

Online Appendix to “Incentives to Fake Reviews in
Online Platforms” (not for publication)

Gustavo Saraiva¹

August 31, 2020

¹Saraiva: Assistant professor, Pontificia Universidad Católica de Chile, gqs007@gmail.com.

Contents

A Bayesian Updating	2
B Proof of proposition 2.1	3
C Relaxing Rational Expectations	4
D Jaccard similarity index	6
E Naïve Bayes estimate of text reliability	8
F Regressions that correct for classification error	9
G Detecting anomalous peaks on the volume of 5 star reviews	10
H Alternative database	12
H.1 Fake review detection	13
H.2 Regressions	15
I Solicitation of fake feedback	17

A Bayesian Updating

Assume that $(\varepsilon_t)_{t=1}^{\infty}$ is iid with $\varepsilon_t \sim N(0, \sigma^2)$. Then, if in period 1 the firm was to choose $\eta = \eta_H$ when $q = 1$, and $\eta = \eta_L$ when $q = 0$, we would have from Bayes' rule that consumers' updated beliefs that the firm is of high quality ($q = 1$) after observing v_1 should be given by:

$$\mu_1 = \frac{\mu_0 e^{-\frac{(v_1 - 1 - \eta_H)^2}{2\sigma^2}}}{\mu_0 e^{-\frac{(v_1 - 1 - \eta_H)^2}{2\sigma^2}} + (1 - \mu_0) e^{-\frac{(v_1 - \eta_L)^2}{2\sigma^2}}},$$

In general, denoting $\boldsymbol{\eta} : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}^+$ as consumers' guess regarding the effort level on review manipulation chosen by the seller as a function of its type $q \in \{0, 1\}$ and customers' beliefs $\mu_{t-1} \in [0, 1]$; we have that, provided such guess is correct, then starting at initial beliefs μ_0 , customers' beliefs and signals obey the following Markov process:

$$\mu_t = B(\mu_{t-1}, v_t, \boldsymbol{\eta}) \equiv \frac{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}} + (1 - \mu_{t-1}) e^{-\frac{(v_t - \boldsymbol{\eta}(0, \mu_{t-1}))^2}{2\sigma^2}}},$$

for all $t = 0, 1, 2, \dots$, where

$$v_t = q + \boldsymbol{\eta}(q, \mu_{t-1}) + \varepsilon_t$$

is the signal consumers observe in period t , $(\varepsilon_t)_{t=1}^{\infty}$ is iid with $\varepsilon_t \sim N(0, \sigma^2)$, q is the quality of the firm that is defined initially in period 0, and it is equal to 1 with probability μ_0 , and 0 with probability $1 - \mu_0$, and $B(\cdot)$ is the Bayesian updating function of beliefs.

Notice that we have made the high level assumption that consumers expect the strategy chosen by the seller, $\boldsymbol{\eta} : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}^+$, to only depend on the seller's type and on the previous beliefs μ_{t-1} held by customers. As shown in section 2.2.2, given those beliefs, q and μ_{t-1} will indeed be sufficient statistics for the seller's optimal policy in period t .

B Proof of proposition 2.1

Proof: Let $C(X)$ be the set of real bounded continuous functions with the sup norm defined over $[0, \bar{\eta}]$. If $V(q, \cdot) \in C(X)$, then applying the following transformation T to $V(q, \cdot)$:

$$T(V(q, \mu)) \equiv \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv \quad (1)$$

$$s.t. \quad \mu' = \frac{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}}}{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}} + (1-\mu) e^{-\frac{(v-\eta(0,\mu))^2}{2\sigma^2}}}, \quad (2)$$

we have that $T(V(q, \cdot))$ also belongs to $C(X)$. Indeed, because $\tilde{\eta} \in [0, \bar{\eta}]$, the expression $\lambda\tilde{\eta}^2$ is bounded. In addition, because $q \in \{0, 1\}$, we have that $0 \leq \mu \leq 1$, so that $\omega(\mu') = (1 + \mu)^2/4$ is also bounded. And finally, by assumption, $V(q, \cdot)$ is bounded, which implies that $\delta V(q, \cdot)$ is also bounded. So if we aggregate all these terms to form the function $X(\mu, v, \tilde{\eta}) \equiv \omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')$ defined over $[0, 1] \times \mathbb{R} \times [0, \bar{\eta}]$ (where μ' is obtained by constraint 2), we have that X is bounded. Therefore, there exists $\underline{x}, \bar{x} \in \mathbb{R}$ such that $\underline{x} \leq X(\mu, y, \tilde{\eta}) \leq \bar{x}$ for any $(\mu, v, \eta) \in [0, 1] \times \mathbb{R} \times [0, \bar{\eta}]$. This implies that

$$T(V(q, \mu)) = \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} X(\mu, v, \tilde{\eta}) dv \in [\underline{x}, \bar{x}], \quad \forall \mu \in [0, 1],$$

so that $T(V(q, \cdot))$ is bounded.

The continuity of $T(V(q, \mu))$ follows from the fact that the function $f : [0, 1] \times [0, \bar{\eta}] \rightarrow \mathbb{R}$ such that

$$f(\mu, \tilde{\eta}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\mu^2 - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv,$$

is continuous,¹ and the set of feasible choices for $\tilde{\eta}$, $[0, \bar{\eta}]$, is compact so that, from the maximum theorem,

$$T(V(q, \mu)) = \max_{\tilde{\eta} \in [0, \bar{\eta}]} f(\mu, \tilde{\eta})$$

is continuous with respect to μ .

Now the operator $T : C(X) \rightarrow C(X)$ clearly satisfies the Blackwell sufficient conditions for a β -contraction. Because $C(X)$ is a Banach space, the contraction mapping theorem guarantees that the operator $T(\cdot)$ has a unique fixed point in $C(X)$. ■

C Relaxing Rational Expectations

The results from the previous section were built under the assumption that consumers knew the strategy undertaken by the seller in equilibrium, and thus, would correctly expect some reviews to be fake. But one can also imagine situations in which customers are unaware of the existence of fraudulent reviews, or at least underestimate how prevalent they are. Motivated by that, this section presents the results from the model in a scenario in which most customers incorrectly believe that the effort on review manipulation is zero for both high and low quality sellers (i.e., consumers believe that $\eta(q, \mu) = 0$ for all $q \in \{0, 1\}$ and $\mu \in [0, 1]$), and only a small number of sophisticated customers with zero mass correctly guess the strategy undertaken by the seller.

¹To show that $f(\mu, \tilde{\eta})$ is continuous, define

$$g(\mu, \tilde{\eta}, v) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\mu^2 - \lambda\tilde{\eta}^2 + \delta V(q, \mu')],$$

and let $(\mu_n, \tilde{\eta}_n)_{n=1}^{\infty}$ be a generic sequence defined on $[0, 1] \times [0, \bar{\eta}]$ such that $(\mu_n, \tilde{\eta}_n) \rightarrow (\mu, \tilde{\eta})$. Because $g(\cdot)$ is continuous (since it is the multiplication of continuous functions), the sequence of functions $h_n : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$h_n(v) = g(\mu_n, \tilde{\eta}_n, v), \quad \forall n \in \mathbb{N} \text{ and } \forall v \in \mathbb{R},$$

converges pointwise to $h(\cdot)$ such that

$$h(v) \equiv g(\mu, \tilde{\eta}, v) \quad \forall v \in \mathbb{R}.$$

Moreover, because $|h_n(v)| \leq l(v) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} \max\{\bar{x}, -\underline{x}\}$ for all $n \in \mathbb{N}$ and all $v \in \mathbb{R}$, and because $l(\cdot)$ is integrable, we have from Lebesgue's Dominated Convergence Theorem that

$$\lim_{n \rightarrow \infty} f(\mu_n, \tilde{\eta}_n, v) = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g(\mu_n, \tilde{\eta}_n, v) dv = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} h_n(v) dv = \int_{-\infty}^{\infty} h(v) dv = f(\mu, \tilde{\eta}, v).$$

Figure I displays the equilibrium strategy from both high and low quality sellers in this new environment, together with the equilibrium strategy from the standard version of the model where all consumers have rational expectations. At least for the set of parameters that we tested, we find that when consumers are unaware of the existence of fake reviews, high quality sellers tend to engage in less review fraud, while low quality sellers end up faking more reviews.

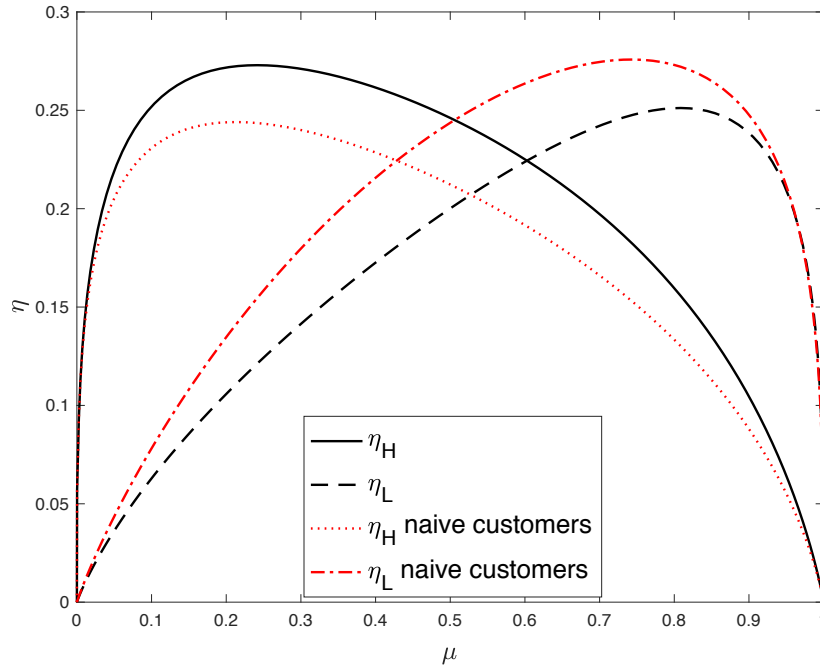


Figure I: Equilibrium as a function of μ , when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$, when consumers know the strategy taken by the seller, and when consumers are naive and believe the seller does not engage in review manipulation (i.e., they believe $\eta(q, \mu) = 0$ for all q, μ).

As this diminishes the gap between the signals from high and low quality sellers, sophisticated consumers take longer to learn the true type from the seller, as displayed in the red lines from figure II. As to naive customers, the green lines from the figure show that their expectation regarding the quality from both high and low quality sellers increase, as they do not suspect ratings to have been inflated by the seller.

A similar pattern emerges by allowing the fraction of sophisticated consumers to have a positive mass, as depicted in figure III. For all combination of parameters that we tried we obtained the same result: the greater the mass of consumers that are unaware of the existence of fake reviews, the lower is the effort of review manipulation from high quality sellers, and the higher is the effort of review manipulation from low quality sellers. So in

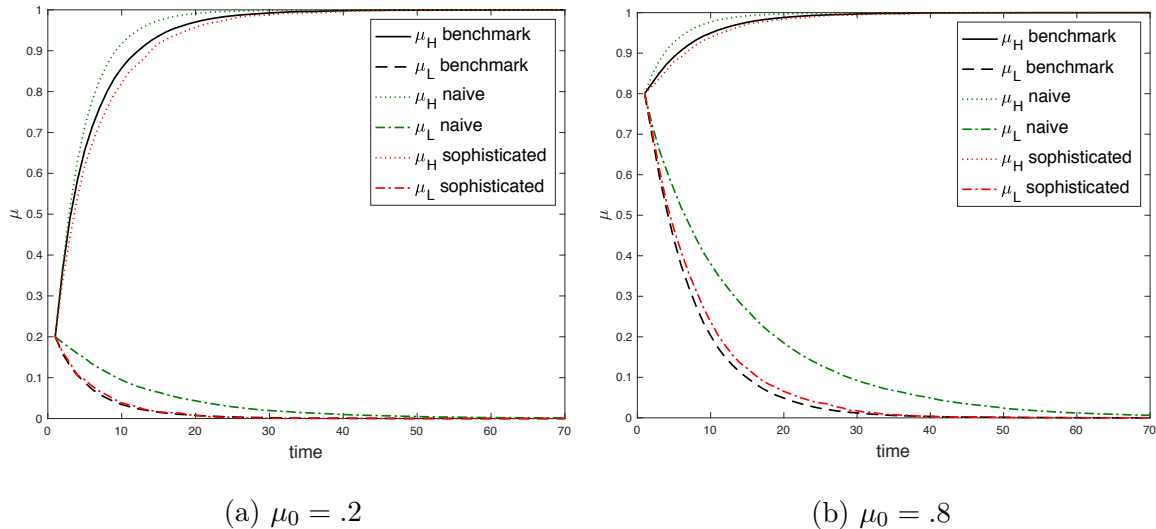


Figure II: The average simulated evolution of reputation when the seller is faced with naive customers and a small fraction of sophisticated customers, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. μ_H corresponds to the average reputation from high quality sellers, whereas μ_L corresponds to the average reputation from low quality sellers.

principle, educating naive customers about the existence of fake reviews in the platform could increase the gap of signals generated by high vs low quality sellers, thus increasing the speed with which sophisticated customers learn the true quality of the seller.

D Jaccard similarity index

To compute the Jaccard similarity index, we first generate all sequences of 4 words from each review. We call those sequences as “shingles”. As an example, consider the following hypothetical review:

“These wireless earphones are the best!”

The shingles from the above sentence are:

1. “These wireless earphones are”
2. “wireless earphones are the”
3. “**earphones are the best**”

Now doing the same process with the following sentence:

“Those earphones are the best I ever had!”,

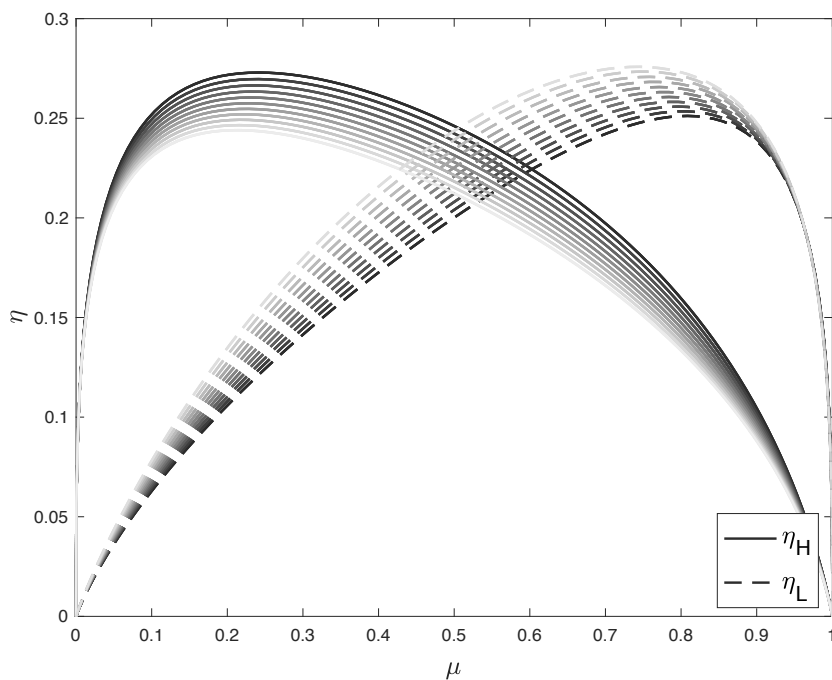


Figure III: Optimal strategy from high and low types, when a fraction $\alpha \in [0, 1]$ of customers uniformly distributed between $[0, 1]$ are naive and the remaining customers are sophisticated. Lighter gray lines correspond to higher values of α .

we get the shingles

I “Those earphones are the”

II “**earphones are the best**”

III “are the best I”

IV “best I ever had”

The Jaccard similarity between those two reviews is given by the number of shingles that intersect divided by the added number of shingles from each review. So in the current example, one can see that shingles 3 and II are the only ones that match. So the Jaccard similarity between those reviews is given by $2/(3 + 4) = 0.29$.

While computing the Jaccard similarity index is computationally feasible for a pair of small reviews, doing so for thousands of potentially large reviews is computationally infeasible.² In those cases, hashing algorithms can be used to consistently estimate the

²For one of my samples, I would need to make $4.17e+10$ of those computations.

actual Jaccard index in a computationally feasible way. For the purposes of this research I used the MinHash algorithm.³

E Naïve Bayes estimate of text reliability

As mentioned earlier, text similarity was build by using a Naïve Bayes classifier. At a high level, the process consists on computing the frequency from each word that appears among fake and real reviews, and then using those frequencies to estimate the probability that a certain sequence of words was generated from a legitimate or a fraudulent review. The process can be employed using content from both review text and review title.

More precisely, let $text = (w_1, w_2, \dots, w_n)$ represent a generic sequence of words used to review a product. Then it follows from Bayes' rule that:

$$P(\text{fake}|\text{text}) = \frac{P(\text{text}|\text{fake})P(\text{fake})}{P(\text{text})},$$

and

$$P(\text{real}|\text{text}) = \frac{P(\text{text}|\text{real})P(\text{real})}{P(\text{text})},$$

where the notation is self explanatory.

So conditional on its content, a review is more likely to be fake iff

$$\begin{aligned} P(\text{fake}|\text{text}) &> P(\text{real}|\text{text}) \\ \iff P(\text{text}|\text{fake})P(\text{fake}) &> P(\text{text}|\text{real})P(\text{real}) \\ \iff \log(P(\text{text}|\text{fake})) + \log(P(\text{fake})) &> \log(P(\text{text}|\text{real})) + \log(P(\text{real})). \end{aligned} \quad (3)$$

Getting an unbiased and consistent estimate of $P(\text{fake})$ is relatively easy: one only needs to compute the fraction of reviews in the sample that are fake (though in practice one actually uses the fraction of reviews in the sample that are *classified* as fake, as it is virtually impossible to perfectly distinguish fake reviews from real ones). But creating an unbiased and consistent estimator of $P(\text{text}|\text{fake})$ and $P(\text{text}|\text{real})$ requires imposing stronger restrictions in the data generating process (DGP) governing review texts.

The Naïve Bayes classifier approach simplifies the DGP from review texts by assuming that words are generated randomly and independently. Though this assumption is clearly not very realistic, as words need to be put in a logical order in order to convey meaning, it

³Details of the algorithm can be found at ?.

greatly simplifies the process of finding a reliable estimate of $P(\text{text}|\text{fake})$. Indeed, letting $\text{text} = (w_1, w_2, \dots, w_n)$ denote the sequence of words from a review, this assumption implies that

$$P(\text{text}|\text{fake}) = \prod_{i=1}^n P(w_i|\text{fake}).$$

Because the probabilities $P(w_i|\text{fake})$ can be consistently estimated by computing the proportion of times each word w_i appears on the set of words used to write fake reviews, one can consistently estimate $P(\text{text}|\text{fake})$ by multiplying those estimated probabilities.⁴ The same approach can be applied to estimate $P(\text{text}|\text{real})$.

So the aforementioned procedure was used to estimate the left and righthand side of inequality 3. If the estimated $P(\text{real}|\text{text})$ was greater than $P(\text{fake}|\text{text})$, then the dummy variable “*Reliability index from review text*” would assume value 1, else it would assume value 0.

F Regressions that correct for classification error

The logit model estimated in section 3.5 implicitly assumed that the variable $y_{i,s,t}$ used to classify reviews as fake or real was flawless, i.e., that there were no instances in which some fake reviews were incorrectly classified as real, and vice versa. But in practice the researcher can not determine with absolute certainty whether a review is fake or not, so that one should expect a certain degree of misclassification to be present in the dataset. In our case, even though reviews were only classified as fake when very strong evidence supported that those reviews were in fact fake (see section 3.2), it is very likely that some of the fake reviews from our sample were incorrectly classified as real. So in essence our variable of interest $y_{i,s,t}$ is not observable. What is observable instead is $y_{i,s,t}^o$, an indicator variable that equals 1 if the researcher classified review i from product s posted at time t as fake, and zero otherwise, where occasionally we may have $y_{i,s,t}^o \neq y_{i,s}$.

Because the presence of misclassifications of the dependent binary variable causes the Probit and Logit estimates to be biased and inconsistent, I use an estimation approach proposed by ? that corrects for endogenous misclassifications. Formally, let $\mathbf{z}_{i,t,s}$ be a vector of covariates that can predict whether or not a review is fake, such as whether

⁴As a standard approach, *stop words*, such as “I”, “there”, “but”, etc., were removed from the reviews before conducting the Naïve Bayes estimation.

the review had a verified purchase tag, or whether it contained a picture or a video, etc. Then we assume that the probability that a review is classified as fake when the review is in indeed fake conditional on the vector of covariates $\mathbf{z}_{i,s,t}$ is given by:

$$Prob(y_{i,s,t}^o = 1 | y_{i,s,t} = 1, \mathbf{z}_{i,s,t}) = F_o(\mathbf{z}_{i,s,t}\gamma),$$

where $F_o(\cdot)$ is a cdf. Because reviews from our sample were classified as fake only when very strong evidence supported that they were so, I assume that a real review from our sample is never incorrectly classified as fake, i.e.,

$$Prob(y_{i,s,t}^o = 1 | y_{i,s,t} = 0, \mathbf{z}_{i,s,t}) = 0.$$

So letting $\mathbf{x}_{i,s,t}$ denote the vector of explanatory variables of interest, namely, the time it took for the review to be posted and the product's current reputation level, and letting $\mathbf{z}_{i,t,s}$ be the vector of covariates used to control for classification error, we have that the conditional probability of observing $y_{i,s,t}^o = 1$ is given by

$$\begin{aligned} Prob(y_{i,s,t}^o = 1 | \mathbf{x}_{i,s,t}, \mathbf{z}_{i,s,t}) &= Prob(y_{i,s,t} = 1 | \mathbf{x}_{i,s,t}) Prob(y_{i,s,t}^o = 1 | y_{i,s,t} = 1, \mathbf{z}_{i,s,t}) \\ &= F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma) \end{aligned}$$

With these probabilities, we can then build the loglikelihood function

$$l(\beta, \gamma) = \sum_{i,s,t} [y_{i,s,t}^o \log(F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma)) + (1 - y_{i,s,t}^o) \log(1 - F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma))],$$

and maximize it to obtain estimates of β and γ .

The results from this regression are depicted in table I. Again, the results from the regression are very similar to the ones obtained earlier in section 3.5 and depicted in table 3. Looking at the variables of interest, they exhibit the same patterns as the ones derived earlier: older reviews are more likely to be fake, and the probability of a review being fake is smaller for very low or very high levels of reputation $\mu_{i,s,t}$.

G Detecting anomalous peaks on the volume of 5 star reviews

Detecting spikes on the number of 5 star reviews received by a seller was done using an STL (seasonal trend decomposition) approach. The process consists on first estimating

	variable	estimates		
x	constant	-2.58 (1.674)	-11.34*** (1.0445)	1.556*** (0.135)
	$\tilde{\mu}$	27.42*** (6.02)	48.7*** (4.1122)	
	$\tilde{\mu}^2$	-37.86*** (5.86)	-54.85*** (4.11)	
	time	-0.012*** (0.0014)	-0.0136*** (0.00078)	-0.0154*** (0.00098)
	constant	2.345*** (0.2995)	7.969*** (7.801)	2.571*** (0.2833)
z	Dummy for text reliability	-3.942*** (0.26)		-4.147*** (0.2456)
	Numb. helpful feedback	0.03*** (0.00495)		0.0334*** (0.004996)
	Verified Purchase	-0.58*** (0.0777)		-0.85*** (0.0785)
	has images or videos	0.573*** (0.0994)		0.6557*** (0.0985)
	Observations:	18,440	18,440	18,440
pseudo R^2 :	0.38203	0.1277	0.3614	

Table I: Logit regression after correcting for endogenous classification errors.

the expected number of positive reviews that a seller should receive at a particular day as a function of trend, seasonal effects and covariates. If the estimated prediction was sufficiently distant from the realization of positive reviews on that period, a dummy would classify all the 5 star reviews that the seller received on that day as anomalous.

More precisely, reviews were aggregated on a daily level to create a panel data. Let $X_{i,t,p,s}$ be the number of 5 stars that a product p from seller s received at date t , during its i 'th period since it entered the market (notice that t is the actual date it received a review, whereas i corresponds to the number of days since that product got its first review). $X_{i,s,p,t}$ was regressed against its lagged components, trend, seasonal dummies,

and seller fixed effect, likewise:

$$X_{i,t,p,s} = \beta_0 X_{i-1,t-1,p,s} + \beta_1 t + \beta_2 t^2 + \sum_{j=1}^{12} \gamma_j D_{j,t} + \varepsilon_{i,t,p,s},$$

where $\{D_{j,t}\}_{j=1}^{12}$ are the dummies for the corresponding month,⁵ and $\varepsilon_{i,t,p,s}$ is an iid random term.

After estimating the model using OLS, it was determined that if a residual term was 3 standard deviations above or below the average residual, then that day for the corresponding seller would be flagged as anomalous, in which case all the 4 and 5 star reviews that the seller received on such days would be flagged as fake.⁶

H Alternative database

As mentioned at the beginning of section 3, the database collected from sellers who were either caught soliciting fake reviews or were flagged by users for their involvement in suspicious activity may suffer from selection bias. Indeed, by focusing the analysis on those sellers, the resulting sample may end up with an overrepresentation of fake reviews, which could then affect the resulting estimates from the regressions. Moreover, restricting the analysis to those sellers may limit the overall sample size, as manually finding suspicious sellers is a tedious and time consuming process. And finally, the resulting dataset is highly heterogenous, as it includes several different types of products, ranging from cheap electronic devices to children’s toys, which can lead our model to be misspecified.

To address these issues, I collected a separate dataset comprised exclusively of wireless headsets sold at Amazon, not targeting any seller in particular from such category throughout the sampling process. The reason I chose wireless headsets is because one can find evidence in the news that fake reviews on those products are prolific on Amazon, thus making the analysis for this market niche economically relevant.⁷ The dataset was then used to estimate regressions similar to the ones presented in section 3.5.

The dataset is comprised of 278,829 reviews from 1,134 different headphone products. So sellers on average received approximately 246 reviews, which is significantly higher

⁵Similar results were obtained by replacing those dummies by weekly dummies.

⁶4 star reviews were also flagged as fake during anomalous periods to account for the fact that sellers may try to avoid detection by adding some 4 star reviews into their mix of fraudulent reviews.

⁷See for instance *How merchants use Facebook to flood Amazon with fake reviews (The Washington Post, April 23, 2018)*.

than the 109 average number of reviews from the previous sample. But this can be partially explained by the fact that this sample contains a higher volume of positive unverified purchase reviews, as depicted in table II below:

	All Reviews	Verified Purchase	Unverified Purchase
Number of Reviews	278,829	178,187	100,642
% 5 star reviews	73%	61%	93%

Table II: Descriptive statistics of the wireless earphone sample.

From the table, one can also notice the striking difference between the percentage of 5 stars among verified vs non verified purchases. The high volume of five stars unverified purchases is an early indication that review manipulation in this sample is even more pronounced than in the previous one.

Regarding the distribution of stars, they are similar for both samples as depicted in figure IV. This J shaped configuration of reviews, with most reviews being either highly positive or highly negative, with the great majority being highly positive, is actually quite common in different platforms (?). The wireless headphone dataset has, however, disproportionally more positive reviews, which is mainly attributed to its high volume of 5 star unverified purchase reviews.

H.1 Fake review detection

Because we do not have prior evidence that a generic seller from this new dataset solicited fake reviews through other platforms such as facebook or Rapidworkers, I no longer employ criterion number II presented in section 3.2 to detect fake reviews. Instead, I rely on the following criteria:

- I) If two different reviews were sufficiently similar to one another in terms of their text Jaccard similarity index, and the reviews in question had more than 10 words, and they were both from products in which fake review solicitation happened in the same online platform, then those reviews were classified as fake.
- II**) If a positive reviewer was posted during a day with an anomalously high volume of positive reviews, it was classified as fake. Details are presented in section G from this appendix.

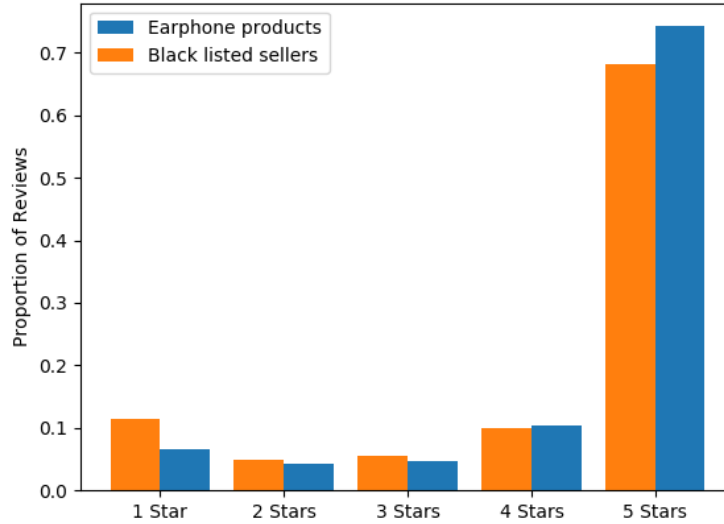


Figure IV: Histogram of the number of stars per sample. The bars in blue correspond to the sample of wireless earphone products, whereas the one in orange corresponds to the sample described in section 3.1 generated by targeting suspicious products that were either soliciting reviews in online platforms, or were flagged as suspicious on Amazon forums.

While the first criterion was already employed in the previous sample, criterion II**) is new and exploits the fact that this sample exhibits several anomalous busts in the volume of positive reviews received by certain products, as depicted in figure V. The figure displays the number of 5 star reviews received by 2 different products throughout time. From the figure, it is evident that most of the positive reviews from these products were concentrated around a few days. Moreover, more than 99% of those reviews were 5-star unverified purchase reviews, thus adding evidence that they were most likely fake. So adding detection criterion II**) enables the identification of a large volume of fake reviews that would not have been detected otherwise by criterion I) alone.

Table III below describes the proportion of reviews in the sample that were classified as fake following those two criteria. Comparing it with table 2 from the main article, we observe that this sample has an even higher proportion of fake reviews, especially among unverified purchase reviews, representing a striking 89% of those.

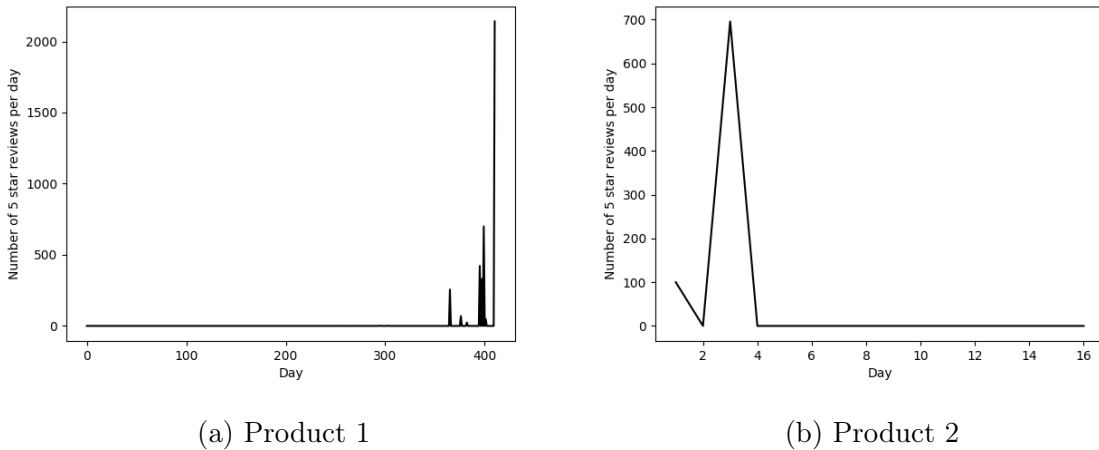


Figure V: Number of 5 star reviews received by a couple of products per day. Product 1 is no longer sold at Amazon, perhaps because Amazon detected suspicious activity surrounding its reviews and thus had the product removed. Regarding product 2, as I write this on May 16, 2019, though it is still sold on Amazon, all its positive reviews (4 and 5 stars) have been removed.

	All Reviews	Verified Purchase	Unverified Purchase
% fake among 4 and 5 star reviews	43%	10%	89%
% fake among 4 star reviews	9.4%	8.9%	15%
% fake among 5 star reviews	48%	11%	90%

Table III: Proportion of reviews classified as fake per category for the wireless earphone sample.

H.2 Regressions

Table IV reports the results from logit regressions. The results are mostly similar to the ones obtained with the previous sample, except for the coefficients of reputation, which are occasionally statistically insignificant due to multicollinearity between this variable and the verified purchase dummy.

But regarding the time variable, all regressions consistently indicate that fake reviews are more pronounced at the initial stages following a seller’s entrance (or potentially reentrance) into the market.

Table IV: Simple Logit regressions using the earphone dataset.

	<i>Dependent variable:</i>			
	$y = \mathbb{1}(\text{review is fake})$			
	(1)	(2)	(3)	(4)
Constant	3.4366*** (0.051)	-2.411*** (4.315e-02)	-1.428*** (4.38e-02)	3.469*** (1.795e-02)
μ	-0.226 (0.171)	9.955*** (1.526e-01)	10.13*** (1.545e-01)	
μ^2	0.541*** (0.164)	-9.724*** (1.413e-01)	-9.947*** (1.45e-01)	
time	-0.0032*** (0.0001)	-5.759e-03*** (8.578e-05)	-5.581e-03*** (8.81e-05)	-2.999e-03*** (9.556e-05)
Dummy for text reliability	-1.695*** (0.016)		-1.317*** (1.05e-02)	-1.694*** (1.610e-02)
Numb. helpful feedback	0.0016*** (0.0003)		-4.543e-03*** (5.43e-04)	1.594e-03*** (2.958e-04)
Verified Purchase	-4.306*** (0.0155)			-4.294*** (1.508e-02)
Has images or videos	-0.291*** (0.039)		-1.703*** (3.27e-02)	-2.914e-01*** (3.857e-02)
Observations	201,393	201,393	232,176	232,176
Log Likelihood	-72,584.41	-148,141.1	-137,984.5	-72,609.94
Pseudo R^2	0.541782	0.06480051	0.1289179	0.5416209
<i>Note:</i>			*p<0.1; **p<0.05; ***p<0.01	

I Solicitation of fake feedback

Amazon HELPFUL VOTES Needed


Work done: 3/50
 You will earn: \$0.11
 This task takes less than 5 minutes to finish
 Campaign ID : 59e38c46-b098-4511-85f7-4eda3257911a
 Campaign Name : Amazon HELPFUL VOTES Needed

You can accept this job if you are from THESE COUNTRIES ONLY:

International

Campaign isn't working?

Campaign isn't working? If a Campaign does not work, please report that immediately. Include Campaign name and Campaign ID [Click to Report](#)

 **What is expected from workers?**

THIS IS AN AMAZON.COM HELPFUL VOTE. YOU CAN DO THIS USING ANY AMAZON ACCOUNT THAT CAN POST REVIEWS.

YOU MUST VOTE YES

Use the link below

Please VOTE YES ON ALL REVIEWS LISTED BELOW

https://www.amazon.com/gp/customer-reviews/R3SLQR3BGP3JTZ/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B075LP9BSY https://www.amazon.com/gp/customer-reviews/R12I5TIT2QHFAG/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B075JRHLK
https://www.amazon.com/review/R1U7S7YH6LQ5FL/ref=pe_1098610_137716200_cm_rv_eml_rv0_rv
https://www.amazon.com/review/R1A0I17I65VF7N/ref=cm_cr_srp_d_rdn_nerm?ie=UTF8

Figure VI: An example of a seller soliciting positive feedback to reviews praising its products.