Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud

Michael Luca
Harvard Business School
<mluca@hbs.edu>

Georgios Zervas
Boston University Questrom School of Business
<zg@bu.edu>

July 20, 2015

Abstract

Consumer reviews are now part of everyday decision-making. Yet, the credibility of these reviews is fundamentally undermined when businesses commit review fraud, creating fake reviews for themselves or their competitors. We investigate the economic incentives to commit review fraud on the popular review platform Yelp, using two complementary approaches and datasets. We begin by analyzing restaurant reviews that are identified by Yelp's filtering algorithm as suspicious, or fake – and treat these as a proxy for review fraud (an assumption we provide evidence for). We present four main findings. First, roughly 16% of restaurant reviews on Yelp are filtered. These reviews tend to be more extreme (favorable or unfavorable) than other reviews, and the prevalence of suspicious reviews has grown significantly over time. Second, a restaurant is more likely to commit review fraud when its reputation is weak, *i.e.*, when it has few reviews, or it has recently received bad reviews. Third, chain restaurants – which benefit less from Yelp – are also less likely to commit review fraud. Fourth, when restaurants face increased competition, they become more likely to receive unfavorable fake reviews. Using a separate dataset, we analyze businesses that were caught soliciting fake reviews through a sting conducted by Yelp. These data support our main results, and shed further light on the economic incentives behind a business's decision to leave fake reviews.

1 Introduction

Consumer review websites such as Yelp, TripAdvisor, and Angie's List have become increasingly popular over the past decade, and now exist for nearly every product and service. Yelp alone contains more than 70 million reviews of restaurants, barbers, mechanics, and other services, and has a market capitalization of roughly four billion dollars. Moreover, there is mounting evidence that these reviews have a direct influence on product sales (see Chevalier and Mayzlin (2006), Luca (2011)).

As the popularity of these platforms has grown, so have concerns that the credibility of reviews can be undermined by businesses leaving fake reviews for themselves or for their competitors. There is considerable anecdotal evidence that this type of cheating is endemic in the industry. For example, the New York Times recently reported on the case of businesses hiring workers on Mechanical Turk – an Amazon-owned crowdsourcing marketplace – to post fake 5-star Yelp reviews on their behalf for as little as 25 cents per review.[1] In 2004, Amazon.ca unintentionally revealed the identities of "anonymous" reviewers, briefly unmasking considerable self-reviewing by book authors.[2]

Despite the major challenge that review fraud poses for firms, consumers, and review platforms alike, little is known about the economic incentives behind it. In this paper, we assemble two novel and complementary datasets from Yelp – one of the industry leaders – to estimate the incidence of review fraud and to understand the conditions under which it is most prevalent. In the first dataset, we focus on reviews that have been written for restaurants in the Boston metropolitan area. Empirically, identifying fake reviews is difficult because the econometrician does not directly observe whether a review is fake. As a proxy for fake reviews, we use the results of Yelp's filtering algorithm that predicts whether a review is genuine or fake. Yelp uses this algorithm to flag suspicious reviews, and to filter them off of the main Yelp page (we have access to all reviews that do not directly violate terms of service, regardless of whether they were filtered.) The exact algorithm is not public information, but the results of the algorithm are. With this in hand, we can analyze the patterns of review fraud on Yelp. In the second data set, we analyze businesses that were caught soliciting fake reviews through a sting conducted by Yelp. We use the second dataset

[1] See "A Rave, a Pan, or Just a Fake?" by David Segal, May'11, available at `http://www.nytimes.com/2011/05/22/your-money/22haggler.html`.

[2] See "Amazon reviewers brought to book" by David Smith, Feb.'04, available at `http://www.guardian.co.uk/technology/2004/feb/15/books.booksnews`.

both to provide support for our use of filtered reviews as a proxy for review fraud, and also to shed further light on the incentives to leave fake reviews.

Overall, roughly 16% of restaurant reviews are filtered by Yelp. While Yelp's goal is to filter fake reviews, the filtering algorithm is imperfect. Therefore, there are both false positives (*i.e.*, filtered reviews that are not fake) and false negatives (*i.e.*, fake reviews that were not filtered). Such misclassification affects our interpretation of filtered reviews in two important ways. First, the rate of fake reviews on Yelp could potentially be higher or lower than the 16% that are filtered. Second, the existence of false positives implies that perfectly honest restaurants may sometimes have their reviews filtered. Similarly, there may be restaurants with no filtered reviews that have successfully committed review fraud. Hence, we do not use filtered reviews to identify specific businesses that committed review fraud. Instead, our main focus is on the economic incentives to commit review fraud. In § 2.4, we provide further empirical support for using filtered reviews as proxy for review fraud by using data on businesses that were known to have committed review fraud.

What does a filtered review look like? We first consider the distribution of star ratings. The data show that filtered reviews tend to be more extreme than published reviews. This observation relates to a broader literature on the distribution of opinion in user-generated content. Li and Hitt (2008) show that the distribution of reviews for many products tends to be bimodal, with reviews tending toward 1- and 5-stars and relatively little in the middle. Li and Hitt (2008) argue that this can be explained through selection if people are more likely to leave a review after an extreme experience. Our results suggest that fake reviews also help to explain the observed prevalence of extreme reviews.

Does review fraud respond to economic incentives, or is it driven mainly by a small number of restaurants that are intent on gaming the system regardless of the situation? If review fraud is driven by incentives, then we should see a higher concentration of fraudulent reviews when the incentives are stronger. Theoretically, restaurants with worse (or less established) reputations have a stronger incentive to game the system. Consistent with this, we find that a restaurant's reputation plays an important role in its decision to leave a fake review. Implementing a difference-in-differences approach, we find that restaurants are less likely to engage in positive review fraud when they have more reviews and when they receive positive shocks to their reputation.

We also find that a restaurant's "offline" reputation matters. In particular, Luca (2011) finds that consumer reviews are less influential for chain restaurants, which already have firmly established reputations built by extensive marketing and branding. Jin and Leslie (2009) find that organizational form also affects a restaurant's performance in hygiene inspections, suggesting that chains face different incentives. We find that chain restaurants are less likely to leave fake reviews relative to independent restaurants. This contributes to our understanding of the ways in which a business's reputation affects its incentives to engage in fraud.

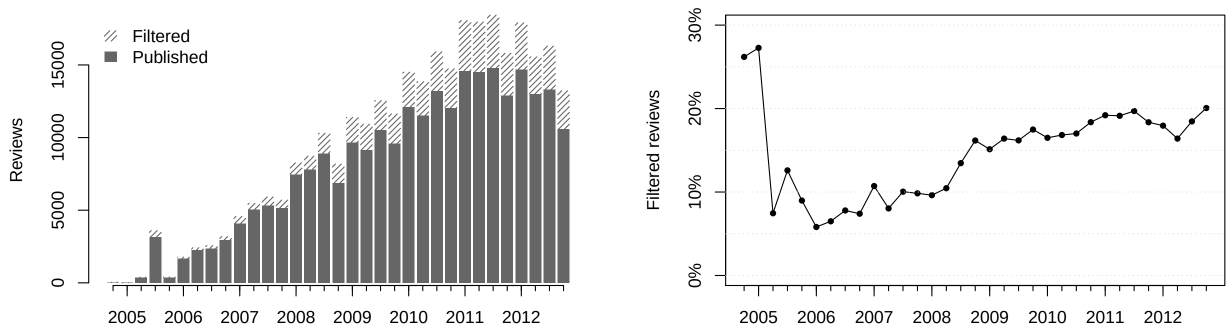
In addition to leaving reviews for itself, a restaurant may commit review fraud by leaving a negative review for a competitor. Again using a difference-in-differences approach, we find that restaurants are more likely to receive negative filtered reviews when there is an increase in competition from independent restaurants serving similar types of food (as opposed to increases in competition by chains or establishments serving different types of food). The entry of new restaurants serving different cuisines has no effect. Our chain results are also consistent with the analysis of Mayzlin et al. (2014) who find that hotels with independently-owned neighbors are more likely to receive negative fake reviews. Overall, our results suggest that independent restaurants are more likely to leave positive fake reviews for themselves, and that fake negative reviews are more likely to occur when a business has an independent competitor. However, it is not necessarily the *same* independent restaurants that are more likely to engage in both positive and negative review fraud.

To reinforce our main interpretation of the results, we then collect a second data set consisting of businesses that were known to have submitted fake reviews. We exploit the fact that Yelp recently conducted a series of sting operations to catch businesses in the act of committing review fraud. During these stings, Yelp responded to businesses that were soliciting fake reviews online. Businesses that were caught soliciting fake reviews were issued a *consumer alert*, a banner which is prominently displayed on offenders' Yelp pages for three months. We identified 126 businesses that had received a consumer alert over the preceding three months as of March 2014. There are no chains among the businesses that were caught leaving fake reviews, which provides strong support for our chain result. Moreover, cheating businesses have significantly lower Yelp ratings and fewer reviews relative to other businesses, consistent with our main findings. The cheating businesses also have much higher rates of algorithmically identified fake reviews relative to our main sample, which provides further support for the validity of using Yelp's algorithmic indicator of fake reviews as a proxy for review fraud. Overall, this supplemental analysis provides a very different methodological approach but yields results that reinforce our main analysis.

Our results contribute to a small but growing literature on the challenges to the quality of user-generated content. The most established literature on the topic of review fraud comes from computer science, and focuses on developing data-mining algorithms that leverage observable review characteristics, such as textual features, and reviewers' social networks, to identify abnormal reviewing patterns (for example, see Feng et al. (2012), Akoglu et al. (2013), Mukherjee et al. (2011, 2012), and Jindal et al. (2010)). A related strand of the literature has focused on constructing a "gold standard" for fake reviews that can be used as training input for fake review classifiers. For example, Ott et al. (2012) construct such a fake review corpus by hiring users on Mechanical Turk – an online labor market – to explicitly write fake reviews. Within the social sciences, researchers have identified a set of issues around the types of reviews that are left. For example, Li and Hitt (2008) investigate selection issues in reviewing, where people who purchase a product choose not to review it. Selection can lead to upward bias for two reasons. First, by revealed preference, people who purchase a product on average like the product better than those who choose not to purchase it. Second, people who choose to leave a review may have different preferences than those who choose not to. In work that is concurrent to but independent from ours, Anderson and Simester (2014) show that in fact, some people who choose not to buy a product still leave a review. Despite these issues with online reviews, Gao et al. (2015) find a high correlation between an offline measure of physician quality and the same physicians' online ratings, suggesting that online reviews – at least in their setting – approximate the distribution of consumer opinion. As mentioned above, our results are also complementary to findings by Mayzlin et al. (2014), who investigate promotional reviews in the context of hotels, doing a cross platform comparison between Expedia and TripAdvisor.

In contrast with prior work, our analysis considers a panel data set, which allows us to perform a variety of new analyses including looking at the growth of review fraud over time, and the role of changes in market structure, competition, and reputation. Ours is also the first study to analyze businesses that are confirmed to have solicited fake reviews, drawing on Yelp's sting and providing a first look at the characteristics of businesses that are known to have attempted review fraud.

Taken in aggregate, our findings suggest that positive review fraud is driven by changes in a



(a) Published and filtered review counts by quarter.

(b) Percentage of filtered reviews by quarter.

Figure 1: Reviewing activity for Boston restaurants from Yelp's founding through 2012.

restaurant's own reputation, while negative review fraud is driven by changing patterns of competition. For platforms looking to curtail gaming, this provides insights into the extent of gaming, as well as the circumstances in which this is more prevalent. Our findings also shed light on the ethical decisions of firms, which are committing fraud in response to changes in economic incentives. Finally, our work is closely related to the literature on organizational form, showing that incentives by independent restaurants are quite different from incentives of chains.

2 Empirical Context and Data

2.1 About Yelp

Our analysis investigates reviews from Yelp , which is a consumer review platform where users can review local businesses such as restaurants, bars, hair salons, and many other services. At the time of this study, Yelp receives approximately 130 million unique visitors per month, and counts over 70 million reviews in its collection. It is the dominant review site for restaurants. For these reasons, Yelp is a compelling setting in which to investigate review fraud. For a more detailed description of Yelp in general, see Luca (2011).

In this analysis, we focus on restaurant reviews in the metropolitan area of Boston, MA. We include in our analysis every Yelp review that was written from the founding of Yelp in 2004 through 2012, other than the roughly 1% of reviews that violate Yelp's terms of service (for example, reviews that contain offensive or discriminatory language). In total, our dataset contains 316,415 reviews for 3,625 restaurants. Of these reviews, 50,486 (approximately 16%) have been filtered by Yelp. Figure 1a displays quarterly totals of published and filtered reviews on Yelp. Yelp's growth in terms of the number of reviews that are posted, and the increasing number of reviews that are filtered, are both evident in this figure. While 16% of reviews are identified as suspicious in our data, there is likely significant variation across platforms and even within a platform in the extent of suspicious activity. This depends on factors including the extent to which reviews influence demand in a given platform as well as the extent to which platforms make it difficult to leave a review. Our findings here are higher than some previous estimates such as Ott et al. (2012). This difference is likely in part due to construction of the training data set used by Ott et al. (2012) (a limitation the authors acknowledge); the fact that Ott et al. (2012) rely only on text features to identify fraudulent reviews whereas Yelp's algorithm uses more markers and can potentially catch more fake reviews; and, finally due to the fact that Ott et al. (2012) use an earlier data sample, which in our data would have had lower rates of review fraud.

2.2 Fake and Filtered Reviews

The main challenge in empirically identifying review fraud is that we cannot directly observe whether a review is fake. The situation is further complicated by the lack of single standard for what makes review "fake." The Federal Trade Commission's truth-in-advertising rules[3] provide some useful guidelines: reviews must be "truthful and substantiated," non-deceptive, and any material connection between the reviewer and the business being reviewed must be disclosed. For example, reviews by the business owner, his or her family members, competitors, a disgruntled ex-employee, or reviewers that have been compensated violate these guidelines unless these connections are disclosed. Not every review can be as unambiguously classified. The case of a business owner nudging consumers by providing them with instructions on how to review his business is in a legal grey area. Most review sites – whose objective is to collect reviews that are as objective as possible – discourage business owners from incentivizing real customers to leave reviews. The reason is simple: businesses are likely to only encourage customers who are having a good experience to leave reviews, resulting – just like review fraud – in a positively biased review sample.

[3]See "Guides Concerning the Use of Endorsements and Testimonials in Advertising," available at `http://ftc.gov/os/2009/10/091005revisedendorsementguides.pdf`.

Our results in this paper are likely driven in part by both processes, which share similar underlying economic incentives and result in an overall positive review bias.

To work around the limitation of not observing fake reviews, we begin by exploiting a unique Yelp feature: Yelp is the only major review site we know of that allows access to *filtered* reviews – reviews that Yelp has classified as illegitimate using a combination of algorithmic techniques, simple heuristics, and human expertise. Filtered reviews are not published on Yelp's main listings, and they do not count towards calculating a business' average star-rating. Nevertheless, a determined Yelp visitor can see a business' filtered reviews after solving a puzzle known as a CAPTCHA.[4] Filtered reviews are, of course, only imperfect indicators of fake reviews. Our work contributes to the literature on review fraud by developing a method that uses an imperfect indicator of fake reviews to empirically identify the circumstances under which fraud is prevalent. This technique translates to other settings where such an imperfect indicator is available, and relies on the following assumption: that the proportion of fake reviews is strictly smaller among the reviews Yelp publishes than among the reviews Yelp filters. We consider this to be a modest assumption whose validity can be qualitatively evaluated. In § 3, we formalize the assumption, suggest a method of evaluating its validity, and use it to develop our empirical methodology for identifying the incentives of review fraud.

2.3 Characteristics of filtered reviews

To the extent that Yelp is a content curator rather than a content creator, there is a direct interest in understanding reviews that Yelp has filtered. While Yelp purposely makes the filtering algorithm difficult to reverse engineer, we are able to test for differences in the observed attributes of published and filtered reviews.

Figure 1b displays the proportion of reviews that have been filtered by Yelp over time. The spike in the beginning results from a small sample of reviews posted in the corresponding quarters. After this, there is a clear upward trend in the prevalence of what Yelp considers to be fake reviews. Yelp retroactively filters reviews using the latest version of its detection algorithm. Therefore, a Yelp

[4] A CAPTCHA is a puzzle originally designed to distinguish humans from machines. It is commonly implemented by asking users to accurately transcribe a piece of text that has been intentionally blurred – a task that is easier for humans than for machines. Yelp uses CAPTCHAs to make access to filtered reviews harder for both humans and machines. For more on CAPTCHAs, see Von Ahn et al. (2003).

review can be initially filtered, but subsequently published (and vice versa.) Hence, the increasing trend seems to reflect the growing incentives for businesses to leave fake reviews as Yelp grows in influence, rather than improvements in Yelp's fake-review detection technology.

Should we expect the distribution of ratings for a given restaurant to reflect the unbiased distribution of consumer opinions? The answer to this question is likely no. Empirically, Hu et al. (2006) show that reviews on Amazon are highly dispersed, and in fact often bimodal (roughly 50% of products on Amazon have a bimodal distribution of ratings). Theoretically, Li and Hitt (2008) point to the fact that people choose which products to review, and may be more likely to rate products after having an extremely good or bad experience. This would lead reviews to be more dispersed than actual consumer opinion. This selection of consumers can undermine the quality of information that consumers receive from reviews.

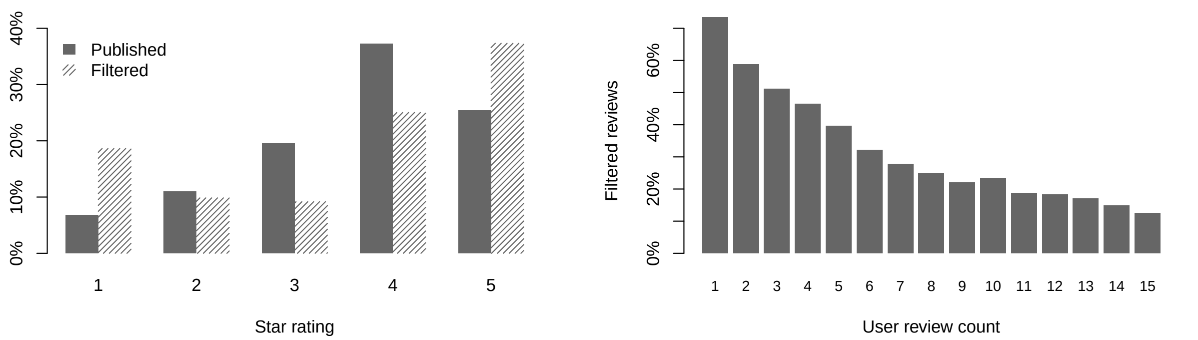
We argue that fake reviews may also contribute to the large dispersion that is often observed in consumer ratings. To see why, consider what a fake review might look like: fake reviews may consist of a business leaving favorable reviews for itself, or unfavorable reviews for its competitors. There is little incentive for a business to leave a mediocre review. Hence, the distribution of fake reviews should tend to be more extreme than that of legitimate reviews. Figure 2a shows the distributions of published and filtered review on Yelp. The contrast between the two distributions is consistent with these predictions. Legitimate reviews are unimodal with a sharp peak at 4 stars. By contrast, the distribution of fake reviews is bimodal with spikes at 1 star and 5 stars. Hence, in this context, fake reviews appear to exacerbate the dispersion that is often observed in online consumer ratings.

In Figure 2b we break down individual reviews by the total number of reviews their authors have written, and display the percentage of filtered reviews for each group. Yelp users who have contributed more reviews are less likely to have their reviews filtered.

We estimate the characteristics of filtered reviews in more detail with the following linear probability model:

$$\text{Filtered}_{ij} = b_i + x_{ij}'\beta + \epsilon_{ij}, \tag{1}$$

where the dependent variable Filtered_{ij} indicates whether the j^{th} review of business i was filtered, b_i is a business fixed effect, and x_{ij} is vector of review and reviewer characteristics including: star rating, (log of) length in characters, (log of) total number of reviewer reviews, and a dummy for



(a) Distribution of stars ratings by published status. (b) Percentage of filtered reviews by user review count.

Figure 2: Characteristics of filtered reviews.

the reviewer having a Yelp-profile picture. We present these results in the first column in Table 1. In line with our observations so far, we find that reviews with extreme ratings are more likely to be filtered – all else equal, 1- and 5-star review are roughly 3 percentage points more likely to be filtered than 3-star reviews. We also find that Yelp's review filter is sensitive to the review and reviewer attributes included in our model. For example, longer reviews, or reviews by users with a larger review count are less likely to be filtered. Beyond establishing some characteristics of Yelp's filter, this analysis also points to the need for controlling for potential algorithmic biases when using filtered reviews as a proxy for fake reviews. We explain our approach in dealing with this issue in § 3.

2.4 Review fraud sting

Our main analysis takes filtered reviews as a proxy for fake reviews. However, one might be concerned that we are reverse engineering Yelp's algorithm rather than analyzing fraud. To support our interpretation, and to provide further insight into the economics of review fraud, we collect and analyze a second dataset consisting of businesses that were caught in the act of soliciting fake reviews.

This second dataset derives from a series of sting operations that Yelp began performing in October 2012. The goal of these stings was to uncover businesses attempting to buy fake reviews.[5]

[5] Yelp's official announcement of the sting operations: `http://officialblog.yelp.com/2012/10/consumer-alerts-because-you-might-like-to-know.html`

Yelp performed the stings by looking for fake review solicitations on classified ads boards like Craigslist – the sting did not rely on the filter in any way. By responding to these solicitations, Yelp was able to expose the identities of the businesses that were attempting to commit review fraud. Businesses which Yelp determined to be buying fake reviews received a notice on their Yelp pages. The notice, known as a *consumer alert*, lasts for at least 90 days, and may be renewed if a business does not seize its efforts to commit review fraud. Yelp continues to perform sporadic sting operations to the present day.

During March 2014, we collected data from all US businesses listed on Yelp. We identified 126 businesses – none of which are Boston restaurants – that had received a consumer alert over the prior 90 days, and collected their entire review histories. In total, the dataset of businesses that received consumer alerts contains 2,233 published and 8,246 filtered reviews.

We use these data for two purposes. First, we cross-validate Yelp's algorithmic results to support our use of filtered reviews as a proxy for fake reviews (in this secion). Then, in § 4.2, we use this dataset for direct evidence on the role of economic incentives in review fraud.

To cross-validate the accuracy of Yelp's filter, we analyze the rate of filtered reviews among businesses that were caught in the sting (which did not affect the filter). We hypothesize that if filtered reviews are a reasonable proxy for fake reviews, then businesses caught in the sting should also have higher rates of filtered reviews. Our sting dataset supports this hypothesis: among businesses that received a consumer alert the average fraction of reviews that are filtered is 79%, compared to 19% for the average Boston restaurant. This suggests that Yelp's filtering algorithm is doing a reasonable job of identifying review fraud and provides support for our use of filtered reviews as a proxy for review fraud.

Next, because we know for certain that the businesses exposed in stings have attempted review fraud, we check whether the characteristics of filtered reviews on this restricted sample of businesses are similar to filtered reviews in our main sample of Boston restaurants. Both in the sample of all Boston restaurants and in the sample of businesses caught in the sting, we find that filtered reviews are more likely to be written by reviewers with less established reputations, as measured by their number of prior reviews, their number of Yelp friends, and whether or not they have a profile photo posted on Yelp. We report these results in Table 2. In the first and second columns of Table 2, focusing on businesses that received a consumer alert, we compare reviewers whose reviews were published with reviewers whose reviews were filtered. Average users whose reviews were filtered have written fewer reviews than reviewers whose reviews were published (3 *vs* 108). They also have fewer friends on Yelp's social network (2 *vs* 66), and are less likely to have a photo associated with their profile (20% *vs* 77%). Our results for Boston restaurants mirror our results for businesses that received a consumer alert: reviewers whose reviews were filtered have written fewer reviews than reviewers whose reviews were published (10 *vs* 137), have fewer friends on Yelp's social network (5 *vs* 61), and are less likely to have a photo associated with their profile (82% *vs* 4%). Therefore, the observable characteristics of the users behind these filtered reviews coincide with the characteristics of users who likely left fake reviews. The similar patterns that emerge across these two samples provides further support for our use of filtered reviews as a proxy for fake.

In summary, even though review fraud is generally unobservable, Yelp's sting operations have enabled us to observe businesses that are known to have attempted review fraud. Using the reviews of these known cheaters, we show that businesses known to attempt review fraud have higher rates of filtered reviews. Moreover, characteristics of filtered reviews, which we use as a proxy for review fraud in our main analysis, do not look systematically different among the businesses that are known to have committed review fraud relative to the population at large. Overall, these results help to cross-validate the results from Yelp's filter and to provide empirical support for our interpretation of filtered reviews.

3 Empirical Strategy

In this section, we introduce our empirical strategy for identifying the economic incentives behind Yelp review fraud. Ideally, if we could recognize fake reviews, we would estimate the following regression model:

$$f^*_{it} = x'_{it}\beta + b_i + \tau_t + \epsilon_{it} \qquad (i = 1 \ldots N; t = 1 \ldots T), \tag{2}$$

where f^*_{it} is the number of fake reviews business i received during period t, x_{it} is a vector of time-varying covariates measuring a business' economic incentives to engage in review fraud, β are the structural parameters of interest, b_i and τ_t are business and time fixed effects, and the ϵ_{it} is an error term. The inclusion of business fixed effects allows us to control for unobservable time-invariant, business-specific incentives for Yelp review fraud. For example, Mayzlin et al. (2014) find that the management structure of hotels in their study is associated with review fraud. To the extent that management structure is time-invariant, business fixed effects allow us to control for this unobservable characteristic. Hence, when looking at incentives to leave fake reviewers over time, we include restaurant fixed effects. However, we also run specifications without a restaurant fixed effect so that we can analyze time-invariant characteristics as well. Similarly, the inclusion of time fixed effects allows us to control for unobservable, common across businesses, time-varying shocks.

As is often the case in studies of gaming and corruption (*e.g.*, see Mayzlin et al. (2014), Duggan and Levitt (2002), and references therein) we do not directly observe f^*_{it}, and hence we cannot estimate the parameters of this model. To proceed, we assume that Yelp's filter possesses some positive predictive power in distinguishing fake reviews from genuine ones. Is this a credible assumption to make? Yelp appears to espouse the view that it is. While Yelp is secretive about how its review filter works, it states that "the filter sometimes affects perfectly legitimate reviews and misses some fake ones, too," but "does a good job given the sheer volume of reviews and the difficulty of its task."[6] Our analysis of businesses that were caught committing review fraud in § 2.4 provides further empirical evidence supporting our use of filtered reviews as a proxy for review fraud. In addition, we suggest a subjective test to assess the assumption's validity: for any business, one can qualitatively check whether the fraction of suspicious-looking reviews is larger among the reviews Yelp publishes, rather than among the ones it filters.

Formally, we assume that $\Pr[\text{Filtered}|\neg\text{Fake}] = a_0$, and $\Pr[\text{Filtered}|\text{Fake}] = a_0 + a_1$, for constants $a_0 \in [0, 1]$, and $a_1 \in (0, 1 - a_0]$, *i.e.*, that the probability a fake review is filtered is strictly greater than the probability a genuine review is filtered. Letting f^*_{itk} be a latent indicator of the k^{th} review of businesses i at time t being fake, we model the filtering process for a single review as:

$$f_{itk} = \alpha_0(1 - f^*_{itk}) + (\alpha_0 + \alpha_1)f^*_{itk} + u_{itk}, \tag{3}$$

where u_{itk} is a zero-mean independent error term. We relax this independence assumption later.

[6]See "What is the filter?", available at `http://www.yelp.com/faq#what_is_the_filter`.

Summing of over all n_{it} reviews for business i in period t we obtain:

$$\sum_{k=1}^{n_{it}} f_{itk} = \sum_{k=1}^{n_{it}} [\alpha_0(1 - f^*_{itk}) + (\alpha_0 + \alpha_1) f^*_{itk} + u_{itk}]$$

$$f_{it} = \alpha_0 n_{it} + \alpha_1 f^*_{it} + u_{it} \tag{4}$$

where u_{it} is a composite error term. Substituting Equation 2 into the above yields the following model:

$$y_{it} = a_0 n_{it} + a_1 \left(x'_{it}\beta + b_i + \tau_t + \epsilon_{it}\right) + u_{it}. \tag{5}$$

It consists of observed quantities, unobserved fixed effects, and an error term. We can estimate this model using a *within* estimator which wipes out the fixed effects. However, while we can identify the reduced-form parameters $a_1\beta$, we cannot separately identify the vector of structural parameters of interest, β. Therefore, we can only test for the *presence* of fraud through the estimates of the reduced-form parameters, $a_1\beta$. Furthermore, since $a_1 \leq 1$, these estimates will be lower bounds to the structural parameters, β.

3.1 Controlling for biases in Yelp's filter

So far, we have not accounted for possible biases in Yelp's filter related to specific review attributes. But what if u_{it} is endogenous? For example, the filter may be more likely to filter shorter reviews, regardless of whether they are fake. To some extent, we can control for these biases. Let z_{itk} be a vector of review attributes. We incorporate filter biases by modeling the error them u_{itk} as follows:

$$u_{itk} = z'_{itk}\gamma + \hat{u}_{itk} \tag{6}$$

where $\hat{u}_{itk}$ is now an independent error term. This in turn suggests the following regression model

$$y_{it} = a_0 n_{it} + a_1 \left(x'_{it}\beta + b_i + \tau_t + \epsilon_{it}\right) + \sum_{k}^{n_{it}} z'_{itk}\gamma + \hat{u}_{it}. \tag{7}$$

In z_{itk}, we include controls for: review length, the number of prior reviews a review's author has written, and whether the reviewer has a profile picture associated with his or her account. As we saw in § 2, these attributes help explain a large fraction of the variance in filtering. A limitation of our work is that we cannot control for filtering biases in attributes that we do not observe, such as the IP address of a reviewer, or the exact time a review was submitted. If these unobserved attributes are endogenous, our estimation will be biased. Equation 7 constitutes our preferred specification.

4 Review Fraud and Own Reputation

This section discusses the main results, which are at the restaurant-month level and presented in Tables 4 and 5. Restaurants in the Boston area receive, on average, approximately 0.1 1-star published reviews per month, and 0.37 published 5-star reviews. Table 3 contains detailed summary statistics – at the restaurant-month level – of all variables in subsequent analyses. Because some of our analyses compare independent and chain restaurants we also report separate summary statistics for these two types of restaurants.

We focus on understanding the relationship between a restaurant's reputation and its incentives to leave a fake review, with the overarching hypothesis that restaurants with a more established or more favorable reputation have less of an incentive to leave fake reviews. While there isn't a single variable that fully captures a restaurant's reputation, there are several relevant metrics, including the number of recent positive and negative reviews, and an indicator for whether the restaurant is a chain, or an independent business. One important feature of Yelp is that it provides more direct measures of a business's reputation than would otherwise be observable.

Our main empirical strategy is a difference-in-differences approach. By analyzing changes in a restaurant's reputation over time, we difference out market-level changes in the propensity to leave a fake review. Specifically, we estimate Equation 7, where we include in the vector x_{it} the following parameters: the number of 1, 2, 3, 4, and 5 star reviews received in period $t-1$; the log of the total number of reviews the business had received up to and including period $t-1$; and the age of the business in period t, measured in (fractional) years. To investigate the incentives of positive review fraud, we estimate specifications with the number of filtered 5-star reviews per restaurant-month as the dependent variable. These results are presented in Table 4. Similarly, to investigate negative review fraud, we repeat the analysis with the number of filtered 1-star reviews per restaurant-month as the dependent variable. We present these results in Table 5. Next, we discuss our results in detail.

4.1 Results: worsening reputation drives positive review fraud

Low ratings increase incentives for positive review fraud, and high ratings decrease them As a restaurant's rating increases, it receives more business (Luca 2011), and hence may have less incentive to game the system. Consistent with this hypothesis, in the first column of Table 4, we observe a positive and significant impact of receiving 1- and 2-star reviews in period $t-1$ on the extent of review fraud in the current period. Conversely, 4- and 5-star published reviews in the previous period lead to a drop in the prevalence of fake reviews in the current period. In other words, a positive change to a restaurant's reputation – whether the result of legitimate, or fake reviews – reduces the incentives of engaging in review fraud, while a negative change increases them.

One way to gauge the economic signficance of these effects is by comparing the magnitudes of the estimated coefficients to the average value of the dependent variable. For example, on average, restaurants in our dataset received approximate 0.1 filtered 5-star reviews per month. Meanwhile, the coefficient estimates in the first column of Table 4 suggest that an additional 1-star review published in the previous period is associated with an extra 0.01 filtered 5-star reviews in the current period, *i.e.*, an increase constituting approximately 10% of the observed monthly average. Furthermore, recalling that most likely $a_0 + a_1 < 1$ (that is to say, Yelp does not identify every single fake review), this number is a conservative estimate for the increase in positive review fraud.

To assess the robustness of these results, we re-estimate the above model including the 6-month leads of published 1, 2, 3, 4, and 5 star reviews counts. We hypothesize that while to some extent restaurants may anticipate reputational shocks, we should see little to know correlation between current review fraud and future shocks to reputation. Column 2 of Table 4 suggests that this is indeed the case. The coefficients of the 6-month lead variables are near zero, and not statistically at conventional significance levels, with the exception of the 6-month lead of 5 star reviews ($p < .05$). Our experiments with short and longer leads did not yield substantially different conclusions.

Having more reviews reduces incentives for positive review fraud As a restaurant receives more reviews, the benefit to each additional review decreases (since Yelp focuses on the average rating). Hence, we expect restaurants to have stronger incentives to submit fake reviews when they have relatively few reviews. To test this hypothesis, we include the logarithm of the current number of reviews a restaurant has in our model. Consistent with this, we find that there exists a negative, statistically significant association between the total number of reviews a business has received up to previous time period, and the intensity of review fraud during the current. Table 4 suggests that restaurants are more likely to engage in positive review fraud earlier in their life-cycles. The coefficient of log Review Count is negative, and statistically significant across all four specifications. These results are consistent with the theory of Branco and Villas-Boas (2011), who predict that market participants whose eventual survival depends on their early performance are more likely to break rules as they enter the market.

Chain restaurants leave fewer positive fake reviews Chain affiliation is an important source of a restaurant's reputation. Local and independent restaurants tend to be less well-known than national chains (defined in this paper as those with 15 or more nationwide outlets). Because of this, chains have substantially different reputational incentives than independent restaurants. In fact, Jin and Leslie (2009) find that chain restaurants maintain higher standards of hygiene as a consequence of facing stronger reputational incentives. Luca (2011) finds that the revenues of chain restaurants are not significantly affected by changes in their Yelp ratings, since chains tend to rely heavily on other forms of promotion and branding to establish their reputation. In addition to the fact that chains receive less benefit from reviews, they may also incur a larger cost if they are caught committing review fraud because they entire brand could be hurt. For example, if one McDonald's gets caught submitting a fake review, all McDonald's may suffer as a result. This observation is consistent with the mechanism identified by Mayzlin et al. (2014). Hence, chains have less to gain from review fraud.

In order to test this hypothesis, we exclude restaurant fixed effects, since they prevent us from identifying chain effects (or, any other time-invariant effect for this matter.) Instead, we implement a random effects (RE) design. One unappealing assumption underlying the RE estimator is the orthogonality between observed variables and unobserved time-invariant restaurant characteristics, *i.e.*, that $E[x'_{it}b_i] = 0$. To address this issue, we follow the approach proposed by Mundlak (1978), which allows for (a specific form) correlation between observables and unobservables. Specifically, we assume that $b_i = \bar{x}_i\gamma + \zeta_i$, and we implement this correction by incorporating the group means of time-variant variables in our model. Empirically, we find that chain restaurants are less likely to engage in review fraud. The estimates of the time-varying covariates in the model remain essentially unchanged compared to the fixed effects specification in the first column of Table 4, suggesting, as Mundlak (1978) highlights, that the RE model we estimate is properly specified. With all controls, the chain coefficient equates to roughly a 5% lower rate of review fraud among chain restaurants.

Other determinants of positive review fraud Businesses can claim their pages on Yelp after undergoing a verification process. Once a business page has been claimed, its owner can respond to consumer reviews publicly or in private, add pictures and information about the business (*e.g.* opening hours and menus), and monitor the number of visitors to the business' Yelp page. 1,964 of all restaurants had claimed their listings by the time we collected our dataset. While we do not observe when these listings were claimed, we expect that businesses with a stronger interest in their Yelp presence, as signaled by claiming their pages, will engage in more review fraud.

To test this hypothesis, we estimate the same random effects model as in the previous section with one additional time-invariant dummy variable indicating whether a restaurant's Yelp page has been claimed or not. The results are shown in the fourth column of Table 4. In line with our hypothesis, we find that businesses with claimed pages are significantly more likely to post fake 5-star reviews. While this finding doesn't fit into our reputational framework, we view it as an additional credibility check that enhances the robustness our analysis.

Negative review fraud Table 5 repeats our analysis with filtered 1-star reviews as the dependent variable. The situations in which we expect negative fake reviews to be most prevalent are qualitatively different from the situations in which we expect positive fake reviews to be most prevalent. Negative fake reviews are likely left by competitors (see Mayzlin et al. (2014)), and may be subject to different incentives (for example, based on the proximity of competitors). We have seen that positive fake reviews are more prevalent when a restaurant's reputation has deteriorated or is less established. In contrast, our results show that negative fake reviews are less responsive to a restaurant's recent ratings, but are still somewhat responsive to the number of reviews that have been left. In other words, while a restaurant is more likely to leave a favorable review for itself as its reputation deteriorates, this does not drive competitors to leave negative reviews. At the same time, both types of fake reviews are more prevalent when a restaurant's reputation is less established, *i.e.* when it has fewer reviews.

Column 2 of Table 5 incorporates 6-month leads of 1, 2, 3, 4, and 5 star review counts. As for the case of positive review fraud, we hypothesize that future ratings should not affect the present incentives of a restaurant's competitors to leave negative fake reviews. Indeed, we find that the coefficients of all 6 lead variables are near zero, and not statistically significant at conventional levels.

As additional robustness checks, we estimate the same RE models as above, which include chain affiliation, and whether a restaurant has claimed its Yelp page as dummy variables. A priori, we expect no association between either of these two indicators and the number of negative fake reviews a business attracts from its competitors. A restaurant cannot prevent its competitors from manipulating its own reviews by being part of chain, or claiming its Yelp page. Indeed, our results, shown in columns 2 & 3 of Table 5, indicate that neither effect is significant, confirming our hypothesis.

4.2 Robustness check: Determinants of fraud using sting data

Our main analysis suggests that a business is more likely to commit positive review fraud when its reputation is weak. To provide further evidence on this, we investigate the reputation of known fraudsters relative to other businesses. In Table 8, we present the average star-rating, published and filtered review counts, and percentage of filtered reviews for businesses that received consumer alerts. Overall, we find that the characteristics of these businesses match our predictions. Consistent with our main analysis, we find that known fraudsters have low ratings and relatively few reviews – on average, 2.6 stars and 18 reviews. In contrast, the average Boston restaurant, which is a priori less likely to have committed review fraud, has 3.5 stars and 86 published reviews. This comparison supports a connection between economic incentives and review fraud. In addition, we observe no chains among the businesses that were caught leaving fake reviews through the sting, providing further support for our chain result. Overall, the sting data reinforce the interpretation of our results by showing that the types of businesses that were caught committing review fraud match the predictions of our main empirical analysis.

5 Review Fraud and Competition

We next turn our attention to analyzing the impact of competition on review fraud. The prevailing viewpoint on negative fake reviews is that they are left by a restaurant's competitors to tarnish its reputation, while we have no similar prediction about the relationship between positive fake reviews and competition.

5.1 Quantifying competition between restaurants

To identify the effect of competition on review fraud, we exploit the fact that the restaurant industry has a relatively high attrition rate. While anecdotal and published estimates of restaurant failure rates vary widely, most reported estimates are high enough to suggest that over its lifetime an individual restaurant will experience competition of varying intensity. In a recent study, Parsa et al. (2005) put the one-year survival probability of restaurants in Columbus, OH at approximately 75%, while an American Express study cited by the same authors estimates it at just about 10%. At the time we collected our dataset, 17% of all restaurants were identified by Yelp as closed.

To identify a restaurant's competitors, we have to consider which restaurant characteristics drive diners' decisions. While location is intuitively one of the factors driving restaurant choice, Auty (1992) finds that food type and quality rank higher in the list of consumers' selection criteria, and therefore, restaurants are also likely to compete on the basis of these attributes. These observations, in addition to the varying incentives faced by chains, motivate a breakdown of competition by chain affiliation, food type, and proximity. To determine whether two restaurants are of the same type we exploit Yelp's fine-grained restaurant categorization. On Yelp, each restaurant is associated with up to three categories (such as Cambodian, Buffets, Gluten-Free, *etc.*) If two restaurants share at least one Yelp category, we deem them to be of the same type.

Next, we need to address the issue of proximity between restaurants and spatial competition. One straightforward heuristic involves defining all restaurants within a fixed threshold distance of each other as competitors. This approach is implemented by Mayzlin et al. (2014), who define two hotels as competitors if they are located with half a kilometer of each other. Bollinger et al. (2010) employ the same heuristic to identify pairs of competing Starbucks and Dunkin Donuts. However, this simple rule may not be as well-suited to defining competition among restaurants. On one hand, location is likely a more important criterion for travelers than for diners. This suggests using a larger threshold to define restaurant competition. On the other hand, the geographic density of restaurants is much higher than that of hotels, or that of Starbucks and Dunkin Donuts branches.[7] Therefore, even a low threshold might cast too wide a net. For example, applying a half kilometer cutoff to our dataset results, on average, in approximately 67 competitors per restaurant. Mayzlin et al. (2014) deal with this issue by excluding the 25 largest (and presumably highest hotel-density) US cities from their analysis. Finally, it is likely that our results will be more sensitive to a particular choice of threshold given that restaurants are closer to each other than hotels. Checking the robustness of our results against too many different threshold values raises the concern of multiple hypothesis testing. Taken together, these observations suggest that a single, sharp threshold rule might not adequately capture the competitive landscape in our setting.

In response to these concerns, a natural alternative is to weigh competitors by their distance. Distance-based heuristics can be generalized using the idea smoothing kernel weights. Specifically, let the impact of restaurant j on restaurant i be:

$$w_{ij} = K\left(\frac{d_{ij}}{h}\right), \tag{8}$$

where d_{ij} is the distance between the two restaurants, K is a kernel function, and h is a positive parameter called the kernel bandwidth. Note that weights are symmetric, *i.e.*, $w_{ij} = w_{ji}$. Then, depending on the choice of K and h, w_{ij} provides different ways to capture the relationship between distance and competition. For example, the threshold heuristic can be implemented using a uniform kernel:

$$K_U(u) = \mathbf{1}_{\{|u|\leq 1\}}, \tag{9}$$

[7]Yelp reports 256 hotels in the Boston area, compared to almost four thousand restaurants.

where $\mathbf{1}_{\{...\}}$ is the indicator function. Using a bandwidth of h, K_U assigns unit weights to competitors within a distance of h, and zero to competitors located farther away.[8]

Similarly, we can define the Gaussian kernel:

$$K_\phi(u) = e^{-\frac{1}{2}u^2}, \tag{10}$$

which produces spatially smooth weights that are continuous in u, and follow the pattern of a Gaussian density function. The kernel bandwidth determines how sharply weights decline, and in empirical applications it is often a subjective, domain-dependent choice. We note that there exists an extensive theoretical literature on optimal bandwidth selection to minimize specific loss functions which is beyond the scope of this work (*e.g.*, see Wand and Jones (1995) and references within).

We approximate the true operating dates of restaurants using their first and last reviews as proxies. Specifically, we take the date of the first review to be the opening date, and if a restaurant is labeled by Yelp as closed, we take the date of the last review as the closing date. While this method is imperfect, we expect that any measurement error it introduces will only attenuate the measured impact of competition. To see this, consider a currently closed restaurant that operated past the date of its last review. Then, any negative fake reviews its competitors received between its miscalculated closing date and its true closing date cannot be attributed to competition. We acknowledge, but consider unlikely, the possibility that restaurants sharply change the rate at which they manipulate reviews during periods we misidentify them as being closed. In this case, measurement error can introduce bias in either direction when estimating competition effects.

Putting together all of the above pieces, we can now operationalize the competition faced by restaurant i. We break down competitors into four categories: same cuisine-type independents, same cuisine-type chains, different cuisine-type independents, and different cuisine-type chains. Let w_{it} be a vector containing these four measures of different kinds of competition. Its first

[8]Kernel functions are usually normalized to have unit integrals. Such scaling constants are inconsequential in our analysis, and hence we omit them for simplicity.

element, which measures competition by independent restaurants of the same type, is defined as:

$$w_{it}^{(1)} = \sum_{i \neq j} w_{ij} \mathbf{1}_{\{\text{independent}_j\}} \mathbf{1}_{\{\text{same type}_{ij}\}} \mathbf{1}_{\{\text{open}_{jt}\}}. \tag{11}$$

The successive indicator functions denote whether j is an independent restaurant, whether i and j share a Yelp category, and whether j is operating at time t. We define the remaining three elements of w_{it} capturing the impact of different type independent restaurants, and same and different type chains in a similar manner.

5.2 Results: competition encourages negative review fraud

We now estimate the impact of competition on review fraud by augmenting our base model of Equation 7 with the vector w_{it}, a set of four time-varying variables measuring competition intensity for restaurant i at time t:

$$y_{it} = a_0 n_{it} + a_1 \left(x_{it}'\beta + w_{it}'\gamma + \sum_{j=1}^{n_{it}} z_{itj}'\gamma + b_i + \tau_t + \epsilon \right), \tag{12}$$

As before, we employ a fixed-effects estimator to rule out any time-invariant endogenous effects. We are especially concerned by purely spatial endogeneity that could arise by incorporating location-dependent variables in our model. For example, restaurants collocated in a shopping mall could exhibit correlated behavior because of their shared location. Our methodology precludes these issues.

We report our results in Table 6. The inclusion of w_{it} does not significantly alter our estimates of the remaining coefficients (as reported in Tables 4 & 5), and hence we omit them from this table for simplicity of presentation. Our first specification estimates the effect of competition on 1-star review fraud using a Gaussian kernel with bandwidth 1km. Using this particular bandwidth, the weight of a restaurant half a kilometer away is approximately 0.6 times the weight of restaurant in exactly the same location, while the weight of any restaurant at a distance of at least 3km becomes negligibly small.

Several interesting patterns emerge in our analysis. First, we find that increased competition from same-type independent restaurants is associated with increased negative review fraud.

Specifically, a unit increase in the competition measure associated with same-type, independent restaurants – which would result, for example, from a same-type, independent competitor opening across the street – is associated with 0.0016 ($p < 0.001$) additional 1-star filtered reviews per month. For the average business, which receives 0.05 1-star filtered reviews each month, this figure roughly translates to a sustained 3% increase. By contrast, a unit increase in competition by chains is associated with -0.0025 ($p < 0.01$) fewer 1-star filtered reviews per month. Empirically, having a new competitor enter can reflect a wide variety of factors, which makes a direct interpretation of these effects difficult. However, whether that competitor is independent or a chain is plausibly exogenous. While there could be other factors that are correlated with an indicator for whether a new competitor enters a neighborhood, we are able to identify the difference between a chain competitor entering and an independent competitor entering. Therefore, while we are hesitant to interpret these as standalone coefficients, we can see that having an independent competitor enter leads to more fake reviews relative to having a chain competitor enter. Overall, our results suggest that increased competition by similar, nearby, independent restaurants has a statistically significant and economically substantive impact on negative review fraud.

By contrast, we find that the effect of increased competition by different food-type independent restaurants is statistically insignificant. In other words, in line with the findings of Auty (1992), our results suggest that restaurants compete for reputation on the basis of both location *and* the type of food they serve. This is in contrast to the results of Mayzlin et al. (2014), who find that hotels compete with their neighbors, but not on the basis of their quality-tiers. While this inconsistency is likely due to differences between the two industries, we cannot reject a methodological explanation since Mayzlin et al. (2014) work in a cross-section setting. Similarly to Mayzlin et al. (2014), we find that, regardless of food type, increased competition by chains has a moderating effect on negative review fraud.

To assess the robustness of our results to kernel choice, we estimate the same model using a uniform kernel with a 1km bandwidth. Our results, shown in the second column of Table 6, remain largely unchanged. Another set of robustness checks with a bandwidth of 0.5km, shown in Table 7, did not yield substantially different outcomes.

In our final set of specifications, reported in the third and fourth columns of Table 6, we test the effect of competition on positive review fraud. Unlike negative fraud, we find no statistically significant relationship between competition intensity and positive fraud regardless of food-type, proximity, or chain affiliation. As before, these results are robust to kernel and bandwidth choice.

6 Conclusion

As crowdsourced information becomes increasingly prevalent, so do concerns that the quality of information can be undermined when businesses game the system. In this paper, we have empirically analyzed review fraud on the popular review website Yelp – both documenting the problem and investigating the conditions under which it is most likely to occur. We show that the problem is widespread – nearly one out of five reviews is marked as fake, by Yelp's algorithm. These reviews tend to be more extreme than other reviews, and are written by reviewers with less established reputations. Moreover, our findings suggest that economic incentives heavily factor into the decision to commit fraud. Organizations are more likely to game the system when they are facing increased competition and when they have poor or less established reputations. In this section, we discuss the implications of our work and directions for future research.

6.1 Improving Fraud Detection Algorithms

Research within computer science has developed a series of tools to identify fake reviews (*e.g.*, see Jindal et al. (2010), Akoglu et al. (2013)). These tools have used features of reviews and reviewers to detect fraud. We see our findings as contributing to this in two ways. First, our findings can directly enter into fraud detection algorithms. For example, a fraud detection algorithm should be quicker to go off in situations where a business is most likely to commit fraud. Second, our findings can help to validate algorithms that are designed to identify fake reviews. Specifically, consider an algorithm that uses text to identify fake reviews. If an algorithm is not identifying fake reviews in the situations where businesses are most likely to leave a fake review (for example, when there is increased competition), then this might raise concerns about the quality of the algorithm.

Clearly, removing low quality content from review platforms is a laudable goal. Yet, one challenge to implementing a review filtering algorithm is that businesses may feel that it is unfair when a legitimate review is removed. This type of complaint is often made toward review platforms by seemingly honest businesses. While we have cross-validated the results of Yelp's algorithm, such an algorithm is necessarily an imperfect science – leading both to false positives and false negatives. This issue is further complicated by businesses that have anecdotally claimed that review filters are stricter on non-advertisers than on advertisers.

While a complete analysis of this issue is beyond the scope of this paper, the results from the second column of Table 1 may shed some light on the issue. In that specification, we include an indicator for whether a restaurant advertises with Yelp, and also interact this variable with the main independent variables in the specification – the characteristics of the review, and the characteristics of the reviewer. While we use this mainly as a control (to allay concerns that Yelp's algorithm may differentially affect advertisers), these interaction variables can also help to provide more direct evidence of how advertisers are treated relative to non-advertisers. We find that none of the advertiser interaction effects are statistically significant, while the remaining coefficients are essentially unchanged in comparison to those in Equation 1. While we cannot reject hypotheses related to differential treatment, we find that neither 1- nor 5-star reviews were significantly more or less likely to be filtered for businesses that were advertising on Yelp at the time we collected our dataset. Another limitation of this interpretation is that we do not observe the complete historic record of which businesses have advertised on Yelp, and hence we can only test for discrimination in favor of (or, against) current Yelp advertisers. As a robustness check, we repeat this analysis using only 2012 data, so that the advertising indicator and the reviews we analyze are in line. Our results, presented in the third column of Table 1, remain unchanged.

An important direction for future research is to investigate both real and perceived fairness issues related to algorithmic fraud detection.

6.2 Mechanisms to Reduce Fraud

The findings of this paper suggest that there is an extensive amount of systematic review fraud in online review platforms. It is therefore of first-order importance for platform designers to develop mechanisms to reduce the amount of fraud. While there is no perfect mechanism that will eliminate all review fraud, we believe that there are several main mechanisms that can help to reduce fraud. Here, we suggest four possible approaches to do so. We hope that future work will develop these and other mechanisms.

First, fraud detection algorithms can identify fake reviews, which allows the market designer to eliminate the offending reviews (and potentially to punish the offending business). The advantage of this is that it is a simple, low-cost approach. However, algorithms can seem to lack transparency and are inherently noisy. Second, the market designer can only allow verified purchases to leave a review (Mayzlin et al. 2014). While this can reduce the prevalence of fake reviews, it may also have the unintended consequence of reducing the amount of legitimate content. Third, market designers can implement stings (such as the ones studied in this paper). However, this is costly to implement and can only be done in situations in which businesses are soliciting fake reviews (rather than writing them themselves). Fourth, one can imagine leveraging behavioral economics to reduce fraud. For example, if businesses do not inherently see review fraud as an ethical decision, the platform can highlight the ethical element of decision making every time someone wants to leave a review.

References

Akoglu, Leman, Rishi Chandy, Christos Faloutsos. 2013. Opinion Fraud Detection in Online Reviews by Network Effects. *ICWSM*.

Anderson, Eric T, Duncan I Simester. 2014. Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research* **51**(3) 249–269.

Auty, Susan. 1992. Consumer choice and segmentation in the restaurant industry. *Service Industries Journal* **12**(3) 324–339.

Bollinger, B., P. Leslie, A. Sorensen. 2010. Calorie Posting in Chain Restaurants. Tech. rep., National Bureau of Economic Research.

Branco, Fernando, J Miguel Villas-Boas. 2011. Competitive vices. *Available at SSRN 1921617* .

Chevalier, Judith a, Dina Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of marketing research* **43**(3) 345–354.

Duggan, Mark, Steven D Levitt. 2002. Winning isn't everything: Corruption in sumo wrestling. *The American Economic Review* **92**(5) 1594–1605.

Feng, Song, Ritwik Banerjee, Yejin Choi. 2012. Syntactic Stylometry for Deception Detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 171–175.

Gao, Guodong Gordon, Brad N Greenwood, Jeff McCullough, Ritu Agarwal. 2015. Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? *Management Information Systems Quarterly (Forthcoming)* .

Hu, Nan, Paul A Pavlou, Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. *Proceedings of the 7th ACM conference on Electronic commerce*. ACM, 324–330.

Jin, G.Z., P. Leslie. 2009. Reputational Incentives for Restaurant Hygiene. *American Economic Journal: Microeconomics* **1**(1) 237–267.

Jindal, Nitin, Bing Liu, Ee-Peng Lim. 2010. Finding unusual review patterns using unexpected rules. *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 1549–1552.

Li, Xinxin, Lorin M Hitt. 2008. Self-Selection and Information Role of Online Product Reviews. *Information Systems Research* **19**(4) 456–474.

Luca, Michael. 2011. Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School NOM Unit Working Paper* (12-016).

Mayzlin, Dina, Yaniv Dover, Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104**(8) 2421–55.

Mukherjee, Arjun, Bing Liu, Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*. ACM, 191–200.

Mukherjee, Arjun, Bing Liu, Junhui Wang, Natalie Glance, Nitin Jindal. 2011. Detecting group review spam. *Proceedings of the 20th international conference companion on World wide web*. ACM, 93–94.

Mundlak, Yair. 1978. On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society* 69–85.

Ott, Myle, Claire Cardie, Jeff Hancock. 2012. Estimating the Prevalence of Deception in Online Review Communities. *Proceedings of the 21st international conference on World Wide Web*. ACM, 201–210.

Parsa, HG, John T Self, David Njite, Tiffany King. 2005. Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly* **46**(3) 304–322.

Shleifer, Andrei. 2004. Does competition destroy ethical behavior? *American Economic Review* **94**(2) 414–418.

Von Ahn, Luis, Manuel Blum, Nicholas J Hopper, John Langford. 2003. CAPTCHA: Using hard AI problems for security. *Advances in Cryptology-EUROCRYPT 2003*. Springer, 294–311.

Wand, Matt P, M Chris Jones. 1995. *Kernel smoothing*, vol. 60. Chapman & Hall/CRC.

Table 1: Characteristics of Boston restaurant filtered reviews.

	(1) All reviews	(2) All reviews	(3) 2012 reviews
Stars (the reference level is 3 stars)			
1	0.035***	0.035***	0.052***
	(8.50)	(8.29)	(6.49)
2	−0.022***	−0.021***	−0.018***
	(−10.10)	(−9.67)	(−3.55)
4	0.0031*	0.0030*	0.0085*
	(2.24)	(2.12)	(2.44)
5	0.026***	0.027***	0.026***
	(11.38)	(11.21)	(5.79)
log(Review length)	−0.016***	−0.016***	−0.028***
	(−10.83)	(−10.03)	(−8.17)
log(User review count)	−0.096***	−0.097***	−0.081***
	(−84.45)	(−81.19)	(−36.56)
User has photo	−0.41***	−0.41***	−0.35***
	(−156.60)	(−152.17)	(−82.91)
Stars × Yelp Advertiser			
1		−0.0054	0.036
		(−0.34)	(1.40)
2		−0.0058	−0.014
		(−0.61)	(−0.72)
4		0.00076	−0.0076
		(0.13)	(−0.58)
5		−0.011	−0.012
		(−1.46)	(−0.75)
log(Review length) × Yelp Advertiser		−0.011	−0.012
		(−1.72)	(−1.04)
log(User review count) × Yelp Advertiser		0.0073	0.0083
		(1.80)	(1.09)
User has photo × Yelp Advertiser		−0.0012	−0.0075
		(−0.11)	(−0.45)
N	316415	316415	66174
R^2	0.43	0.43	0.33

Note: The dependent variable is a binary indicator of whether a specific review was filtered. All models include business fixed effects. Cluster-robust t-statistics (at the individual business level) are shown in parentheses.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 2: Characteristics of Yelp users whose reviews were published compared to Yelp users whose reviews were filtered.

	Consumer alert recipients		Boston restaurants	
	Published	Filtered	Published	Filtered
User review count	108	3	137	10
User friend count	66	2	61	5
User has profile photo	77%	20%	82%	4%

Table 3: Summary statistics at the restaurant-month level.

	Independents	Chains	All
Published reviews per month			
1-star	0.10	0.10	0.10
	(0.37)	(0.37)	(0.37)
5-star	0.39	0.17	0.37
	(1.03)	(0.59)	(1.00)
Filtered reviews per month			
1-star	0.05	0.03	0.05
	(0.28)	(0.20)	(0.27)
5-star	0.11	0.05	0.10
	(0.43)	(0.26)	(0.42)
Reviews with user profile photo per month			
1-star	0.07	0.08	0.07
	(0.29)	(0.31)	(0.29)
5-star	0.31	0.14	0.29
	(0.86)	(0.51)	(0.84)
User review count			
1-star	8.30	9.99	8.45
	(62.98)	(65.73)	(63.24)
5-star	40.64	21.63	38.93
	(165.47)	(113.15)	(161.55)
Sum of review lengths (in characters) per month			
1-star	127.72	105.90	125.76
	(532.31)	(492.06)	(528.85)
5-star	321.96	124.85	304.22
	(961.84)	(492.84)	(931.09)
Independent competitors 0.5km			
Same food type	7.52	6.77	7.46
	(15.59)	(11.30)	(15.26)
Different food type	41.59	54.43	42.75
	(43.08)	(49.04)	(43.81)
Chain competitors within 0.5km			
Same food type	0.77	1.80	0.86
	(2.05)	(3.40)	(2.23)
Different food type	5.59	8.46	5.85
	(7.88)	(8.85)	(8.01)
Observations	167594	16572	184166

Table 4: The effect of own reputation on positive (5-star) review fraud.

	(1)	(2)	(3)	(4)
Dependent variable:	5-star filtered reviews per month	5-star filtered reviews per month	5-star filtered reviews per month	5-star filtered reviews per month
1 month lag				
1-star reviews	0.012***	0.013***	0.012***	0.012***
	(4.78)	(4.77)	(4.69)	(4.70)
2-star reviews	0.007**	0.006**	0.007**	0.007**
	(3.10)	(2.92)	(3.18)	(3.18)
3-star reviews	−0.000	0.000	−0.000	−0.000
	(−0.17)	(0.29)	(−0.19)	(−0.19)
4-star reviews	−0.004***	−0.004**	−0.004***	−0.004***
	(−3.31)	(−3.03)	(−3.38)	(−3.38)
5-star reviews	−0.014***	−0.011***	−0.015***	−0.015***
	(−4.83)	(−4.48)	(−5.00)	(−5.00)
log Review count	−0.021***	−0.020***	−0.020***	−0.020***
	(−7.89)	(−7.37)	(−7.95)	(−7.88)
6 month lead				
1-star reviews		−0.004		
		(−1.80)		
2-star reviews		0.002		
		(1.06)		
3-star reviews		−0.000		
		(−0.22)		
4-star reviews		−0.001		
		(−1.26)		
5-star reviews		−0.007*		
		(−2.55)		
Business age (years)	0.006*	0.005*	0.031***	0.031***
	(2.57)	(2.35)	(3.55)	(3.54)
Chain restaurant			−0.008**	−0.008**
			(−3.28)	(−3.28)
Claimed Yelp listing				0.012***
				(4.80)
Model	Fixed effects	Fixed effects	Random effects	Random effects
N	180912	162063	180912	180912
R^2	0.66	0.68	0.67	0.67

Note: Cluster-robust t-statistics (at the individual business level) are shown in parentheses. All specifications contain controls for various review attributes which are not shown. The number of observations N is smaller than that reported in Table 3 since lag and lead variables are included.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 5: The effect of own reputation on negative (1-star) review fraud.

	(1)	(2)	(3)	(4)
Dependent variable:	1-star filtered reviews per month	1-star filtered reviews per month	1-star filtered reviews per month	1-star filtered reviews per month
1 month lag				
1-star reviews	0.001	0.002	−0.000	−0.000
	(0.31)	(0.70)	(−0.18)	(−0.18)
2-star reviews	−0.003	−0.003	−0.003*	−0.003*
	(−1.71)	(−1.96)	(−2.00)	(−2.00)
3-star reviews	−0.000	−0.000	−0.001	−0.001
	(−0.29)	(−0.19)	(−0.56)	(−0.56)
4-star reviews	−0.000	−0.001	−0.001	−0.001
	(−0.59)	(−1.09)	(−0.80)	(−0.80)
5-star reviews	0.001	0.001	0.001	0.001
	(1.26)	(1.28)	(0.94)	(0.94)
log Review count	−0.003*	−0.003**	−0.004***	−0.004***
	(−2.51)	(−2.64)	(−3.35)	(−3.30)
6 month lead				
1-star reviews		0.002		
		(1.13)		
2-star reviews		0.000		
		(0.30)		
3-star reviews		−0.000		
		(−0.20)		
4-star reviews		0.001		
		(1.47)		
5-star reviews		0.000		
		(0.67)		
Business age (years)	0.002	0.001	−0.000	−0.000
	(1.43)	(1.17)	(−0.00)	(−0.01)
Chain restaurant			−0.002	−0.002
			(−1.83)	(−1.82)
Claimed Yelp listing				0.001
				(0.87)
Model	Fixed effects	Fixed effects	Random effects	Random effects
N	180912	162063	180912	180912
R^2	0.68	0.69	0.68	0.68

Note: Cluster-robust t-statistics (at the individual business level) are shown in parentheses. All specifications contain controls for various review attributes which are not shown. The number of observations N is smaller than that reported in Table 3 since lag and lead variables are included.

Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 6: The effect of competition on review fraud (kernel bandwidth 1km.)

	(1) Gaussian	(2) Uniform	(3) Gaussian	(4) Uniform
Dependent variable:	1-star filtered reviews per month	1-star filtered reviews per month	5-star filtered reviews per month	5-star filtered reviews per month
Independent competitors				
Same food type	0.0016***	0.0013***	0.00094	0.00065
	(3.33)	(3.32)	(1.34)	(1.06)
Different food type	0.000074	0.000068	−0.00029	−0.00013
	(0.64)	(0.77)	(−1.43)	(−0.79)
Chain competitors				
Same food type	−0.0030*	−0.0025**	−0.0023	−0.0023
	(−2.53)	(−2.64)	(−1.20)	(−1.43)
Different food type	−0.0011*	−0.0011**	0.00076	0.00028
	(−2.18)	(−2.78)	(0.88)	(0.42)
N	180912	180912	180912	180912
R^2	0.68	0.68	0.66	0.66

Note: Cluster-robust *t*-statistics (at the individual business level) are shown in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 7: The effect of competition on review fraud (kernel bandwidth 0.5km.)

	(1) Gaussian	(2) Uniform	(3) Gaussian	(4) Uniform
Dependent variable:	1-star filtered reviews per month	1-star filtered reviews per month	5-star filtered reviews per month	5-star filtered reviews per month
Independent competitors				
Same food type	0.0012***	0.00094**	0.00044	0.00030
	(3.43)	(3.26)	(0.89)	(0.77)
Different food type	0.000043	−0.000058	−0.00031	−0.00022
	(0.41)	(−0.67)	(−1.63)	(−1.40)
Chain competitors				
Same food type	−0.0026**	−0.0016*	−0.0021	−0.00082
	(−2.74)	(−2.29)	(−1.38)	(−0.72)
Different food type	−0.0010*	−0.00049	0.0013	0.0010
	(−2.33)	(−1.49)	(1.75)	(1.80)
N	180912	180912	180912	180912
R^2	0.68	0.68	0.66	0.66

Note: Cluster-robust *t*-statistics (at the individual business level) are shown in parentheses.
Significance levels: * p<0.05, ** p<0.01, *** p<0.001.

Table 8: Characteristics of Boston restaurants compared to businesses that have received a Yelp consumer alert.

	Boston restaurants	Consumer alert recipients
Rating	3.5	2.6
Review count	86	18
Filtered count	17	65
Pct. filtered	19%	79%