# MS&E 334 Final Paper

Je-ok Choi
jchoi89@stanford.edu

Kevin Li
kevindli@stanford.edu

Zhaonan Qu
zhaonanq@stanford.edu

December 11, 2022

## 1 Introduction

Online marketplaces have transformed the way that buyers and sellers find each other and transact. Instead of trading at a bazaar or through a long chain of intermediaries, a user can browse through their potential trading partners or even let the marketplace choose one for them and make a transaction with a click of a button. In order to facilitate satisfactory transactions, these platforms have been using extensively user reviews to rate and compare the market participants. However, there has recently been a fair amount of discussion on how useful these reviews are when they are closely clustered around the maximum possible rating due to wide-scale inflation.

Rating inflation has become nearly ubiquitous among a wide variety of online marketplaces. On eBay, where each transaction is rated as either positive or negative, sellers have an average positive percentage of 99.3% with a median of 100% [NT15]. Even on service-based platforms where user rating is one of very few comparison metrics available, ratings are severely inflated. For instance, for overall ratings on their accommodations on Airbnb, guests submit a five-star rating 74% of the time and a four-star rating 20% of the time [FGH17]. On Uber, which does not officially provide public information about ratings, the average rating is estimated to be around 4.8, and it has been repeatedly reported that drivers with average ratings lower than 4.6 are at the risk of deactivation [Coo15].

Such rating inflation can be detrimental for online marketplaces in several ways. First of all, it is hard to distinguish among several different products or services when every rating is clustered around the maximum possible value. Hence, the significance of ratings as an informative metric for comparison diminishes. This may also lead to lower user trust of these metrics and eventually, the platform. Moreover, on a platform like eBay, every seller's average rating is less impacted after receiving a low rating.

There have been several theoretical and empirical studies that identify and examine the nature of rating inflation. In this paper, we first discuss a subset of such papers and motivate the research questions we want to explore. Then, we talk about an algorithm, `FairJudge`,

which estimates the fairness of users and goodness of products, which we utilize in our analysis. Lastly, we analyze the fairness and goodness of users on Yelp to investigate our research questions.

## 2 Previous Literature

The setting for our study is online marketplaces/platforms where we examine user ratings. Due to a plethora of successful online platforms, which have emerged in the last decade or so, many scholars have explored user rating data from several different marketplaces as well as developed theoretical frameworks to characterize user behaviors with regards to ratings. In this section, we review a few papers which motivate our research questions that we seek to answer.

[F+17] examines the reputation system of a large online labor market. On this platform, they find that the fraction of workers receiving five-star ratings increased dramatically from 32% in 2007 to 85% in 2016, and also found clear evidence that feedback scores have been steadily increasing. They consider the following two potential reasons for these increases to be:

- Rater satisfaction has increased.
- Raters are lowering their standards.

Using the written feedback that accompanies numerical scores as an alternative measure of rater satisfaction, they find that although predicted ratings based on the alternative measure have increased over time, they do not increase as much as the actual ratings do, providing clear evidence that there might be inflation, where raters are lowering their standards over time.

Moreover, using a stylized model of reputation dynamics, [F+17] demonstrate that the rating inflation is due to the costs that raters incur when leaving negative feedback, such as the fear of retaliation or unwillingness to harm the rated individual, which both increase over time. In addition, from a quasi-experiment with the platform's introduction of a new, "private" feedback mechanism, they find raters are far more truthful about "bad" performances when their opinions will not be shared publicly, confirming their findings from their theoretical model.

[FGH17] uses a field experiment and data from Airbnb to show that strategic reciprocity causes rating inflation. For instance, 6% of guests who anonymously answered that they would not recommend their host gave a five-star rating on the public review. In the field experiment in which the reviews of the treatment group are hidden until both parties (guests and and hosts) submit a review, the simultaneous revealing of reviews increased the review rates of guests by 1.8% while the proportion of five-star reviews decreased by 1.5%.

While [F+17] and [FGH17] show the possible effect of retaliation and reciprocity, few studies delve into how the effect differs depending on the characteristics of rated individuals/businesses. [ZPB15] compare the ratings of properties on Airbnb and TripAdvisor and find that while the average rating on Airbnb of 4.7 stars is higher than TripAdvisor's 3.8 stars, the difference was much smaller when only considering vacation rentals on TripAdvisor, suggesting that the fear of retaliation for raters may be greater for individual businesses rather than for hotel chains. Furthermore, [LZ16] study restaurant data on Yelp and find that independent (typically family-owned) restaurants are more likely to have fake reviews than chain restaurants, and the prevalence of fake reviews increases over time, which enhances the rating inflation of small and independent businesses caused by the fear of retaliation.

# 3 Research Questions and Initial Analysis

We consider the following questions in this study:

- Is rating inflation a real problem?
- If so, what causes rating inflation? Are businesses getting better over time or are reviewers giving higher ratings than they should?
- How does rating inflation differ across various kinds of services, e.g. restaurants vs. beauty spas/home services?
- How can we solve the rating inflation problem?

To study these questions, we use the Yelp Dataset, which has users/businesses/ratings data over 2006-2017, comprised of data from 12 metropolitan areas in the US, Canada, UK, and Germany. As discussed in the introduction, since different trends in ratings have been observed for different business types, we look at two types of businesses: restaurants and personal services (beauty and spas, home services). By examining the yearly average ratings and percentage of ratings, we see that personal services have witnessed a sharp increase in both the average rating and the percentage of five star ratings after 2012, while the average rating of restaurants and percentage of five star ratings did not rise as drastically.

We conjecture that the there exists fear of retaliation: people are afraid to give owners of businesses, whom they potentially could interact with again, poor ratings. As a result, ratings go up, but the fairness of reviewers goes down: reviewers are less reliable. Goodness stays roughly equivalent: the quality of these businesses does not actually change. For Yelp, we think that businesses offering personal services will exhibit the trend above, while restaurants will not.

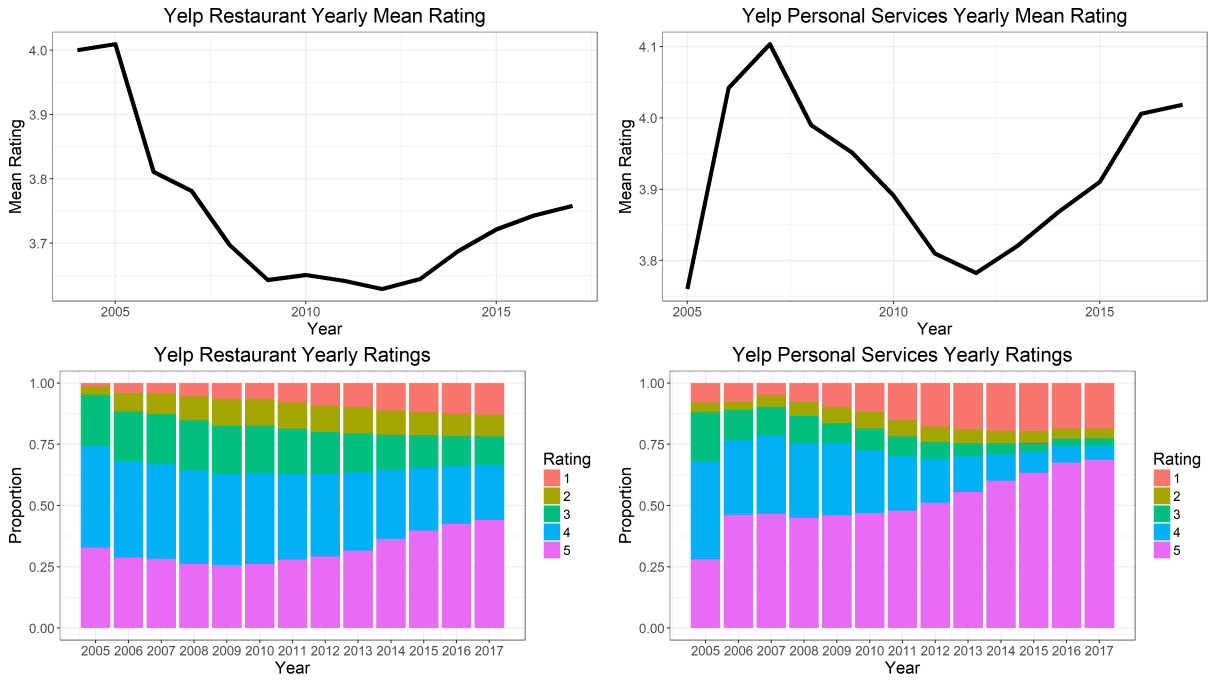


*Figure 3.1: Yelp Yearly (Proportional, Average) Ratings of Restaurants and Personal Services Businesses*

# 4 Methods

The challenge to analyzing rating inflation is the following. A restaurant usually has a wide range of ratings from different users. Some of these rating differences may be due to the personal tastes of customers. That is, a user has different preferences regarding type of service and restaurants. Some users are more likely to give higher reviews across all restaurants, i.e. they have an intrinsic bias. On the other hand, each restaurant has some intrinsic quality, and the resulting ratings that we see of a restaurant by different users is a combination, in some sense, of these factors.

How do we extract a meaningful measure of the intrinsic quality of restaurants, and the intrinsic bias of users? We employ the `FairJudge` framework proposed by [Kum+17], which uses three quantities that recursively define each other. For each restaurant, we have a score called goodness that captures its intrinsic quality, which is based on the reviews it receives, but based on how reliable each review is.

For each user, there is a score termed fairness that describes how fair a particular user is when reviewing restaurants. This can be interpreted as the intrinsic bias of a user. A fair user gives ratings that are usually close to the goodness of a business, while an unfair user deviates from the goodness frequently. The practical problem in [Kum+17] is to identify fraudulent users who are employed to give fake good or bad reviews. In this framework, such users will have low fairness score.

A user may have category-dependent biases. For example, a user may be a good judge of Thai restaurants, but not beauty spas. In our analysis we differentiate between two categories of businesses, restaurants and personal services (beauty and spas, home services), to analyze this intuition.

The key to linking the goodness of restaurants and fairness of customers are the reviews. More precisely, the `FairJudge` framework proposes that each rating from a customer to a restaurant has a reliability score, which depends on how reliable a customer is at giving reviews, but also how close the review is to the intrinsic quality of the customer, since even a fair customer may give out an unfair review from time to time.

**Definition 4.1.** *(`FairJudge` framework a la [Kum+17]) The rating network is modeled as a bipartite network, where user $u$ gives a rating $(u, p)$ to product $p$, where score $s(u, p)$, scaled to be between -1 and 1.*
*Let $O(u)$ be the set of ratings given by a user $u$ and $I(p)$ be the set of ratings received by a business.*

$$\textit{Fairness of User: } F(u) = \frac{\sum_{(u,p)\in O(u)} R(u,p)}{|O(u)|}$$

$$\textit{Goodness of Business: } G(p) = \frac{\sum_{(u,p)\in I(p)} R(u,p)\cdot s(u,p)}{|I(p)|}$$

$$\textit{Reliability of rating: } R(u,p) = \frac{1}{2}(F(u) + (1 - \frac{|s(u,p) - G(p)|}{2}))$$

It is clear that the three quantities recursively define each other. The fairness of a customer is the average of reliability scores of reviews he or she gives. The goodness of a business is the average of the reviews it receives, with each review weighted by the reliability of that review. We also see that the goodness has range in [-1,1], fairness has range in [0,1], and reliability has range in [0,1].

There are several practical issues that are addressed in the paper by [Kum+17]. More specifically, there is the common cold start problem. A user who has given few ratings is hard to evaluate. This is because a fraudulent user could be masking as a normal user at first in order to gain credibility. To address this problem, Bayesian priors are assigned to each user's fairness score, and similarly for each business' ratings. The second issue is that the behavior of users is also an important aspect that can improve the evaluation of fairness and goodness. For example, a fraudulent user tends to post several ratings in a very short timespan. Incorporating such behaviors will greatly improve the calculation of fairness and hence goodness. This is very important since fraudulent ratings could contribute to rating inflation, something that needs to be studied further. These two issues were addressed in the `FairJudge` framework, and we refer to the paper for details.

Another important result of the paper is an iterative algorithm with exponential convergence rate that calculates fairness, goodness, and reliability scores. We summarize it stylistically here, referring to the paper for details.

**Lemma 4.1.** *Given $F^t(u)$, $R^t(u,p)$, $G^t(p)$, the following recursive algorithm*

$$G^{t+1}(p) = \frac{\sum_{(u,p)\in I(p)} R^t(u,p)\cdot s(u,p)}{|I(p)|}$$
$$R^{t+1}(u,p) = \frac{1}{2}(F^t(u) + (1 - \frac{|s(u,p) - G^{t+1}(p)|}{2}))$$
$$F^{t+1}(u) = \frac{\sum_{(u,p)\in O(u)} R^{t+1}(u,p)}{|O(u)|}$$

*converges to a unique solution, where $s(u,p) \in [-1,1]$ is the score that user $u$ gives to product $p$.*
*The difference between scores at time $t$ and the unique scores are bounded above by $(3/4)^t$.*

This guarantees a convergent algorithm that provides a unique set of scores. We will employ this algorithm (and its variants taking into account the cold start problem and behavioral indicators of fake review) to conduct our analysis. This network approach by [Kum+17] can be seen as a complement to some statistical frameworks where users', reviewers', and ratings' scores are modeled using particular probability density functions, and the goal there is to do maximum likelihood estimation on the data.

The main objective of the `FairJudge` (later revised to be called `rev2`) framework is to solve the problem of detecting fraudulent activities in online platforms with rating systems. The main idea of this project is to apply the fairness and goodness framework as a method to extract intrinsic quality and fairness to study the dynamic problem of rating inflation, by calculating fairness and goodness scores for each calendar year, and investigating the trends of these scores over time. More specifically, we apply the `FairJudge` algorithm to rating data on Yelp bucketed by year to produce a time series of fairness and goodness, then analyze it to see if the fairness scores for reviewers of personal services differ from those of restaurants, as we conjectured.

# 5 Analysis

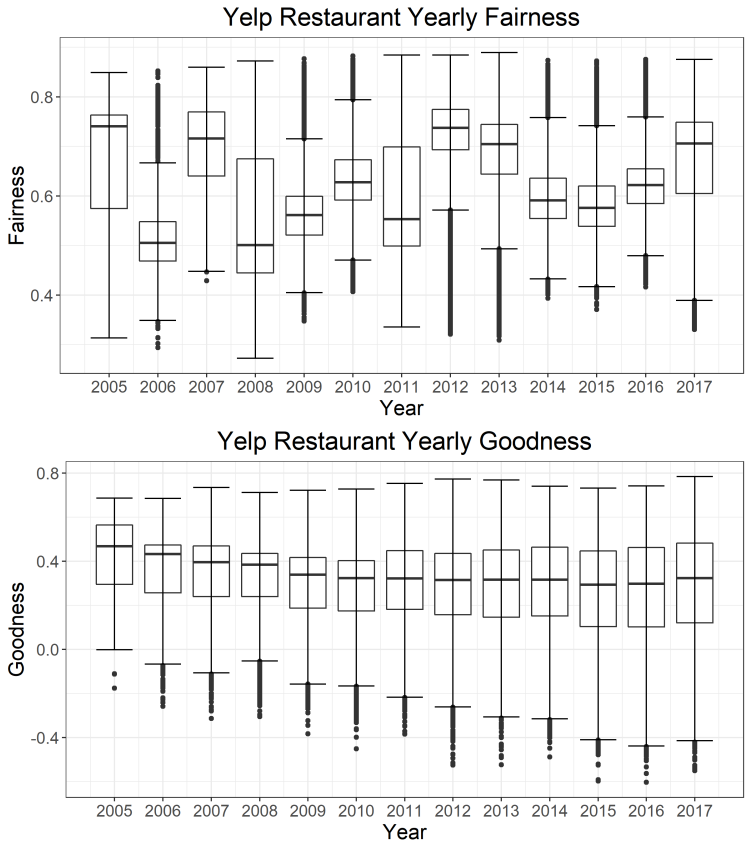


*Figure 5.1: Fairness and Goodness of Restaurants*

We first discuss rating trends for restaurants. First of all, recall that there is continuous decline in average ratings all the way until 2012, when average ratings finally began to climb. Taking a closer look at Figure 3.1, we see that the decline in average rating until 2012 is due to a combination of slight shrinkage of proportion of 5-star ratings and an increase in the proportion of 1-star to 3-star ratings. Thus, on the surface it seems that reviewers of Yelp restaurants became more critical. This is interesting in its own right: did the proportion of 1-star ratings increase because there are more bad restaurants, or because users on Yelp are more and more likely to give out 1-star ratings because they have become less fair? If the latter, what could be the cause? After 2012, the proportion of 5-star ratings starts to climb, while the proportion of 1-star ratings also continues to climb, resulting in a more polarized rating distribution. Again, this phenomenon is interesting. We have some conjectures as to why this might be true. Perhaps fraudulent behavior on Yelp's rating system is becoming more rampant: polarized reviews increase as restaurants hire services to post fake good reviews about themselves or, even worse, fake bad reviews of their competitors. There is evidence for this phenomenon in the Amazon rating system, for example. Or, it could just be that reviewers are authentic, but they are just becoming less objective and ready to

give out overly lavish reviews to average restaurants or lash out at decent restaurants over occasional under-performances. If either scenario is the cause, then the observed fairness of users should decrease from 2012 onwards. Looking at the time series for average fairness of restaurant reviewers on Yelp, we see that the fairness indeed drops from 2012 to 2015. However, between 2015 and 2017 it increases again. The average goodness of restaurants has stayed around the same level over the years, but with growing variance. The time series studies of goodness and fairness are inconclusive, although the shrinking proportion of 2-, 3-, and 4-star ratings, i.e. the polarization of Yelp ratings, is an interesting phenomenon. The growing variance in the quality of restaurants could be what is causing the polarization.

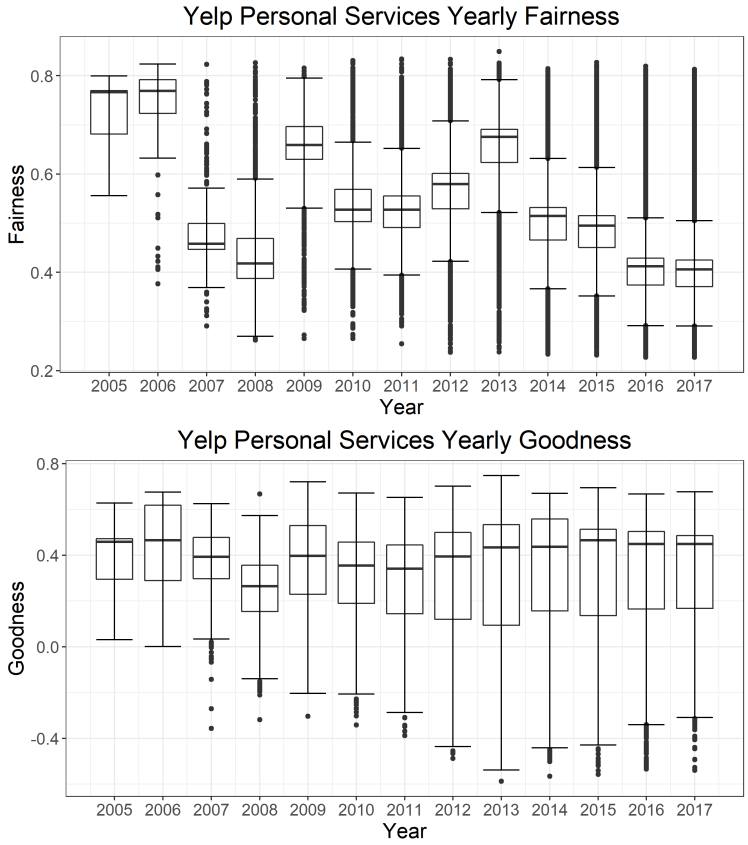


*Figure 5.2: Fairness and Goodness of Personal Services Businesses*

On the other hand, the time series of personal services provide some more firm evidence. First off, the average goodness also stays around the same level, although the variance also increases. There is a steady decrease in fairness starting from 2013. Around the same time, the average rating of personal services witnessed a dramatic increase. What is more striking is the degree of polarization in the distribution of ratings. 2-, 3-, and 4-star ratings account for less than 10 percent of the total ratings. The rest are mostly 1 star and 5 star ratings, both of which witnessed a steady increase in proportion starting in 2012. In this case, the evidence seems to support the conjecture that users of personal services are becoming less and less fair over the years when they review personal service businesses and that this is the

cause of the increase in ratings. Why are reviewers becoming less fair, and in particular, why is there spike in percentage of 5-star ratings, if the overall quality of businesses have stayed the same? This could again be partially explained by the increase in the variance of qualities. There are more high-quality personal services, which could contribute to the increase in 5-star ratings. On the other hand, psychological studies have suggested that personal interactions make it difficult for people to leave negative reviews. This can either be explained by people's fear of retaliation, or increased empathy after experiencing negative experience. This fits our personal experiences: if we go to a beauty spa and receive decent service, but service that may not deserve a 5-star rating, we may end up giving a 5-star rating because we either plan to go back to the service later, so that we want to keep a good relationship with the owner, or because more personal interactions with these businesses make us more likely to forgive mistakes in their services. On the other hand, the steady increase in the proportion of 1-star ratings, which by the time series analysis suggests is at least partially due to users giving out 1-star ratings to businesses that do not actually deserve them, also fits with our experiences. If we go to a beauty spa and have a bad experience there, we are very unlikely to return. In such a case, there is no fear for retaliation and so we are likely to give out a 1-star rating out of dissatisfaction and anger, even though the service may not be of 1-star quality. In any case, the personal nature of personal services seems to be a very relevant factor when trying to explain the observed fairness drop and increase in polarization of rating distribution and average ratings. Whether a causal relationship exists should be the subject of further studies, although comparison with the case of restaurant ratings provides some evidence. In the absence of personal interactions, the proportion of 2-, 3-, and 4-star ratings is still substantial, and the fairness of reviewers of restaurants is definitely not dropping as much as that of personal services.

# 6 Conclusion

We find that for the Yelp dataset, there exists some evidence of rating inflation for both restaurants and businesses offering personal services, on a yearly basis from 2012–2017. In addition, the inflation rate for personal services is greater than that of restaurants. We also find that for restaurants, there is quite a bit of fluctuation with fairness, but it remains around 0.6, while goodness remains constant. For personal services, however, we find that fairness decreases from 2013–2017, while goodness remains roughly equivalent over that time period. This indicates that there is am mostly unclear pattern in fairness for reviewers of restaurants while a downward trend in fairness coupled with an increase in ratings indicates that there likely is the presence of *more* fraudulent, or untrustworthy reviews for businesses offering personal services, supporting our hypothesis.

## 6.1 Limitations

Although we do find some evidence supporting our hypothesis, we do note some limitations. Firstly, we only use the Yelp dataset for this analysis, and it is only limited to a select number

of cities and is certainly insufficient to make any general claims about Yelp as a whole. In addition, it is possible that only Yelp exhibits this trend while other previously mentioned platforms/marketplaces such as eBay, Airbnb, and TripAdvisor do not. Another limitation is that we do not have any indication that this relationship is causal; that is, the decrease in fairness *causes* the increase in ratings–perhaps they are not particularly related. Also, we only used one algorithm to measure fraud; perhaps other algorithms demonstrate that this behavior is relatively normal. Lastly, we only did yearly analysis; perhaps analyzing ratings, fairness, and goodness on a monthly basis would reveal that there is no real relationship, or the relationship is stronger than what we found in our analysis.

# 7 Future Direction

The ultimate goal in studying rating inflation is to be able to identify unfair/fraudulent ratings and to adjust the rating of businesses based on them. The dual objective is to identify fair reviewers whose reviews reveal a lot of information about the true quality of businesses. As discussed before, for future work, it would also be interesting to compare the network-based approach proposed by [Kum+17] to extract quality scores from ratings to methods used in the the statistics and econometrics literature, where quality and user bias are modeled using parametric models.

The problem of rating inflation is also related to fraudulent behavior in rating systems, and as suggested in presentation, a different approach to tackling the fraudulent user detection problem is to apply machine learning methods to the content of reviews, in order to extract features (and lack thereof) of reviews that are highly indicative of fraudulent behavior, e.g. fake reviews on Yelp tend to lack description of physical locations of the business. In addition, it will be interesting to see how one may combine the network-based approach to the natural language processing approach to improve fraud detection.

# References


[Coo15] J. Cook. "Uber's internal charts show how its driver-rating system actually works". In: *Business Insider* (Feb. 2015). URL: http://www.businessinsider.com/leaked-charts-show-how-ubers-driver-rating-system-works-2015-2.

[F+17] A. Filippas, J. Horton, J. M. Golden, et al. *Reputation in the Long-Run*. Tech. rep. CESifo Group Munich, 2017.

[FGH17] A. Fradkin, E. Grewal, and D. Holtz. *The determinants of online review informativeness: Evidence from field experiments on Airbnb*. Tech. rep. Working Paper, 2017.

[Kum+17] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahamanian. "FairJudge: Trustworthy User Prediction in Rating Platforms". In: *arXiv preprint arXiv:1703.10545* (2017).

[LZ16] M. Luca and G. Zervas. "Fake it till you make it: Reputation, competition, and Yelp review fraud". In: *Management Science* 62.12 (2016), pp. 3412–3427.

[NT15] C. Nosko and S. Tadelis. *The limits of reputation in platform markets: An empirical analysis and field experiment.* Tech. rep. National Bureau of Economic Research, 2015.

[ZPB15] G. Zervas, D. Proserpio, and J. Byers. "A first look at online reputation on Airbnb, where every stay is above average". In: (2015).