Fake It Till You Make It:

Reputation, Competition, and Yelp Review Fraud

Michael Luca
Harvard Business School
mluca@hbs.edu

Georgios Zervas[†]
Boston University
zg@bu.edu

November 8, 2013

Abstract

Consumer reviews are now a part of everyday decision-making. Yet the credibility of reviews is fundamentally undermined when business-owners commit review fraud, either by leaving positive reviews for themselves or negative reviews for their competitors. In this paper, we investigate the extent and patterns of review fraud on the popular consumer review platform Yelp.com. Because one cannot directly observe which reviews are fake, we focus on reviews that Yelp's algorithmic indicator has identified as fraudulent. Using this proxy, we present four main findings. First, roughly 16 percent of restaurant reviews on Yelp are identified as fraudulent, and tend to be more extreme (favorable or unfavorable) than other reviews. 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 leave unfavorable reviews for competitors. Taken in aggregate, these findings highlight the extent of review fraud and suggest that a business's decision to commit review fraud responds to competition and reputation incentives rather than simply the business's ethics.

[†]Part of this work was completed while the author was supported by a Simons Foundation Postdoctoral Fellowship.

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 any product or service imaginable. Yelp alone contains more than 30 million reviews of restaurants, barbers, mechanics, and other services, and has a market capitalization in excess of 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), Zhu and Zhang (2010)).

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]

In recent years, review fraud has emerged as the preeminent threat to the sustainability of this type of user generated content. Despite the major challenge that review fraud poses for firms and consumers alike, little is known about the economic incentives behind it. In this paper, we assemble a novel dataset from Yelp – one of the industry leaders – to estimate the incidence of review fraud among restaurants in the Boston metropolitan area, and to understand the conditions under which it is most prevalent. Specifically, we provide evidence on the impact of restaurants' evolving reputation, and the competition it faces, and its decision to engage in positive and negative review fraud.

A growing computer science literature has developed data-mining algorithms which leverage observable review characteristics, such as textual features, and reviewers' social networks, to identify abnormal reviewing patterns (for example, see Akoglu et al. (2013)). 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

[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`.

review corpus by hiring users on Mechanical Turk – an online labor market – to explicitly write fake reviews.

In this paper, we analyze fake reviews from a different perspective, and investigate a business's incentives to engage in review fraud. We analyze these issues using data from Yelp.com, focusing 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 fake 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.

Overall, roughly 16% of reviews are identified by Yelp as fake and are subsequently filtered. 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 tend to be bimodal, with reviews tending toward 1- and 5-stars and relatively little in the middle. Their argument is that this can be explained through selection if people are more likely to leave a review after an extreme experience. Our results suggest another factor that increases the proportion of extreme reviews is the prevalence of fake reviews.

Does review fraud respond to economic incentives, or is it driven mainly by a small number of unethical 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 is a strong predictor of its decision to leave a fake review. We also find that restaurants are more likely to engage in fraud when they have fewer reviews. This result motivates an additional explanation for the often-observed concentration of high ratings earlier in a business' life-cycle. Previous work attributed this concentration to interactions between a consumer's deci-

sion to contribute a review, and prior reviews by other consumers (see Moe and Schweidel (2012), Godes and Silva (2012).) Our work suggests businesses' increased incentives to manipulate their reviews early on as an additional explanation for this observation. We also find that restaurants that have recently received bad ratings engage in more positive review fraud. By contrast, we find no link between negative review fraud and reputation. To some extent this is to be expected, since these are fake reviews that are presumably left by a business' competitors.

We also find that a restaurant's "offline" reputation is a determinant of its decision to seek positive fake reviews. 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. Consistent with this, 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' reputation affects its incentives with respect to review fraud.

In addition to leaving reviews for itself, a restaurant may commit review fraud by leaving a negative review for a competitor. Empirically, we find that increased competition by nearby independent restaurants serving similar types of food is positively associated with negative review fraud. Interestingly, increased competition by nearby restaurants serving different types of food are not a significant predictor of negative review fraud. A potential explanation for this finding is that restaurants tend to compete based on a combination of location and cuisine. This explanation echoes the survey of Auty (1992) in which diners ranked food type and quality highest among a list of restaurant selection criteria, suggesting that restaurants do no compete on the basis of location alone. Our results are also consistent with the analysis of Mayzlin et al. (2012) who find that hotels with independently-owned neighbors are more likely to receive negative fake reviews.

Overall, our findings suggest that positive review fraud is primarily driven by changes in a restaurant's own reputation, while negative review fraud is primarily driven by changing patterns of competition. For platforms looking to curtail gaming, our results provide insights both into the extent of gaming, as well as the circumstances in which this is more prevalent. Our results also provide insight into our understanding of ethical decision making by firms, which we show to be a function of economic incentives rather than a function of unethical firms. 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 Related work

There is now extensive evidence that consumer reviews have a causal impact on demand in industries ranging from books to restaurants to hotels, among others (Chevalier and Mayzlin (2006), Luca (2011), Ghose et al. (2012)). However, there is considerably less agreement about whether reviews contain trustworthy information that customers should use. For example, Li and Hitt (2008) argue that earlier reviewers tend to be more favorable toward a product relative to later reviewers, making reviews less representative of the typical buyer and hence less reliable. Looking at movie sales and reviews, Dellarocas et al. (2010) provide evidence that consumers are more likely to review niche products, but at the same time are more likely to leave a review if many other reviewers have contributed, suggesting other types of bias that may appear in reviews. In principle, if one knows the structure of any given bias that reviews exhibit, Dai et al. (2012) argue that the review platform can improve the reliability of information by simply adjusting and reweighing reviews to make them more representative.

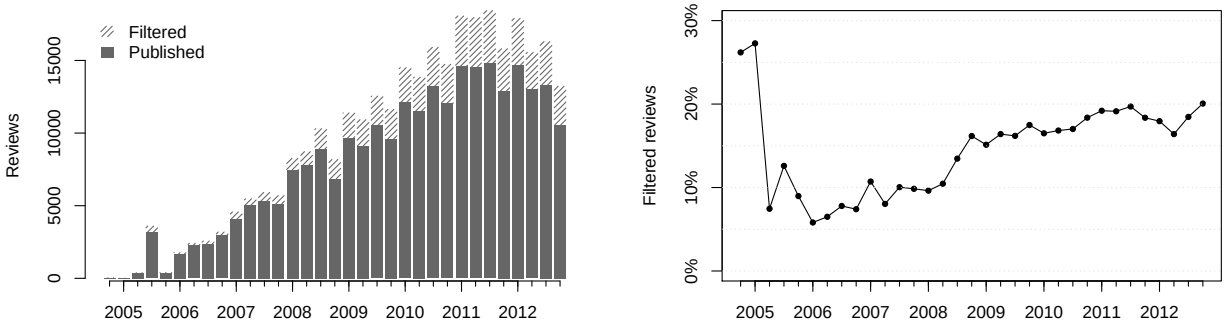
Perhaps the most direct challenge to the reliability of online information is the possibility of leaving fake reviews. Theoretically, Dellarocas (2006) provides conditions under which reviews can still be informative even if there is gaming. In concurrent but independent work, Anderson and Simester (2013) show that many reviews on an online apparel platform are written by customers who have no purchase record. These reviews tend to be more negative than other reviews. In contrast with our setting - where we look for economic incentives to commit review from - their work highlights reviews that are written by people without any clear financial incentive to leave a fake review. Within computer science, a growing literature has focused on the development of machine learning algorithms to identify review fraud. Commonly these algorithms rely either on data mining techniques to identify abnormal reviewing patterns, or employ supervised learning methods trained on hand-labelled examples. Some examples are Ott et al. (2012), Feng et al. (2012), Akoglu et al. (2013), Mukherjee et al. (2011, 2012), Jindal et al. (2010).

To the best of our knowledge, there exists only one other paper that analyzes the economic incentives of review fraud, and uses a different empirical approach and setting. Mayzlin et al.

(2012) exploit an organizational difference between Expedia and TripAdvisor (which are spin-offs of the same parent company with different features[3]) to study review fraud by hotels: while anyone can post a review on TripAdvisor, Expedia requires that a guest has "paid and stayed" before submitting a review. The authors observe that Expedia's verification mechanism increases the cost of posting a fake review. The study finds that independent hotels tend to have a higher proportion of five-star reviews on TripAdvisor relative to Expedia and competitors of independent hotels tend to have a higher proportion of one-star reviews on TripAdvisor relative to Expedia. Their argument is that there are many reasons that TripAdvisor reviews may be different from Expedia reviews, and many reasons that independent hotels may receive different reviews from chain hotels. However, they argue that if independent hotels receive favorable treatment relative to chains on TripAdvisor but not on Expedia, then this suggests that these reviews on TripAdvisor are fraudulent. The validity of this measure rests on the assumption that differences in the distributions of ratings across the two sites, and between different types of hotels are due to review fraud. In our work, we do not rely on this assumption as we are identifying our effects entirely within a single review platform. Our work also difference in that we are able to exploit the panel nature of our data, and analyze the within restaurant role of reputation in the decision to commit review fraud. This allows us to show not only that certain types of restaurants are more likely to commit review fraud, but also that even within a restaurant, economic conditions influence the ultimate decision to commit review fraud.

Finally, we briefly mention the connection between our work, the literature on statistical fraud detection (*e.g.*, see Bolton and Hand (2002)), and the related line of research on statistical models of misclassification in binary data (*e.g.*, see Hausman et al. (1998)). These methods have been applied to uncover various types of fraud, such as fraudelent insurance claims, and fraudulent credit card transactions. The key difference of our work is that our end goal isn't to identify individual fraudulent reviews – instead we wish exploit a noisy signal of review fraud to investigate the incentives behind it.

[3]See `http://www.iac.com/about/history`.



(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.

3 Description of Yelp and Filtered Reviews

3.1 About Yelp

Our analysis investigates reviews from the website Yelp.com, 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 100 million unique visitors per month, and counts over 30 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. Both Yelp's growth in terms of the number of reviews that are posted on it, and the increasing number of reviews that are filtered are evident in this figure.

3.2 Fake and Filtered Reviews

The main challenge in empirically identifying review fraud is that we not 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[4] 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, reviewers that have been compensated, or a disgruntled ex-employee are considered fake (and, by extension illegal) 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 – frown upon such interventions, and encourage business owners to avoid them.

To work around the limitation of not observing fake reviews we exploit 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.[5] 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 the reviews Yelp filters. We consider this to be a modest assumption whose validity can be qualitatively evaluated. In § 4, we formalize the assumption, suggest a method of evaluating its validity, and use

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

[5]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).

it to develop our empirical methodology for identifying the incentives of review fraud.

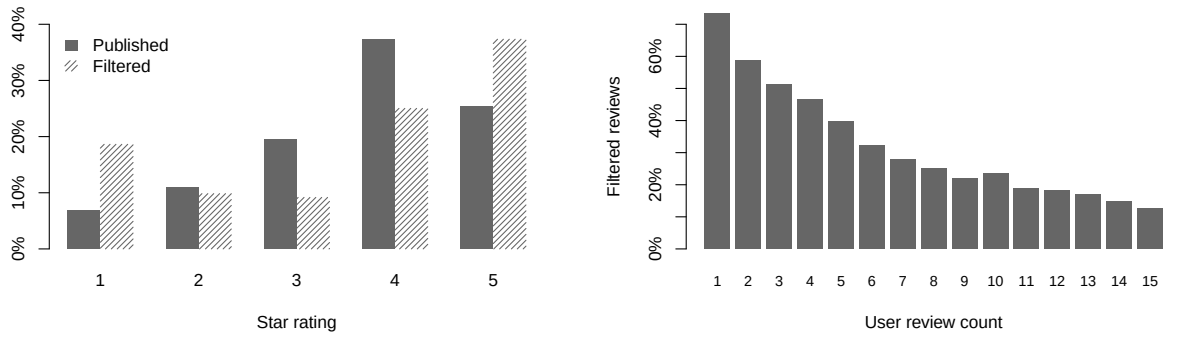
3.3 Characteristics of filtered reviews

To the extent that Yelp is a content aggregator rather than a content creator, there is a direct interest in understanding reviews that Yelp has filtered. While Yelp purposely makes the filtering algorithm hard 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's retroactively filters reviews using the latest version of its detection algorithm. Therefore, a Yelp 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,



(a) Distribution of stars ratings by published status.

(b) Percentage of filtered reviews by user review count.

Figure 2: Characteristics of filtered reviews.

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. The trend we observe suggests that 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 by using 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 the reviewer having a Yelp-profile picture. The estimation results are shown in the first column of 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 § 4.

3.4 Filtered Reviews and Advertising on Yelp

Local business advertising constitutes Yelp's major revenue stream. Advertisers are featured on Yelp search results pages in response to relevant consumer queries, and on the Yelp pages of similar, nearby businesses. Furthermore, when a business purchases advertising, Yelp removes competitors' ads from that business' Yelp page. Over the years, Yelp has been the target of repeated complaints alleging that its filter discriminates in favor of advertisers, going in some cases as far as claiming that the filter is nothing other than an extortion mechanism for advertising revenue.[6] Yelp has denied these allegations, and successfully defended itself in court when lawsuits have been brought against it (for example, see Levitt v. Yelp Inc., and Demetriades v. Yelp Inc.) If such allegations were true, they would raise serious concerns as to the validity of using filtered reviews as a proxy for fake reviews in our analysis.

Using our dataset we are able to cast further light on this issue. To do so we exploit the fact that Yelp publicly discloses which businesses are current advertisers (it does not disclose which businesses were advertisers historically.) Specifically, we augment Equation 1 by interacting the x_{it} variables with a dummy variable indicating whether a business was a Yelp advertiser at the time we obtained our dataset. The results of estimating this model are shown in second column of Table 1. 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. This suggests, for example, 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.

A limitation of this analysis 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. To address this limitation, we obtained a complete record of advertising contracts for the businesses in our study directly through Yelp. Specifically, for each business in our dataset, the Yelp-supplied dataset contains the starting, and ending date of all advertising contracts the business entered.[7] In Table 2, we repeat our analysis of which reviews

[6] See "No, Yelp Doesnt Extort Small Businesses. See For Yourself.", available at: `http://officialblog.yelp.com/2013/05/no-yelp-doesnt-extort-small-businesses-see-for-yourself.html`.

[7] We have no information on the pricing associated with these contracts.

are likely to be filtered replacing the indicator of current Yelp advertisers, with a time-varying indicator of whether a business was a Yelp advertiser at the time each of its reviews were submitted. Since this new advertising indicator is time-varying we can include it separately in a fixed-effects specification. Our results are displayed in the first column of Table 2. The advertising indicator is statistically, and substantively insignificant. In the second column of Table 2, we expand our analysis by interacting the advertising indicator with various review characteristics. Our results remain virtually unchanged. Neither the main effect nor any of its interactions are significant predictors of a review being filtered.

These analyses, using both data we collected, and data Yelp provided us with, suggest that once we control for observable review characteristics, Yelp's current implementation of the filtering algorithm does not treat advertisers' reviews in a manner different to non-advertisers' reviews. While we have no direct knowledge of how Yelp's filtering algorithm works, the lack of filtering biases associated with advertising increases our confidence in using filtered reviews as an unbiased, albeit imperfect, proxy for fake reviews.

4 Empirically Strategy

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

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

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. (2012) 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. (2012), 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's 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."[8] 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 $a_1 \in (0, 1]$, *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:

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

where u_{itk} is a zero-mean independent error term. We relax this independence assumption later. Summing of over all n_{it} reviews for business i in period t we obtain:

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

$$f_{it} = \alpha_0 n_{it} + \alpha_1 f^*_{it} + u_{it} \tag{3}$$

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

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

model

$$y_{it} = a_0 n_{it} + a_1 \left(x'_{it}\beta + b_i + \tau_t + \epsilon_{it} \right) + u_{it}. \tag{4}$$

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.

4.1 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 in by modeling the error them u_{itk} as follows

$$u_{itk} = z'_{itk}\gamma + \hat{u}_{itk} \tag{5}$$

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

$$y_{it} = a_0 n_{it} + a_1 \left(x'_{it}\beta + b_i + \tau_t + \epsilon_{it} \right) + \sum_{k}^{n_{it}} z'_{itk}\gamma + \hat{u}_{it}. \tag{6}$$

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 § 3 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 6 constitutes our preferred specification.

5 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. The particular choice of aggregation granularity is driven by the frequency of reviewing activity. Restaurants in the Boston area receive on average approximately 0.1 1-star published reviews per month, and 0.37 5-star reviews. Table 3 contains detailed summary statistics of the these variables.

We focus on understanding the relationship between a restaurant's reputation and the 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 that we feel comfortable analyzing, including its number of recent positive and negative reviews, and an indicator for whether the restaurant is a chain, or independent business. These are empirically well-founded metrics of a restaurant's reputation.

Specifically, we estimate Equation 6, 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 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 1-star reviews per restaurant-month as the dependent variable. We present these results in Table 5. Next, we discuss our results in detail.

5.1 Results: worsening reputation drives positive review fraud

Low ratings increase incentives for positive review fraud, and high ratings decrease them One measure of a restaurant's reputation is its rating. 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 association between the number of published 1- and 2-star reviews a business received in period $t-1$, and review fraud in the current period. Conversely, we observe a negative, statistically

significant association between review fraud in the current period, and the occurrence of 4- and 5-star published reviews in the previous 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.

Beyond the statistical significance of these results we also interested in their substantive economic impact. One way to gauge this, is to compare 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 $a_1 < 1$ (that is to say Yelp does not identify every single fake review), this number is a lower bound for the increase in positive review fraud.

To assess the robustness of the relationship between recent reputational shocks and review fraud we re-estimated 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 good or bad reviews and engage in review fraud in advance, the effect should be much smaller compared to the effect of recently received reviews. Our results, shown in column 2 of Table 4 suggest that this is indeed the case. The coefficients of the 6-month lead variables are near zero, and not statistically at conventional significance levels. The only exception is the coefficient for the 6-month lead of 5 star reviews ($p < .05$). This is not surprising as to some extent restaurant owners should be able to predict their future ratings based on past performance. 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. First, if a business is looking to manipulate its average rating, then the impact of an additional review is higher when the business has a small number of reviews. Theoretically, the impact of the n^{th} review on the average rating of business is $O(1/n)$. Second, to the extent that the number of reviews that business has signals quality, we would expect that the marginal benefit of an additional review to be lower for

well-reviewed businesses. 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.

Restaurants with fewer reviews are more likely to engage in positive review fraud Our results in Table 4 suggest 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. Furthermore, our results are consistent with the theoretical predictions of Branco and Villas-Boas (2011) who show 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 that 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. 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 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.

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 third 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. (2012)), 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 responsive – albeit to a lesser degree – 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 affect the present incentives of a restaurant's competitors to leave negative fake reviews. Indeed, we find that the coefficients of all 6 leads variables are near zero, and not statistically significant at conventional levels.

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

6 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.

6.1 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 a 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 break down 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. (2012) 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.[9] Therefore, even a low threshold might cast too wide a net. For example, applying a half kilometer cutoff to our dataset results, on average, to approximately 67 competitors per restaurant. Mayzlin et al. (2012) 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,

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

let the impact of restaurant j on restaurant i be:

$$w_{ij} = K\left(\frac{d_{ij}}{h}\right), \tag{7}$$

where d_{ij} is the distance between the two restaurants, K is a kernel function, and h is 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:

$$K_U(u) = \mathbf{1}_{\{|u|\leq 1\}}, \tag{8}$$

where $\mathbf{1}_{\{\dots\}}$ 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.[10]

Similarly, we can define the Gaussian kernel:

$$K_\phi(u) = e^{-\frac{1}{2}u^2}, \tag{9}$$

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 labelled 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

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

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. Let w_{it} be a four-dimensional vector whose first element measures competition by independent restaurants of the same type:

$$w_{it}^{(1)} = \sum_{i \neq j} w_{ij} \mathbf{1}_{\{\text{independent}_j\}} \mathbf{1}_{\{\text{same type}_{ij}\}} \mathbf{1}_{\{\text{open}_{jt}\}}. \tag{10}$$

The indicator functions successively 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.

6.2 Results: competition encourages negative review fraud

We now estimate the impact of competition on review fraud by augmenting our base model of Equation 6 with the vector w_{it}, a set of four time-varying variables measuring competition intensity for restaurant i at time t:

$$y_{it} = a_0 n_{it} + a_1 \left(x'_{it}\beta + w'_{it}\gamma + \sum_{j=1}^{n_{it}} z'_{itj}\gamma + b_i + \tau_t + \epsilon \right), \tag{11}$$

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 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. In addition, since not all fake reviews are successfully filtered by Yelp, this is percentage is a lower bound. Overall, our results suggest that increased competition by similar, nearby restaurants has a statistically significant, 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 Mayzlin et al. (2012) 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 can not reject a methodological explanation since Mayzlin et al. (2012) work in a cross-section setting. Similarly to Mayzlin et al. (2012), 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 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.

7 Conclusion

As crowdsourced information becomes increasingly prevalent, so do incentives for businesses to 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.

Our findings suggest that unethical decision making is a function of incentives, rather than of unethical businesses. Organizations are more likely to game the system when they are facing increased competition and when they have poor, or less established reputations. For managers, policymakers, and even end-users investigating review fraud, our work highlights the situations where reviews are most likely to be fraudulent, as well as the economic incentives that lead organizations to violate ethical, and legal norms.

References

Akoglu, Leman, Rishi Chandy, Christos Faloutsos. 2013. Opinion fraud detection in online reviews by network effects .

Anderson, Eric, Duncan Simester. 2013. Deceptive reviews .

Anderson, Michael, Jeremy Magruder. 2012. Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal* **122**(563) 957–989.

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

Bennett, Victor Manuel, Lamar Pierce, Jason A Snyder, Michael W Toffel. 2013. Customer-driven misconduct: How competition corrupts business practices. *Management Science* .

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

Bolton, Richard J, David J Hand. 2002. Statistical fraud detection: A review. *Statistical Science* 235–249.

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.

Dai, Weijia, Ginger Z Jin, Jungmin Lee, Michael Luca. 2012. Optimal aggregation of consumer ratings: An application to yelp. com. Tech. rep., National Bureau of Economic Research.

Dellarocas, C., G. Gao, R. Narayan. 2010. Are Consumers More Likely to Contribute Online Reviews for Hit or Niche Products? *Journal of Management Information Systems* **27**(2) 127–158.

Dellarocas, Chrysanthos. 2006. Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms. *Management Science* **52**(10) 1577–1593.

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.

Ghose, A., P.G. Ipeirotis, B. Li. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. *Marketing Science* **31**(3) 493–520.

Godes, David, José C Silva. 2012. Sequential and temporal dynamics of online opinion. *Marketing Science* **31**(3) 448–473.

Hausman, Jerry A, Jason Abrevaya, Fiona M Scott-Morton. 1998. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics* **87**(2) 239–269.

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, M. 2011. Reviews, Reputation, and Revenue: the Case of Yelp.com. *Com (September 16, 2011). Harvard Business School NOM Unit Working Paper* (12-016).

Mayzlin, D., Y. Dover, J.A. Chevalier. 2012. Promotional Reviews: An Empirical Investigation of Online Review Manipulation. Tech. rep., National Bureau of Economic Research.

Moe, Wendy W, David A Schweidel. 2012. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science* **31**(3) 372–386.

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.

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.

Zhu, Feng, Xiaoquan Zhang. 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing* **74**(2) 133–148.

Table 1: Characteristics of filtered reviews.

	(1)	(2)
Stars (the reference level is 3 stars)		
1	0.035***	0.036***
	(8.50)	(8.36)
2	−0.022***	−0.021***
	(−10.10)	(−9.63)
4	0.0031*	0.0030*
	(2.24)	(2.13)
5	0.026***	0.026***
	(11.38)	(11.19)
log(Review length)	−0.016***	−0.016***
	(−10.83)	(−10.03)
log(User review count)	−0.096***	−0.097***
	(−84.45)	(−81.13)
User has photo	−0.41***	−0.41***
	(−156.60)	(−152.01)
Stars × Current Yelp Advertiser		
1		−0.0058
		(−0.36)
2		−0.0058
		(−0.62)
4		0.00075
		(0.13)
5		−0.011
		(−1.45)
log(Review length) × Current Yelp Advertiser		−0.011
		(−1.72)
log(User review count) × Current Yelp Advertiser		0.0073
		(1.80)
User has photo × Current Yelp Advertiser		−0.0012
		(−0.11)
N	316415	316107
R^2	0.43	0.43

Note: The dependent variable for all models 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: Filtered reviews and advertising on Yelp.

	(1)	(2)
Stars (the reference level is 3 stars)		
1	0.035***	0.036***
	(8.50)	(8.50)
2	−0.022***	−0.021***
	(−10.09)	(−9.83)
4	0.0031*	0.0034*
	(2.23)	(2.44)
5	0.026***	0.026***
	(11.37)	(11.41)
log(Review length)	−0.016***	−0.016***
	(−10.84)	(−10.48)
log(User review count)	−0.096***	−0.096***
	(−84.46)	(−83.38)
User has photo	−0.41***	−0.41***
	(−156.60)	(−155.06)
Yelp Advertiser	−0.0072	0.020
	(−1.53)	(0.83)
Stars × Yelp Advertiser		
1		−0.019
		(−1.28)
2		−0.0088
		(−0.77)
4		−0.0086
		(−1.07)
5		−0.0089
		(−0.77)
log(Review length) × Yelp Advertiser		−0.0070
		(−0.93)
log(User review count) × Yelp Advertiser		0.0035
		(0.74)
User has photo × Yelp Advertiser		−0.0089
		(−0.73)
N	316415	316415
R^2	0.43	0.43

Note: The dependent variable for all models 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 3: Monthly filtered and published reviews rates by star-rating.

	1-star	2-stars	3-stars	4-stars	5-stars
Filtered reviews/month	0.05	0.03	0.03	0.07	0.10
	(0.27)	(0.18)	(0.16)	(0.29)	(0.42)
Published reviews/month	0.10	0.16	0.28	0.54	0.37
	(0.37)	(0.49)	(0.69)	(1.11)	(1.00)
N	184166	184166	184166	184166	184166

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

	(1)	(2)	(3)	(4)
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: The dependent variable for all models is the number of 5-star filtered reviews per month for each business. 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)
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: The dependent variable for all models is the number of 1-star filtered reviews per month for each business. 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-star fraud		5-star fraud	
	(1) Gaussian	(2) Uniform	(3) Gaussian	(4) Uniform
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: The dependent variable is the number of k-star filtered reviews per month for each business (for $k = 1, 5$). 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-star fraud		5-star fraud	
	(1) Gaussian	(2) Uniform	(3) Gaussian	(4) Uniform
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: The dependent variable is the number of k-star filtered reviews per month for each business (for $k = 1, 5$). Cluster-robust t-statistics (at the individual business level) are shown in parentheses.

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