different rating providers [46,47], including local search technologies used by millions of people in the U.S. and elsewhere, such as Apple Maps, Bing, and DuckDuckGo.

To collect data from Google Maps and Yelp, we faced two methodological challenges that have been highlighted in prior work [4]: (1) gathering a sufficiently large dataset and (2) collecting parallel data about restaurants across platforms.

3.1.1 Gathering a Sufficiently Large Dataset of Ratings

A common challenge in studying user-generated content is to gather a sufficient amount of high-quality data from a platform [4]. In our case, although both Google's and Yelp's API services provide extensive information about restaurants, they do not return all the data we are interested in such as the distribution of star ratings for a given restaurant, which was necessary to calculate the restaurant's ranking.

To obtain more complete data than is possible through the platforms' APIs, we first downloaded the Yelp Open Dataset. This dataset is published by Yelp and contains granular rating information for each restaurant [48]. It includes this information for restaurants located in seven metropolitan areas in the U.S., specifically, Phoenix (AZ), Las Vegas (NV), Cleveland (OH), Urbana-Champaign (IL), Madison (WI), Pittsburgh (PA), and Charlotte (NC). We discuss the limitation of this non-random sample further in Limitations. Following prior work [31,34], we removed restaurants that received fewer than ten ratings, resulting in 25316 restaurants. This filtering allowed us to focus on restaurants whose ratings are likely to be deemed reliable by users of local search technologies [28].

3.1.2 Collecting the Parallel Google Maps Data

After processing the Yelp Open Dataset, we faced the second challenge of data collection: collecting these restaurants' parallel Google Maps ratings. To address this challenge, we followed prior work [31] and developed a script that can acquire a restaurant's listing on Google Maps' interface by searching for the combination of the restaurant's Yelp name and address. The script[2] was built upon the open-source Puppeteer library [37]. It slowly iterated through the 25316 restaurants from the Yelp Open Dataset (one restaurant per 60 seconds) to collect each restaurant's rating distribution on Google Maps while avoiding any undue burden to the platform's server [27].

We were able to retrieve 21945 (87%) restaurants from Google Maps out of the 25316 restaurants from the Yelp Open Dataset. Through manual inspection, we determined that most of the missing restaurants are either permanently closed or have addresses that do not lead to a restaurant listing on Google Maps. Similar to our processing of the Yelp Open Dataset, we removed the restaurants that have fewer than ten ratings on Google Maps, leaving us with 21841 restaurants in total.

We took efforts to verify that our script located the correct restaurant listings on Google Maps using the restaurant names and addresses from the Yelp Open Dataset. From the 21841 restaurants with parallel Yelp and Google Maps data, we sampled 50 restaurants randomly and manually inspected whether each restaurant's Google Maps information truly matches with that of Yelp. We found that 49 (98%) restaurants were correctly matched. The only mismatched restaurant was shown to be permanently closed on Yelp but was incorrectly matched with a Google Maps restaurant listing nearby.

3.2 GoogleMinusYelp = Average Google Maps Rating – Average Yelp Rating

We constructed a very straightforward key metric, *GoogleMinusYelp*, to compare a restaurant's average Google Maps rating with its average Yelp rating. We first calculated a restaurant's average rating on each platform as the arithmetic mean of its rating distribution. We then calculated a restaurant's GoogleMinusYelp as its average Google Maps rating subtracted by

[2] https://github.com/hanlinl/restaurant-rating

its average Yelp rating. For a given restaurant, a positive GoogleMinusYelp value means that its average Google Maps rating is higher than its Yelp counterpart, and vice versa.

Notably, the average Google Maps and Yelp ratings we calculated may not be exactly what Google Maps and Yelp display on their interfaces. Yelp is known to round a restaurant's arithmetic mean in ratings to the nearest half-star for display (e.g. a restaurant with an average rating of 3.7 stars will be shown as 3.5 stars) [17]. Therefore, on Yelp, the gap between a restaurant's display rating and its arithmetic mean may be as large as 0.25 stars. In contrast, Google Maps does not publicize how a restaurant's display rating is calculated. However, we found that Google Maps is most likely to use a restaurant's arithmetic mean rounded to one decimal place; after our data collection, we manually inspected 200 randomly sampled restaurants and found their Google Maps display ratings are their arithmetic means rounded to one decimal place. Although anecdotal evidence suggests that some restaurants' Google Maps' display rating are not their rounded arithmetic means [5], for the restaurants in our dataset, the two metrics should be very close, if not identical.

As a restaurant's display rating and average rating may be somewhat different on Yelp, we conducted our analysis in parallel using both rating metrics. Because using average rating leads to more precise comparison and characterization, below we primarily report results with average Yelp rating and report corresponding key statistics using Yelp display rating in footnotes. As will be seen below, the results from the two calculations are consistent.

3.2.1 Characterization Metrics for GoogleMinusYelp

In addition to identifying any pairwise differences in average rating across platforms (using the GoogleMinusYelp metric), we also wanted to characterize these differences. We investigated if any differences were being driven by a certain type of restaurant in particular. To do so, we classified each restaurant according to restaurant properties that are prominent on review platforms and/or used in prior work on restaurant reviews.

Chain Status: We sought to examine any cross-platform differences through the lens of chain status due to a hypothesis that emerged from anecdotal observation of cross-platform ratings. Restaurant chains have two distinct business models, i.e. the chain model and the franchise model [38]. However, restaurant customers commonly refer to restaurants operated under both models as chain restaurants. As such, in this paper, we consider both models as "chain restaurants" for simplicity.

Determining whether a restaurant is affiliated with a chain is not a trivial task [9]. To address this challenge, we first consulted two market research databases, Mintel [49] and Statista [50]. These two databases list large, national restaurant chains across the U.S. based on their numbers of locations. However, the databases turned out to be far from comprehensive. Following prior work [14], we supplemented these databases with a name-repetition approach that counted how many times each unique restaurant name occurs in our datasets. Any restaurant name that occurred more often than the least-repeated name from the two business databases was treated as indicative of a chain. This process resulted in us classifying any name that appeared at least 18 times as belonging to a chain. We identified 90 chains in total, ranging from McDonald's with 342 locations to Which Wich with 19 locations. Overall, by our definition, 21% of the restaurants in our datasets are chain restaurants.

Category: Following prior work on restaurant ratings [11], we also used a restaurant's categories, i.e. type of food (e.g. Thai) or type of service (e.g. buffet) to characterize GoogleMinusYelp. Although the restaurants in our datasets belong to 176 categories in total, many of the categories are uncommon. Therefore, to simplify our analysis, we only considered the most common 31 categories, to ensure a sufficient number of restaurants (at least 400) for each category in our analysis.

Price tier: Also following prior work [11], we included a restaurant's price tier to examine whether restaurants of all price tiers exhibit an equal amount of GoogleMinusYelp.

3.2.2 Characterizing GoogleMinusYelp

Once our key and characterization metrics had been finalized, the remainder of our investigation consisted of straightforward descriptive analysis, statistical hypothesis tests (e.g. one-way ANOVA), and multivariate linear regression. In particular, our multivariate linear regression model predicts a restaurant's GoogleMinusYelp value using its chain status, categories, and price tier. As a restaurant can belong to multiple categories according to Yelp, we describe each category using a dummy variable, with 1 marking the restaurant as belonging to a given category. The correlations among the independent variables in our model are no greater than 0.4. Furthermore, we calculated the variance inflation factor (VIF) for each independent variable and observed that all the values fall between 1.0 and 1.8, suggesting that multicollinearity is not severe in our model. Similarly, chain status is represented with a binary variable, with 1 marking a restaurant as a chain. Price tier is represented as an ordinal variable, ranging from 1 to 4 based on Yelp's definition.

3.3 Relative Ranking

As local search technologies often use a ranked list to show restaurant results, we wanted to gain insights into how a restaurant's ranking relative to other restaurants may vary across platforms. However, it is impossible to conduct a complete audit of local search results without access to local search technologies' ranking algorithms as well as popular search queries. As a step toward answering this question, we leveraged ranking methods used in prior work [20] with information that is available to the public.

3.3.1 Ranking Approaches

Notably, when ranking star ratings, most applications will rank a restaurant with ten five-star ratings lower than a restaurant with 48 five-star ratings and two one-star ratings, even though the former has a higher average rating. As such, we applied two ranking approaches that have been used in prior work to account for this complication. The first approach is straightforward. Given a restaurant's rating distribution, we simply calculated the lower bound of the distribution's mean on a 95% confidence interval [20]. This lower bound value is then used to determine a restaurant's ranking. Our second approach is recommended by Miller, which ranks a restaurant by the lower bound of its Bayesian approximation to the rating distribution's mean [51]. We observed that the two ranking approaches resulted in metrics that are strongly correlated on both platforms (Google Maps: Spearman's $r=0.98$, $p<0.001$; Yelp: Spearman's $r=0.95$, $p<0.001$). Below we show our results using the first ranking approach, i.e. the lower bound of the distribution's mean on a 95% confidence interval.

3.3.2 Comparing Top Restaurants in a Geographic Region Between Google Maps and Yelp

Our ranking metrics are based on the output of the ranking approach mentioned above but defined at different levels. More specifically, for each platform, we calculated each restaurant's ranking at three levels: overall ranking, geographically-bounded ranking, and geographically-bounded category-specific ranking.

At the most general level, each platform's overall ranking represents a restaurant's ranking among the total 21841 restaurants on that platform.

At a more granular level, we defined a restaurant's geographically-bounded ranking as its ranking among all the restaurants located in its geographic region. We used three different geographic units: metropolitan area (the largest unit), zip code, and census tract (the smallest unit). As a result, a restaurant has three geographically-bounded rankings per platform: rankings in the metropolitan area, in the zip code, and in the census tract. The ranking results for a geographic region can be seen as an attempt to approximate the ranking results of Google Maps and Yelp for that region. For example, our ranking of restaurants in a census tract based on Google

Maps ratings can be seen as analogous to querying top restaurants in that neighborhood on Google Maps.

We did not compare every zip code or census tract's ranked list across platforms and only considered zip codes and census tracts whose numbers of restaurants are above a certain threshold. Because the cross-platform comparison of ranked lists requires a non-trivial number of restaurants per a geographic region, we examined zip codes with at least 60 restaurants and census tracts with at least eleven restaurants (the upper quartiles in number of restaurants per zip code and per census tract, respectively.) Additionally, because the Yelp Open Dataset does not have a good coverage of restaurants for the zip codes and census tracts near the edge of each metropolitan area, this process allows us to filter out these zip codes and census tracts. As a result, 130 zip codes satisfied the requirement and contain 67% of the restaurants in our dataset. 628 census tracts satisfied the requirement and contain 69% of the restaurants.

At the most granular level, we calculated a restaurant's geographically-bounded category-specific ranking as an attempt to approximate local search queries such as "Mexican restaurants near me". We considered the most commonly-searched categories, i.e. Mexican, Chinese, Italian, and seafood according to Google Trends [52], and focused on the geographic regions that have a sufficient number of these restaurants. For example, we selected the zip codes whose number of Mexican restaurants are above the upper quartile among all the zip codes in our dataset. For each of these zip codes, we then produced a ranked list of its Mexican restaurants. Because of the lack of restaurants belonging to the same category in a census tract, we omitted the census tract scale for the category-specific ranking analysis.

After calculating all these ranking metrics, we then compared the top n restaurants from each census tract, zip code, and metropolitan area between platforms. For example, we examined how many restaurants in a zip code's top-five list from Google Maps are different than that from Yelp.

3.3.2.1 Sanity Check for Ranking by Metropolitan Area

We examined the ranked lists displayed on the native user interfaces of Google Maps and Yelp as a sanity check to ensure the basic ecological validity of our ranking approach. We collected the top ten restaurants returned by searching "restaurants near me" on Google Maps and Yelp in private browsing mode for the seven metropolitan areas on which our dataset focuses. We also did the same for the top Mexican, Chinese, Italian, and seafood restaurants. The cross-platform differences in top ten restaurants on the native user interfaces are largely consistent with our own ranking approach. However, it is worth noting that restaurant rankings on Google Maps and Yelp are likely to be influenced by various factors, such as personalization, time, and restaurant traffic. Therefore, more data is required to fully audit the rankings returned by Google Maps and Yelp.

4 RESULTS

We first present our results describing pairwise differences in average ratings between Google Maps and Yelp. We then characterize the differences we observed. Finally, we report how the top-n lists differ across the two platforms.

4.1 Average Rating

We see substantial differences in average ratings between Google Maps and Yelp, providing our first evidence that problematizes the universal assessment assumption. The mean of average Yelp ratings for the restaurants in our dataset is 3.5 stars (Median=3.6, SD=0.7, Min=1.0, Max=5.0) [3], while the equivalent figure for Google Maps is 4.2 stars (Median=4.2, SD=0.4, Min=1.7, Max=5.0). Examining the distributions of average ratings on Google Maps and Yelp, we found

[3] The mean in Yelp's display rating (i.e. average rating rounded to the nearest half star) is 3.5 stars (Median = 3.5, SD=0.75, Median=3.5, Min=1, Max=5).

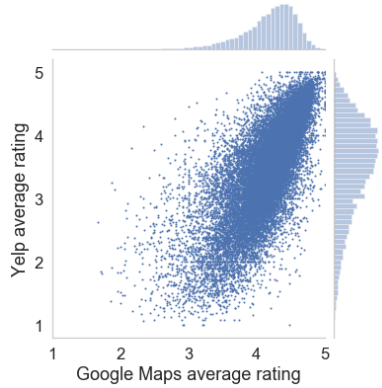

Figure 2: A scatterplot of the 21841 paired average ratings, along with their distributions on Google Maps (at the top) and Yelp (on the right), respectively.

that a four-star average rating is above the 72th percentile on Yelp, but only above the 28th percentile on Google Maps.

In terms of pairwise, restaurant-by-restaurant comparisons of average ratings, we found more evidence that challenges the universal assessment assumption. On average, the restaurants in our dataset are rated 0.7 stars higher on Google Maps than on Yelp. In other words, the mean GoogleMinusYelp metric defined in Methods is 0.7 stars[4] (Median: 0.6, SD=0.5, Min=-1.8, Max=3.5, Q1=0.3, Q3=1.0). 93% of the restaurants in our dataset have a positive GoogleMinusYelp value, and 24% have a GoogleMinusYelp value equal to or greater than one star. In contrast, for only 0.1% restaurants are their average Yelp ratings one star or higher than their Google Maps counterparts.

Figure 2 show that, overall, the restaurants in our dataset have somewhat correlated average ratings across platforms, suggesting that the universal assessment assumption somewhat holds in the ranking of all restaurants. The Pearson's *r* in average rating across two platforms is 0.74 ($p<0.001$)[5] and the Spearman's *r* in overall ranking (a restaurant's ranking relative to all the other restaurants in our dataset as calculated in Methods) is 0.75 ($p<0.001$). Both correlations are well below 1.0 and suggest the higher-rated restaurants on one platform are only somewhat likely to be higher-rated on the other platform as well. Indeed, in our dataset, 20% of the restaurants below the 50th percentile in average Google Maps ratings are above the 50th percentile in average Yelp ratings. As we will see below, this misalignment in ranking becomes worse when it comes to the important scenario of identifying top-ranked restaurants in specific geographic areas.

Figure 2 also highlights that Google Maps and Yelp have substantially different minimum average ratings, a phenomenon that also occurs in hotel review platforms [8]. The lowest-rated restaurant in Google Maps has 1.7 stars, while this figure is 1.0 for Yelp. Interestingly, the potential cause for the disparity in minimum value across hotel review platforms—one platform raising the lowest possible rating a reviewer can give from 1.0 to 2.5—does not apply to our finding here. Google Maps and Yelp have the same lowest possible rating: one star.

[4] When using Yelp's display rating, we found that the cross-platform difference between Yelp display rating and average Google Maps rating remains 0.7 stars on average (Median=0.6, SD=-0.5, Min=-1.8, Max=3.5, Q1=0.3, Q3=1.0).
[5] The Pearson's r between Yelp display rating and average Google Maps rating is slightly lower, 0.72 ($p<0.001$).

Table 1: A multivariate linear regression model to predict GoogleMinusYelp. The ten coefficients with the largest absolute values are shown here. Independent variables include chain status, price tier, and categories. Adjusted R-Squared=0.23

	coef	*std err*	*P>|t|*
chain	0.46	0.01	<0.001
Buffets	0.17	0.02	<0.001
Fast Food	0.12	0.01	<0.001
American Traditional	0.10	0.01	<0.001
Delis	-0.09	0.02	<0.001
Juice Bars Smoothies	-0.09	0.02	<0.001
Desserts	-0.11	0.02	<0.001
Vegetarian	-0.14	0.02	<0.001
Specialty Food	-0.16	0.02	<0.001
Mediterranean	-0.18	0.02	<0.001

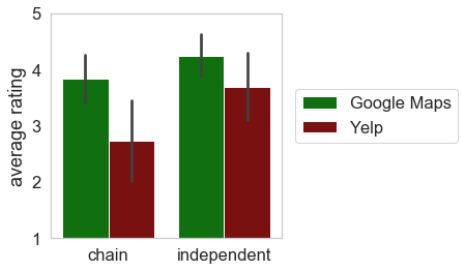

Figure 3: The average ratings from Google Maps and Yelp, faceted by chain status. Error bars indicate standard deviation.

4.1.1 Characterization of GoogleMinusYelp

We found that GoogleMinusYelp has strong associations with certain restaurant properties. Table 1 shows the ten restaurant properties with the largest coefficients in our multivariate linear regression. Below, we unpack the relationships between each restaurant property and GoogleMinusYelp in more detail.

Chain status: As can be seen in Table 1, a chain restaurant's GoogleMinusYelp value is 0.5 stars higher than an independent restaurant in the same price tier and categories. Absent the price and category controls, the difference in mean for GoogleMinusYelp between chain restaurants and independent restaurants is also 0.5 stars. The average GoogleMinusYelp for chain restaurants is 1.1 stars (Median=1.1, SD=0.5, Q1=0.8, Q3=1.4), while it is 0.6 stars (Median=0.5, SD=0.5, Q1=0.3, Q3=0.8) for independent restaurants[6]. Given the nearly equal variance between the two distributions of GoogleMinusYelp, we used a one-way ANOVA to test the significance of their differences and calculated Cohen's d for effect size. We found the difference in GoogleMinusYelp between chain restaurants and independent restaurants is statistically significant ($F(1, 21839)= 4666.1$, $p<0.001$) and of large effect size (Cohen's $d=1.2$).

Figure 3 shows the distribution of chain and independent restaurant ratings across platforms in additional detail. Independent restaurants are rated higher than chain restaurants on both Google Maps ($F(1, 21839)=4105.1$, $p<0.001$) and Yelp ($F(1, 21839)=8685.6$, $p<0.001$). However, relative to independent restaurants, chain restaurants are evaluated differently between two

[6] When we used Yelp's display rating in lieu of average rating for paired, restaurant-to-restaurant comparison, the gap between chain restaurants and independent restaurants remains. Among chain restaurants, Yelp display rating is, on average, 1.1 stars below average Google Maps rating. The equivalent figure for independent restaurants is 0.6 stars.

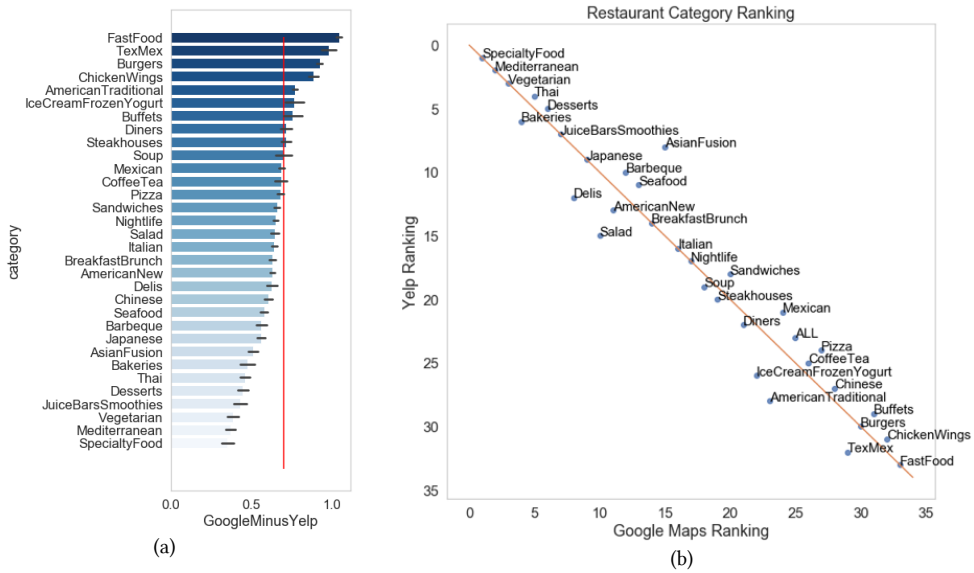

Figure 4: (a) The associations between GoogleMinusYelp and restaurant categories. The red line indicates the mean GoogleMinusYelp value overall (i.e. 0.7). To better estimate each category's mean GoogleMinusYelp value based upon the smaller sample size per category, we calculated the 95% confidence intervals from 1000 bootstrapped replications (marked with error bars).
(b) Category rankings by pooled average rating on Google Maps (x-axis) and Yelp (y-axis).

platforms. On Yelp, the mean of chain restaurants' average ratings is 2.7 stars (Median=2.7, SD=0.7), 1.0 stars below the equivalent figure for independent restaurants—3.7 stars (Median=3.9, SD=0.4). On Google Maps, the mean of chain restaurants' average ratings is 3.8 stars (Median=3.9, SD=0.4), 0.4 stars below the corresponding mean for independent restaurants—4.2 stars (Median=4.3, SD=0.4).

Importantly, however, within chain restaurants we do see some form of universal assessment across platforms. The ordinal ranking of chain brands is roughly consistent across the platforms. The 90 chains' pooled average ratings[7] from Google Maps and Yelp correlate strongly (Pearson's $r=0.94$, $p<0.001$). On both platforms, In-N-Out Burger is the highest-rated chain and KFC is the lowest rated chain.

Categories: Table 1 has the nine category coefficients with the largest absolute effect sizes. Holding all else constant, buffet restaurants are associated with the largest increase in GoogleMinusYelp, while Mediterranean restaurants are associated with the smallest increase. However, the maximum category effect size is less than half of that of chain status.

Figure 4(a), which plots the distribution of GoogleMinusYelp per category, further illustrates the associations between GoogleMinusYelp and restaurant category. The mean GoogleMinusYelp among Mediterranean restaurants is 0.4, 0.3 stars lower than the equivalent figure among non-Mediterranean restaurants. Conversely, the mean GoogleMinusYelp among buffet restaurants is 0.8, 0.1 stars higher than the equivalent figure among non-buffet restaurants. Notably, in this univariate analysis, the fast food category has the largest mean GoogleMinusYelp: 1.0 star.

[7] As a robustness check, we also used the median of average ratings per chain and the results are similar.

Interestingly, Figure 4(b) shows that category ranking, based on each category's pooled average rating[8], remains consistent. In particular, the two platforms' category rankings are strongly correlated (Spearman's r=0.97, p<0.001). This finding provides some support for the universal assessment assumption at the category ranking level.

Price tier: In our regression model, price tier has a trivial effect size in predicting GoogleMinusYelp (coefficient=0.02, p<0.05). With all else constant, one unit of increase in price tier is associated with a 0.02-star increase in GoogleMinusYelp.

4.2 Differences in Top-Ranked Restaurants

In our comparison of top-ranked restaurants by geographic region, we found additional misalignments between Google Maps and Yelp, further problematizing the universal assessment assumption. Below, we report the misalignment in the top-n lists between platforms using our ranking approach, in addition to describing the results from our sanity check, which used the native user interfaces of Google Maps and Yelp at the metropolitan area level.

Figure 5 plots to what extent Google Maps' and Yelp's top-three lists (left) and top-five lists (right) differ. The chart's first row represents the census tract level. For a non-trivial share of the census tracts considered in our study, the two platforms produce very different top-three and top-five lists. 40% of the census tracts have at least two restaurants different between their top-three lists from Google Maps and Yelp. In terms of top-five lists, 20% of these census tracts have at least three restaurants different between Google Maps and Yelp. These findings suggest that at the census tract level, the top-ranked restaurants on Google Maps may not always appear at the top on Yelp.

Figure 5: Percentage of geographic areas per number of restaurants different in top-three (left) and top-five (right) lists between Google Maps and Yelp

Table 2: Number of restaurants that are different in the top-ten lists per metropolitan area between Google Maps and Yelp, using our own ranking approach (the *est.* column) and the native ranking results from two platforms' user interfaces (the *native* column), respectively.

	Phoenix, AZ		Las Vegas, NV		Cleveland, OH		Pittsburgh, PA		Charlotte, NC		Madison, WI		Urbana-Champaign, IL	
	est.	native	est.	native	est.	native	est.	native	est.	native	est.	native	est.	native
All	9	9	7	10	9	8	9	8	9	10	9	6	5	8
Mexican	8	6	6	7	9	7	3	7	6	6	5	6	4	7
Chinese	5	7	8	8	7	7	7	6	4	6	2	4	2	5
Italian	6	6	9	10	9	6	5	6	6	6	4	4	4	4
Seafood	6	5	7	8	6	6	5	6	7	6	5	5	2	4

[8] As a robustness check, we also used the median of average ratings per category and the results are similar.

The second row in Figure 5 shows that a large share of the zip codes in our analysis have very different top-three and top-five lists, further challenging the universal assessment assumption. 65% zip codes' Google Maps top-three lists have at least two restaurants different from their corresponding Yelp lists. 67% zip codes' Google Maps top-five lists have at least three restaurants different from their corresponding Yelp lists, a large increase from 20% at the census tract level.

The remaining rows in Figure 5 consider four restaurant categories. Compared with the second row of the figure, a relatively small share of zip codes have very different top-ranked restaurants between Google Maps and Yelp for each category. In other words, two platforms tend to agree more on the top Mexican, Chinese, Italian, and seafood restaurants than they do on the top restaurants in general.

At the metropolitan area level, the top-three and top-five lists differ greatly between Google Maps and Yelp. The five largest metropolitan areas in our dataset have *entirely unique* top-three lists and top-five lists between Google Maps and Yelp. Overall, the misalignment in the top-n lists between Google Maps and Yelp increases as the geographic region granularity expands. In other words, our results suggest that the universal assessment assumption may become especially problematic when a local search query specifies a large geographic region, such as a metropolitan area.

Table 2 visualizes the comparison of top-ten lists at the metropolitan area level, using our rankings and those from Google Maps' and Yelp's native user interfaces. Based on our rankings, five out of the seven metropolitan areas in our dataset have nine restaurants different between their Google Maps and Yelp top-ten lists. Urbana-Champaign has the minimum number of restaurants different between platforms: five. Our sanity check using Google Maps' and Yelp's native ranking approaches yields consistent results: the top-ten lists from two platforms largely differ from each other and the number of restaurants different ranges from six to ten.

Table 2 also visualizes the substantial misalignment in top-ten lists faceted by popular categories per metropolitan area. The results from our ranking approach show less misalignment in the top-n lists across platforms than the previous category-agnostic comparison in most cases. This trend is also observed in our sanity check.

5 DISCUSSION

Our results suggest the universal assessment assumption that many local search technologies inherently adopt is likely flawed. In this section, we discuss the implications of these results for end users of ratings and local search technologies, as well as several design recommendations. We additionally hypothesize about the potential causes for the pairwise divergences in average ratings between Google Maps and Yelp. Finally, we discuss the implications of our results on rating research, outline the limitations of our study, and highlight opportunities for future work.

5.1 Implications for End Users of Ratings

At the highest level, the cross-platform differences in average rating suggest that average rating is not a universal metric that describes a restaurant across platforms. In other words, there is no such thing as "4.5-star restaurant", only a "4.5-star restaurant on Google Maps", a "4.5-star restaurant on Yelp", and so on. Recall that a restaurant with a four-star average rating on Yelp is ranked above 72% of the other restaurants in our dataset, but a restaurant with a four-star average rating on Google Maps is only ranked above 28% of the other restaurants. More generally, comparing different restaurants with different rating sources will very likely favor the restaurants from Google Maps, which are not necessarily better than the restaurants with lower average Yelp ratings.

For end users who consult both review platforms and who may have become aware of the general cross-platform differences, our results shed more light into the complexity in translating

average ratings between platforms. That is, simply adding 0.7 stars (the average value of GoogleMinusYelp for all restaurants) to a restaurant's average Yelp rating does not make the value comparable with another restaurant's average Google Maps rating. For instance, this adjustment should be 1.1 stars when comparing two chain restaurants and 0.4 stars when comparing two restaurants that are under the specialty food category. These complex translations between average Yelp rating and average Google Maps rating suggest that tools would be needed to perform a translation; simple addition is insufficient.

For end users who rely on a single review platform to compare restaurants, their choice of review platform may influence what top-ranked restaurants they see. Our ranking experiments showed that top-ranked restaurants likely differ between Google Maps and Yelp, especially in large geographic regions, e.g. a metropolitan area. Therefore, even though these two platforms have somewhat correlated average ratings, Yelp users and Google Maps users might still come to different conclusions about which restaurants they expect to outperform the others.

These results paint a challenging picture: both ratings and rankings are inconsistent between Google Maps and Yelp. How, then, should end users find good restaurants? Our results imply a bit of guidance here. First, we suggest that Yelp might be well-suited for customers who prefer independent restaurants strongly. Although chain restaurants are rated lower than independent restaurants on both platforms, the gap is more prominent on Yelp. Therefore, chain restaurants are likely to appear at the bottom of a ranked list on Yelp. For customers who strongly prefer independent restaurants, Yelp may help them effectively avoid chain restaurants, whereas Google Maps may present many chain options mixed with independent restaurants.

Second, our comparison of top-ranked restaurants by geographic region suggests that consulting both Google Maps and Yelp may result in more top-ranked restaurant results, especially when the specified geographic region is large. In the case of needing more restaurant options, end users may consider exploring different review platforms and local search technologies.

5.2 Implications for Local Search Technologies

Below, we first discuss local search technologies' potential business constraints that may drive them to avoid integrating ratings from multiple review platforms. We then categorize our potential design recommendations for local search technologies into three categories based on how flexible these constraints are likely to be.

5.2.1 Tension between Platform Diversity and Business Incentives

The findings above highlight a key tension between the business incentives of companies that run local search technologies and their goal to help their users make well-informed restaurant decisions. Our results demonstrate that users would likely be better informed if local search technologies display ratings from multiple review platforms, but there are a number of business pressures that push local search companies towards single-platform approaches.

First, local search technologies want to minimize the licensing fees necessary to acquire ratings from review platforms. Both Google Maps and Yelp charge a service fee for heavy use of their APIs, which can be a financial burden to local search technologies. As such, local search technologies likely have a financial incentive to source ratings from a single review platform. Put another way, by displaying multiple review platforms' average ratings next to each other, local search technologies could highlight diverse restaurant evaluations to their end users; however, the associated costs may make this approach unfeasible in the near future.

Second, local search technologies' business relationships with review platforms potentially introduce an additional constraint on what rating sources they could use. A company like Google that has its own review platform will want to elevate its own review platform over others when possible. Indeed, Google Search prioritizes Google Maps' ratings over other rating sources (e.g. Yelp, TripAdvisor, and Facebook). In turn, competitors of Google will want to avoid promoting

Google Maps and find other review platforms as rating sources. This can be seen among Apple Maps (competitor of Google Maps) and Bing (competitor of Google Search), which refrain from using Google Maps ratings and leverage other rating sources. Put simply, local search technologies are likely to avoid promoting competing review platforms or platforms owned by competitors as their source of ratings.

More generally, our results can provide another data point to inform the growing discussion about monopoly power in the technology industry [19,26]. The findings above suggest that displaying the ratings of multiple review platforms is beneficial for users, meaning that there likely is at least a small degree of consumer harm when a local search technology spotlights only one platform's reviews. This harm is potentially magnified if the absence (or reduced prominence) of a review platform leads to fewer participants in that platform, a process that has been described as a "death spiral" [21]. Given these risks and the scale of local search technologies' influence, more research is needed to inform policy decisions on the potential monopoly power in local search.

5.2.2 Design Recommendations for Local Search Technologies

The potential business constraints mentioned above have direct impacts on what design recommendations are feasible to implement in local search technologies, at least over the short-term. Below, we discuss design recommendations that range from incremental changes to transformative changes, depending on how dismissible these potential constraints may be.

5.2.2.1 Design Recommendations Under Full Constraints

Chain restaurants' greater cross-platform differences suggest that local search technologies may consider faceting local search results by chain status. By separating independent restaurants from chain restaurants, local search technologies will help customers of independent restaurants, who account for the majority of restaurant businesses *[36]*, see average ratings that more suitably represent evaluations from other platforms. It is worth noting that prior research on faceted search has suggested that this approach may be overwhelming to end users, especially when end users have already specific targets in mind [23]. As such, when implementing faceted search, local search technologies need to consider this trade-off, especially when users are searching for specific restaurant names.

Our study also points to another metric that may be of value in communicating restaurant evaluation for chain restaurants specifically. As shown in our results, reviewers across all platforms have reached a consensus on the best or worst chains. Local search technologies may consider integrating this ranking when displaying chain restaurants, e.g. "In-N-Out Burger (#1 chain)".

Another immediate takeaway from our study is that local search technologies should avoid using different review platforms for different restaurants when presenting ratings. This applies to the current ways that Apple Maps and Bing display ratings. Apple Maps uses Yelp as a rating source for some restaurants, but TripAdvisor for others. Bing does the same but with ratings from Facebook instead of TripAdvisor. These use cases of average ratings promote the notion that average rating is a universal metric across platforms. However, our results on the cross-platform differences in average rating dispute this notion and suggest that average ratings of different restaurants from different review platforms are not comparable.

Lastly, local search technologies may consider adding percentile-based metrics so average ratings are standardized. This will provide helpful context for new users of a local search technology. For example, a long-term Google Search user who is used to seeing higher average ratings may be confused by the consistently lower average ratings after switching to Bing and even think that there are no truly good restaurants nearby. Having a percentile indicating how well a restaurant perform relative to others on Yelp may help to disparage this misunderstanding. That being said, this recommendation does not guarantee to eliminate cross-platform differences

in percentile. As noted above in our results, the percentile metric still does not produce consistent results across platforms and the top-ranked restaurants are largely different. Nonetheless, the percentile may still be a useful and straightforward metric for end users of ratings.

5.2.2.2 Design Recommendations Under Somewhat Flexible Business Constraints

Given people's overwhelming use of local search to find good restaurants [43], local search technologies might want to highlight the diversity in top-ranked restaurants across review platforms, especially when the specified geographic area is large. For example, when responding to a "best restaurants in London" query, Google Search might consider showing the lists from Yelp, TripAdvisor, Facebook, etc. directly in its search result page, in lieu of blue links.

For the restaurant properties that are associated with the largest cross-platform differences in average rating, such as chain and buffet restaurants, local search technologies might consider highlighting the diversity of restaurant assessment for them in particular. For example, when a user searches for chain restaurants, a local search technology may acquire restaurant assessments from various review platforms instead of merely relying on its primary review platform.

5.2.2.3 Design Recommendations Under No Business Constraints

In the case where local search technologies have the full capacity to leverage ratings from all review platforms, local search technologies may consider displaying a restaurant's average ratings from various review platforms in a co-equal fashion to highlight any cross-platform difference. Currently, Google Search puts a restaurant's average Google Maps rating in a knowledge panel (an information box displayed on the right side of search results) and the restaurant's average ratings from other review platforms, e.g. Yelp, TripAdvisor, Facebook, in blue links. Because Google Search users prioritize information in knowledge panels over blue links [19], they may easily overlook information provided by review platforms other than Google Maps. By displaying average ratings from all review platforms in its knowledge panel, Google Search may better inform users of any potential cross-platform differences in average rating.

Taking a step further from displaying multiple average ratings, local search technologies may even implement features to explain any cross-platform difference in average ratings. For example, such features may extract detailed opinions from restaurant reviews [30] about different aspects of restaurant assessment, e.g. "Yelp reviewers like this restaurant's service".

There exists another approach that may help users of local search technologies to find good restaurants without exposing them to a plethora of information from multiple review platforms. Local search technologies may consider what platform's ratings to display based on users' personal preferences. For example, we observed that in aggregate, Yelp reviewers have a strong distaste for chain restaurants relative to Google Maps reviewers. Similarly, in aggregate, Yelp reviewers favor specialty food restaurants and Mediterranean restaurants relative to Google Maps reviewers, as these two categories have the lowest GoogleMinusYelp values. As such, local search technologies may consider whether a user' preferences align better with Yelp reviewers or Google Maps reviewers and personalize the source of ratings for the user.

5.3 Potential Causes for Cross-Platform Differences

As mentioned in Related Work, prior studies on review platforms have suggested ratings can be influenced by various factors, which may have led to different ratings on different review platforms. Although our study does not investigate what led to cross-platform differences in ratings, we reflect on a few potential causes based on prior work.

One potential cause for Google Maps' higher average ratings may be the lack of anonymity on the platform relative to Yelp. In the rating domain, the extent of anonymity on a review platform has been shown to be positively linked with the proportion of negative ratings from reviewers [31]. Due to fear of retaliation and harassment, people often avoid harsh criticism when they believe others can identify them [12]. Although both Google Maps and Yelp require reviewers to set up their profiles before giving a rating, Yelp offers a greater extent of anonymity

than Google Maps. While Google Maps' interface displays a reviewer's full name based on the reviewer's profile, Yelp's interface only displays a reviewer's first name and the first letter of the reviewer's last name. Additionally, a Google Maps reviewer's profile connects to a variety of Google products, including interpersonal communication tools such as Gmail, possibly limiting the reviewer's freedom to use pseudonyms in their profiles. Together, these design decisions may potentially lead Google Maps reviewers to be more disinclined to give low ratings to restaurants than Yelp reviewers, and thereby, raise Google Maps' average ratings.

Another possible explanation for Google Maps' higher average ratings is the platform potentially filtering a smaller share of fake positive reviews, i.e. highly positive reviews not written by real restaurant customers. Empirical work on review fraud has presented evidence of independent restaurants soliciting fake positive reviews for themselves [18]. Between Google Maps and Yelp, Yelp's filtering algorithm has been shown to automatically label and remove a non-trivial share of all ratings submitted as fake (as many as 16% [18]). In contrast, no rigorous evidence about Google Maps' systematic removal of fake ratings has been made public. Future work may consider leveraging established fake review detection mechanisms [22] to explore this hypothesis. Such future studies may start with independent restaurants whose average Yelp ratings reach the theoretical minimum, i.e. one-star, but are higher on Google Maps. As independent restaurants with lower average ratings have stronger motivations to commit review fraud [18], their ratings on Google Maps might consist of a larger share of fake positive ratings than others. However, it is worth noting that fake positive reviews are unlikely a major reason for Google Maps' higher ratings. While independent restaurants' average Google Maps ratings are 0.7 stars higher than their Yelp counterparts, chain restaurants, who are less likely to have fake positive reviews, are rated even higher on Google Maps with an increase of 1.1 stars.

The reputation features on Yelp may have also played a role in lowering the platform's average ratings relative to Google Maps. In 2010, Wang argued that reputation features on Yelp, i.e. votes and Elite badges, are the major reasons for the platform's larger share of prolific reviewers than anonymous review platforms such as Yahoo Local [31]. As prolific reviewers tend to provide lower ratings [29], Yelp's reputation features may have driven its average ratings down. In contrast, Google Maps does not provide as strong reputation-based incentives as Yelp. For example, the equivalent elite badge feature on Google Maps, the Local Guide program, is less selective. While Yelp reviews their users' applications for Elite badges and renew the badges annually, Google Maps' Local Guide badges are simply based on total number of contributions [53].

Lastly, GoogleMinusYelp's associations with chain status and restaurant categories and our results on category ranking support Zervas and colleagues' suggestion that reviewers from different platforms may have different preferences in aggregate [34]. If the two platforms' reviewers share similar tastes, we should expect GoogleMinusYelp to be somewhat consistent across different chain statuses and restaurant categories. However, as detailed above, GoogleMinusYelp is larger among chain restaurants and buffet restaurants, which suggests Google Maps reviewers may favor these restaurants more than Yelp reviewers in aggregate. Similarly, salad and Asian Fusion restaurants are ranked higher on Yelp than on Google Maps, indicating reviewers' different aggregate preferences across platforms.

5.4 Implications for Research

Our results suggest that research on restaurant review platforms should take a multi-site approach. Given the magnitude of cross-platform differences we observed, it reasonable to hypothesize that findings observed on Yelp might not generalize to Google Maps and vice versa. The existing literature on restaurant ratings often focuses on a single review platform, as described in Related Work [2,3,11,17]. Future work on ratings may want to consider validating

prior findings from a single review platform with other platforms. For example, Bakhshi and colleagues found that bad weather is associated with lower ratings on Yelp [2]; will this association occur on Google Maps as well? More generally, our study provides yet more support for social computing researchers' calls for more multi-site studies [4].

5.5 Limitations

One important limitation in our study is that our analyses focused on the local search landscape in the U.S. and thereby may not be generalizable to other geographic contexts. Moreover, given that Yelp and Google Maps may not be popular in other countries, future work should extend its focus beyond these two platforms.

Relatedly, as mentioned in Methods, to collect a restaurant's rating distribution, we took advantage of the Yelp Open Dataset, a decision limiting our data to seven metropolitan areas in the U.S. As a result, our findings may be not representative of restaurants located in rural areas in particular. However, given that the majority of the U.S. population resides in urban areas [6], our results still speak to the likely experience of a large share of the population. Future work may want to consider using population-weighted sampling to construct datasets that are more representative of the distribution of the U.S. population along the rural-urban spectrum.

In processing the Yelp Open Dataset, we filtered out restaurants that have fewer than ten ratings, which may have introduced another limitation to our analyses. Future work may consider experimenting with different filtering strategies and examine their impact on restaurants' cross-platform differences.

6 FUTURE WORK

Given the popularity of local search technologies, our study points to an interesting research topic at the intersection of social computing and spatial computing. Future work may consider conducting a large audit of review platforms using improved simulation of local search queries. Our results on the differences in top-ranked restaurants provide some preliminary insights into this area; more formal, comprehensive audits are needed to understand how restaurant ratings influence what restaurants people visit. For example, inspired by Johnson and colleagues' approach that leverages taxi trip datasets to audit routing algorithms [10], future work could use geotagged social media data (e.g. check-ins at restaurants) to infer where local search queries may have happened. This more ecologically valid approach may uncover real-world differences in local search results.

Researchers studying geotagged user-generated content may want to further their investigation in cross-platform differences. While our study focused on the average rating metric, future work may explore whether a restaurant's geographic location is associated with cross-platform difference in number of ratings. In the Appendix, we provide the key descriptive statistics on number of ratings from our dataset, faceted by chain status, restaurant category, and price tier. Our results suggest that reviewers of Google Maps and Yelp rate different types of restaurants. For example, on Yelp, chain restaurants gain fewer ratings than independent restaurants, whereas on Google Maps, the opposite is true. User-generated content research may further examine this divergence in depth in consideration of geographic factors such as demographics and rurality. For example, future work could investigate whether restaurants located in areas with different demographics exhibit an equal amount of cross-platform differences in number of ratings.

7 CONCLUSION

Examining the two most popular restaurant review platforms in the U.S., Google Maps and Yelp, we found that restaurant ratings on Google Maps are, on average, 0.7 stars higher than those

on Yelp. We observed that this increase is larger among chain restaurants than independent restaurants. Additionally, we found extensive diversity in the top-ranked restaurants for a given metropolitan area between Google Maps and Yelp. Our study problematizes the use of a single, primary review platform as a source for ratings in local search technologies and point to corresponding design recommendations.

ACKNOWLEDGMENTS

The first author would like to thank Nicholas Vincent for his feedback on an early draft of the paper. The authors also appreciate the anonymous reviewers for their helpful feedback and comments. This work was supported by NSF grant IIS-1815507.

REFERENCES

[1] Michael Anderson and Jeremy Magruder. 2012. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *Econ J* 122, 563 (September 2012), 957–989. DOI:https://doi.org/10.1111/j.1468-0297.2012.02512.x

[2] Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. 2014. Demographics, Weather and Online Reviews: A Study of Restaurant Recommendations. In *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14), ACM, New York, NY, USA, 443–454. DOI:https://doi.org/10.1145/2566486.2568021

[3] Saeideh Bakhshi, Partha Kanuparthy, and David A. Shamma. 2015. Understanding Online Reviews: Funny, Cool or Useful? In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '15), ACM, New York, NY, USA, 1270–1276. DOI:https://doi.org/10.1145/2675133.2675275

[4] Taryn Bipat, Susan R. Fussell, Brent Hecht, Charles Kiene, David W. McDonald, and Mark Zachry. 2018. Navigating the Challenges of Multi-Site Research. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '18), ACM, New York, NY, USA, 409–415. DOI:https://doi.org/10.1145/3272973.3273014

[5] Mike Blumenthal. 2014. Why does Google show a 4.8 rating when I have all 5-star reviews? *Local University*. Retrieved January 31, 2020 from https://localu.org/2014/12/22/google-show-4-8-rating-5-star-reviews/

[6] US Census Bureau. New Census Data Show Differences Between Urban and Rural Populations. *The United States Census Bureau*. Retrieved October 15, 2019 from https://www.census.gov/newsroom/press-releases/2016/cb16-210.html

[7] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. In *Eleventh International AAAI Conference on Web and Social Media*. Retrieved April 18, 2019 from https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15697

[8] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User Attitudes Towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), ACM, New York, NY, USA, 494:1–494:14. DOI:https://doi.org/10.1145/3290605.3300724

[9] Andrew Hall, Jacob Thebault-Spieker, Shilad Sen, Brent Hecht, and Loren Terveen. 2018. Exploring the Relationship Between "Informal Standards" and Contributor Practice in OpenStreetMap. In *Proceedings of the 14th International Symposium on Open Collaboration* (OpenSym '18), ACM, New York, NY, USA, 10:1–10:11. DOI:https://doi.org/10.1145/3233391.3233962

[10] I. Johnson, J. Henderson, C. Perry, J. Schöning, and B. Hecht. 2017. Beautiful…but at What Cost?: An Examination of Externalities in Geographic Vehicle Routing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2 (June 2017), 15:1–15:21. DOI:https://doi.org/10.1145/3090080

[11] Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith. 2014. Narrative framing of consumer sentiment in online restaurant reviews. *First Monday* 19, 4 (March 2014). Retrieved September 26, 2018 from http://firstmonday.org/ojs/index.php/fm/article/view/4944

[12] Ruogu Kang, Stephanie Brown, and Sara Kiesler. 2013. Why Do People Seek Anonymity on the Internet?: Informing Policy and Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), ACM, New York, NY, USA, 2657–2666. DOI:https://doi.org/10.1145/2470654.2481368

[13] Senior PR Coordinator Kathleen Liu and February 1 Friday. 2019. Consumer Alerts: No Tolerance for Review Manipulation. *Yelp*. Retrieved September 18, 2019 from https://blog.yelp.com/2019/02/consumer-alerts-no-tolerance-for-review-manipulation

[14] Balázs Kovács, Glenn R. Carroll, and David W. Lehman. 2013. Authenticity and Consumer Value Ratings: Empirical Tests from the Restaurant Domain. *Organization Science* 25, 2 (July 2013), 458–478. DOI:https://doi.org/10.1287/orsc.2013.0843

[15] Stuart E. Levy, Wenjing Duan, and Soyoung Boo. 2013. An Analysis of One-Star Online Reviews and Responses in the Washington, D.C., Lodging Market. *Cornell Hospitality Quarterly* 54, 1 (February 2013), 49–63. DOI:https://doi.org/10.1177/1938965512464513

[16] Michael Luca. 2015. User-Generated Content and Social Media. In *Handbook of Media Economics*, Simon P. Anderson, Joel Waldfogel and David Strömberg (eds.). North-Holland, 563–592. DOI:https://doi.org/10.1016/B978-0-444-63685-0.00012-7

[17] Michael Luca. 2016. *Reviews, Reputation, and Revenue: The Case of Yelp.Com*. Social Science Research Network, Rochester, NY. Retrieved November 25, 2018 from https://papers.ssrn.com/abstract=1928601

[18] Michael Luca and Georgios Zervas. 2016. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science* 62, 12 (December 2016), 3412–3427. DOI:https://doi.org/10.1287/mnsc.2015.2304

[19] Farhad Manjoo. 2018. Stumbles? What Stumbles? Big Tech Is as Strong as Ever. *The New York Times*. Retrieved April 4, 2019 from https://www.nytimes.com/2018/08/01/technology/big-tech-earnings-stumbles.html

[20] Mary McGlohon, Natalie Glance, and Zach Reiter. 2010. Star Quality: Aggregating Reviews to Rank Products and Merchants. In *Fourth International AAAI Conference on Weblogs and Social Media*. Retrieved August 13, 2019 from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1507

[21] Connor McMahon, Isaac L Johnson, and Brent Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *ICWSM*, 142–151.

[22] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013. What Yelp Fake Review Filter Might Be Doing? In *Seventh International AAAI Conference on Weblogs and Social Media*. Retrieved October 15, 2019 from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6006

[23] Xi Niu, Xiangyu Fan, and Tao Zhang. 2019. Understanding Faceted Search from Data Science and Human Factor Perspectives. *ACM Trans. Inf. Syst.* 37, 2 (January 2019), 14:1–14:27. DOI:https://doi.org/10.1145/3284101

[24] Paul Phillips, Krystin Zigan, Maria Manuela Santos Silva, and Roland Schegg. 2015. The interactive effects of online reviews on the determinants of Swiss hotel performance: A neural network analysis. *Tourism Management* 50, (October 2015), 130–141. DOI:https://doi.org/10.1016/j.tourman.2015.01.028

[25] Lee Rainie, Kristen Purcell, Amy Mitchell, and Tom Rosenstiel. 2011. Where people get information about restaurants and other local businesses. *Pew Research Center: Internet, Science & Tech*. Retrieved August 27, 2019 from https://www.pewinternet.org/2011/12/14/where-people-get-information-about-restaurants-and-other-local-businesses/

[26] Kenneth Rogoff. 2019. Big tech has too much monopoly power – it's right to take it on. *the Guardian* (April 2019). Retrieved from https://www.theguardian.com/technology/2019/apr/02/big-tech-monopoly-power-elizabeth-warren-technology

[27] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms.

[28] Jerry O. Talton III, Krishna Dusad, Konstantinos Koiliaris, and Ranjitha S. Kumar. 2019. How Do People Sort by Ratings? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), ACM, New York, NY, USA, 305:1–305:10. DOI:https://doi.org/10.1145/3290605.3300535

[29] Chenhao Tan, Ed H. Chi, David Huffaker, Gueorgi Kossinets, and Alexander J. Smola. 2013. Instant foodie: predicting expert ratings from grassroots. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (CIKM '13), ACM, New York, NY, USA, 1127–1136. DOI:https://doi.org/10.1145/2505515.2505712

[30] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent Aspect Rating Analysis Without Aspect Keyword Supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '11), ACM, New York, NY, USA, 618–626. DOI:https://doi.org/10.1145/2020408.2020505

[31] Zhongmin Wang. 2010. Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews. *The B.E. Journal of Economic Analysis & Policy* 10, 1 (January 2010). DOI:https://doi.org/10.2202/1935-1682.2523

[32] Zheng Xiang, Qianzhou Du, Yufeng Ma, and Weiguo Fan. 2017. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management* 58, (February 2017), 51–65. DOI:https://doi.org/10.1016/j.tourman.2016.10.001

[33] Inc Yelp and March 18 Thursday. 2010. Yelp's Recommendation Software Explained. *Yelp*. Retrieved September 18, 2019 from https://blog.yelp.com/2010/03/yelp-review-filter-explained

[34] Georgios Zervas, Davide Proserpio, and John Byers. 2015. *A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average*. Social Science Research Network, Rochester, NY. Retrieved September 7, 2019 from https://papers.ssrn.com/abstract=2554500

[35] Ye Zhang and Shu Tian Cole. 2016. Dimensions of lodging guest satisfaction among guests with mobility challenges: A mixed-method analysis of web-based texts. *Tourism Management* 53, (April 2016), 13–27. DOI:https://doi.org/10.1016/j.tourman.2015.09.001

[36] 2017. Report: Consumers prefer independent restaurants over chains. *Nation's Restaurant News*. Retrieved April 23, 2019 from https://www.nrn.com/consumer-trends/report-consumers-prefer-independent-restaurants-over-chains

[37] 2019. *GoogleChrome/puppeteer*. GoogleChrome. Retrieved September 16, 2019 from https://github.com/GoogleChrome/puppeteer

[38] 2019. Franchising. *Wikipedia*. Retrieved October 14, 2019 from https://en.wikipedia.org/w/index.php?title=Franchising&oldid=919113784

[39] QSR & restaurant holiday strategy. *Think with Google*. Retrieved January 16, 2020 from https://www.thinkwithgoogle.com/marketing-resources/qsr-restaurant-strategy/

[40] Google Maps: 1 Billion Monthly Users – GPS Business News. Retrieved September 12, 2019 from https://gpsbusinessnews.com/Google-Maps-1-Billion-Monthly-Users_a4964.html

[41] Factsheet. *Yelp*. Retrieved September 16, 2019 from https://www.yelp.com/factsheet

[42] • Top U.S. mapping apps by reach 2018 | Statista. Retrieved October 14, 2019 from https://www.statista.com/statistics/865419/most-popular-us-mapping-apps-ranked-by-reach/

[43] U.S. online review usage for local business research 2017. *Statista*. Retrieved September 13, 2019 from https://www.statista.com/statistics/695736/online-review-site-usage-local-business-research-usa/

[44] Write reviews and add ratings of places - iPhone & iPad - Google Maps Help. Retrieved September 11, 2019 from https://support.google.com/maps/answer/6230175?co=GENIE.Platform%3DAndroid&hl=en&oco=1

[45] review filter. *Yelp*. Retrieved January 30, 2020 from https://blog.yelp.com/search/review+filter

[46] Daily Access Limiting - Yelp Fusion. Retrieved March 31, 2020 from https://www.yelp.com/developers/documentation/v3/rate_limiting

[47] Pricing & Plans | Google Maps Platform. *Google Cloud*. Retrieved March 31, 2020 from https://cloud.google.com/maps-platform/pricing

[48] Yelp Dataset. Retrieved September 16, 2019 from https://www.yelp.com/dataset

[49] Mintel 2018. Retrieved June 5, 2019 from https://store.mintel.com/us-fast-casual-restaurants-market-report

[50] Restaurants in the U.S. *Statista*. Retrieved June 4, 2019 from http://www.statista.com/study/11721/restaurants-in-the-us-statista-dossier/

[51] How Not To Sort By Average Rating – Evan Miller. Retrieved September 18, 2019 from http://www.evanmiller.org/how-not-to-sort-by-average-rating.html

[52] Google Trends. *Google Trends*. Retrieved January 31, 2020 from https://trends.google.com/trends/explore?q=restaurants%20near%20me&geo=US

[53] Local Guides. Retrieved September 8, 2020 from https://maps.google.com/localguides/howto

APPENDIX

This section provides descriptive statistics about cross-platform differences in the number of ratings in our dataset to assist future work in understanding what restaurants people rate.

Descriptive Statistics about Number of Ratings on Google Maps and Yelp

In general, the restaurants in our datasets have more ratings on Google Maps than on Yelp, and their numbers of ratings are weakly correlated across platforms (Spearman's r=0.33, p<0.001). Both platforms' numbers of ratings follow a long-tail distribution (Yelp: Mean=114, Median=55, Min=10, Max=7968; Google Maps: Mean=370, Median=225, Min=10, Max=38211), and the median number of rating on Google Maps is higher than the median on Yelp. Because of the long-tail distribution and our samples being paired, we used a Wilcoxon signed-rank test to test two distributions' difference and found it to be statistically significant (p<0.001).

Figure 6 visualizes whether chain or independent restaurants have more ratings on each platform, using log-transformed number of ratings. While chain restaurants receive more ratings than independent restaurants on Google Maps, we observe an opposite trend on Yelp. A Mann-Whitney U test indicated that the number of ratings is higher for chain restaurants (Median=310)

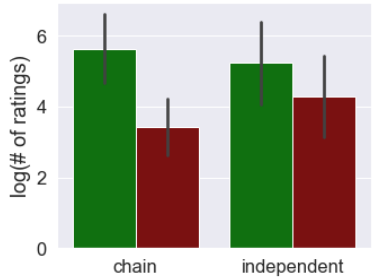

Figure 6: Log-transformed numbers of ratings from Google Maps and Yelp, faceted by chain status. Error bars indicate standard deviation.

Figure 7: Restaurant category ranking by median number of ratings on Google Maps (x-axis) and Yelp (y-axis)

than for independent restaurants (Median=206) on Google Maps ($p<0.001$). On Yelp, however, we observed an opposite trend: a Mann-Whitney U test indicated that chain restaurants receive a smaller number of ratings than independent restaurants ($p<0.001$). The median number of ratings is 27 for chain restaurants and the equivalent figure for independent restaurants is 70.

Moreover, the categories with more (or fewer) ratings on Google Maps are not the ones with more (or fewer) ratings on Yelp. We observed substantial inconsistencies in category ranking based on the median number of ratings. In Figure 7, above the line, we see some categories ranked higher in number of ratings on Yelp than on Google Maps, including Thai, Asian Fusion, Vegetarian, and Japanese. Conversely, below the line are the categories ranked lower on Yelp than on Google Maps, such as Fast Food, Burgers, and Tex-Mex.

Finally, Figure 8 plots the log-transformed number of ratings per price tier on both platforms. On each platform, a Kruskal Wallis Test indicated that there is a statistically significant difference in number of ratings across price tiers (Google Maps: $H(3)=264.6$, $p<0.001$; Yelp: $H(3)=2558.0$, $p<0.001$). The second lowest price tier has the highest median in number of ratings on Google Maps, whereas the highest price tier has the highest median on Yelp.

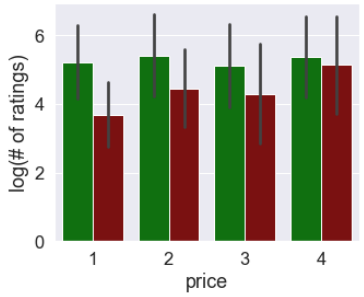

Figure 8: Log-transformed number of ratings from Google Maps and Yelp, faceted by price tier. Error bars indicate standard deviation.