ABSTRACT

With the proliferation of online rating platforms, there has been an increasing concern over the authenticity of reviews posted online. While much effort has been dedicated to improving fake review detection algorithms, little attention has been spent on understanding the incentives that drives some sellers to solicit fake reviews. To fill this gap, this paper develops a theoretical model in which sellers dynamically choose their effort spent on review manipulation. Among other things, the model predicts that sellers' optimal investment in fake reviews is not a monotone function of their reputation. More precisely, sellers that currently possess a very good or very bad history of past reviews have less incentives to solicit fake reviews praising their own products, the intuition being that, for sellers with very bad reputation, it is too costly to pretend that they are high quality sellers; while sellers that have already accumulated a very good reputation do not need to spend much effort in convincing buyers that they are high quality sellers. Another prediction from the model is that, in order to maximize the impact from each fake review, sellers tend to concentrate review manipulation at the initial stages after they have entered the market. Using data collected from Amazon, I was able to observe those two features from the model at the empirical level by estimating a Logit regression that predicts the probability of a review being fake as a function of the product’s reputation and the time it took for the review to be posted since the seller entered the market.

Key words: fake reviews, deception, reputation
1 Introduction

With the proliferation of online rating platforms such as Yelp and TripAdvisor, there has been an increasing concern over the authenticity of reviews posted online. In the news one can find several pieces of anecdotal evidence that fake reviews are prolific and have been increasing over the last years (see for instance, Amazon’s Fake Review Problem Is Now Worse Than Ever, Study Suggests (Forbes, Sep 9, 2017), Facebook fake review factories uncovered by Which? investigation, (The Guardian, Oct., 2018) and ‘The Shed at Dulwich’ was London’s top-rated restaurant. Just one problem: It didn’t exist (Washington Post, Dec 8, 2017)).

Those practices can add distortions to the market through a series of channels, a few of which include: 1) they can lead customers to make suboptimal decisions; 2) they can also lead customers to perceive reviews as a poor measurement of the product’s quality, and it is a well established result that, when buyers do not know the quality of the products sold, adverse selection may occur (Akerlof [1970]); 3) buying fake reviews can be unfair to sellers who refuse to engage in this practice for ethical or legal concerns, etc. These problems have motivated the dissemination of a literature in Computer Science dedicated to detecting and eliminating fake reviews in online platforms (see section 2 for details).

In contrast to computer scientists, economists are usually more interested in understanding the causal effects that lead some sellers to fake reviews more than others. One motivation for understanding these causal effects is that such knowledge can be used to perfect existing algorithms used in the detection process. Indeed, by understanding what causes sellers to fake reviews, one can trim down the set of predictors to be used in the detection algorithm to include only the most pertinent variables.

This paper fits into the latter branch of the literature, as it aims to shed light into the incentives that drive some sellers to fake reviews. For that purpose the paper develops

\textsuperscript{2}Notice that in some markets a suboptimal decision made by customers can be quite costly, such as in the market for private doctors, where the resulting assignment can impact patients’ health. And in those markets the surge of online platforms as the likes of vitals.com, ratemds.com, Yelp, etc., which display reviews from physicians, has been increasingly playing a bigger role in customers’ decisions. Shukla, Gao and Agarwal [2018], for instance, find that the introduction of reviews in a doctor appointment platform in India has increased the number of online appointments for highly rated doctors by roughly 29.6%, while decreasing the number of appointments from unrated doctors.
a novel theoretical framework in which sellers dynamically choose their optimal amount of effort devoted to faking reviews in online rating platforms such as Yelp, Amazon and TripAdvisor, given that faking reviews is costly to sellers and that customers correctly anticipate some reviews may be fake. The model predicts that, in order to maximize the impact from each fake review, high quality sellers have incentives to concentrate review fraud at the initial stages after they have entered the market. As their reputation improves over time, such sellers gradually fake less reviews. For low quality sellers, however, maintaining a good reputation is unsustainable in the long run, as they systematically receive reviews disparaging their products from honest consumers. Once their reputation has been squandered, such sellers exit the market and reenter with a new brandname, always concentrating their efforts in review manipulation right after reentering the market, so as to maximize the impact from each fake review. So in markets wherein changing one’s name is relatively costless, low quality sellers should be expected to disproportionately fake more reviews than high quality sellers in the long run. Verifying this last prediction empirically may be challenging, since the sellers that change their brandnames usually do so in a concealed fashion that prevents the researcher from accessing their previous identities.

Another qualitative prediction from the model is that the effort spent on review fraud is not a monotone function of the seller’s reputation. More precisely, very low or very high reputation levels are usually associated with a low effort on review manipulation, the intuition being that, for sellers with very low reputation, it is too costly to pretend that they are high quality types, which gives them little incentives to fake reviews; similarly, for sellers with very good reputation, the marginal benefit from faking reviews is relatively small since everyone already believes the seller to be of high quality with high probability.

To test some of the predictions from the model, I scraped reviews from different products sold by Amazon that I flagged as suspicious based on the fact that their sellers were (apparently) soliciting fake reviews online (namely, through Facebook and Rapidworkers). I then classified the reviews collected as fake and real based on a series of criteria used in the computer science literature dedicated to detecting fake reviews through supervised learning algorithms. After that I estimated a Logit model to predict the probability of a review being fake conditional on the seller’s reputation, and on the time it took for the review to be posted since the seller entered the market. Consistent with the predictions
from the theoretical model, the results from the regressions suggest that the probability of a review being fake is lower for sellers with very high or very low reputation. Moreover, the probability of a review being fake diminishes with time, which is consistent with the prediction that sellers should focus review manipulation at the initial stages following their entrance (or reentrance with a new brandname) into the market.

2 Related Literature

When it comes to related literature done on the theoretical level, one can cite the work from Mayzlin [2006] and Dellarocas [2006]. These papers have different premises and predictions as to the types of sellers that have most incentives to fake reviews.

Starting with Mayzlin’s paper, it assumes that consumers randomly observe a single opinion and then, based on that single observation, update their beliefs regarding the quality of sellers, knowing that the opinion that they picked may potentially be fake. Her model leads to the prediction that low quality sellers fake more reviews as compared to high quality sellers. However this conclusion relies on the assumption that each buyer only observes a single review extracted from the overall pool of reviews, when in reality most online rating platforms (e.g., Yelp, TripAdvisor, Amazon, etc.) provide summary statistics of all previous reviews. Moreover the model implicitly assumes that consumers know exactly how many legitimate reviews have been posted online, even though they do not know the total number of reviews posted (i.e., the number of fake plus real reviews). In addition, her model imposes a series of technical restrictions on exogenous parameters, such as sellers’ initial reputation, which prevents one from performing some comparative statics analysis, such as how reputation affects the effort on review manipulation. Finally, her model transpires in a single time window, which prevents one from accessing, for example, whether sellers have incentives to concentrate review fraud at the initial stages after they have entered into the market, or smooth review manipulation throughout the periods.

Dellarocas [2006], on the other hand, uses a different specification that abstracts from some of the technical complications associated with modeling fake review optimization. In particular, it assumes that legitimate reviews generate a stochastic signal (common to all customers) that is correlated with the product’s true quality. Because the signal
is stochastic, if sellers invest in the distortion of this signal through fake reviews, buyers
can not perfectly separate which part of the signal was generated by honest reviews, and
which part was explained by fake reviews added to the system.

Different than Mayzlin [2006], Dellarocas [2006] predicts that in some instances high
quality sellers may actually fake more reviews than low quality sellers. But those results
suffer from a few limitations, including: 1) The paper has a few mistakes when it comes
to the derivation of the seller’s profit function. 2) It uses an intractable data generating
process for product quality. Indeed, it assumes that the quality parameter is drawn from a
normal distribution, which implies that the seller’s strategy is a continuous function that
maps its type into effort on review manipulation. This implies that the researcher has
to rely on guess and verify methods to discover the functional format of the equilibrium
strategy. In Dellarocas [2006], the equilibrium they analyze is one in which the seller’s
optimal investment on fake reviews is an affine function of its quality, which would then
imply that a seller with a sufficiently low quality (or sufficiently high quality, depending on
whether this affine function is increasing or decreasing) would choose to receive money to
have its products disparaged (i.e., they would choose a negative amount of fake reviews),
which is a strategy that might be hard to be implementable in practice. Not to mention
that the model could potentially have many other equilibria. 3) Finally, like Mayzlin
[2006], his model transpires in a single time window, thus preventing researches from
accessing whether sellers have incentives to concentrate review fraud at the initial stages
following their entrance into the market.

My theoretical framework is very similar to Dellarocas’ specification. The main dif-
fferences and innovations from my model are that: 1) it corrects some mistakes regarding
the derivation of the profit function from the seller; 2) it assumes that there are only two
types of product quality, high and low, as opposed to a continuum set of types, which
makes the model more tractable, and therefore allows one to compute relevant comparative
statics; 3) it allows the seller to be forward looking and set its effort on review
manipulation dynamically, thus allowing the researcher to derive conclusions regarding
dynamic aspects of review manipulation.

A common feature share by these two papers which is also present in my own model is
that they both have the desirable property that consumers correctly anticipate that some
reviews are fake. So when looking at signals generated by reviews, consumers curb their
expectations by taking into account that some reviews are not perfectly reliable, which is consistent with anecdotal evidence that consumers are aware of the existence of fake reviews.

As to empirical papers that investigate variables that affect review manipulation, one can cite [Luca and Zervas 2016] and [Mayzlin, Dover and Chevalier 2014]. Among other things, these two papers find, using different databases, that chain restaurants (in [Luca and Zervas 2016]) and chain hotels (in [Mayzlin, Dover and Chevalier 2014]) are less likely to fake positive reviews praising their products, since they offer a standard service that already has a solidified reputation.

[Luca and Zervas 2016] also run regressions that seem to support their conjecture that sellers with lower reputation have more incentives to fake reviews, and as their reputation improve, they gradually fake less reviews (hence, the creative title from their paper: “fake it till you make it”). To test this conjecture, they use positive reviews (4 or 5 stars) filtered by Yelp as a proxy to measure the effort of review manipulation spent by restaurant owners, then regress the number of filtered reviews per time interval as a function of the total number of 1, 2, 3, 4 and 5 stars accumulated by the restaurant in previous periods. A limitation from this specification is that, by grouping the number of filtered reviews into time intervals the researchers lose pertinent information regarding each individual review that could be used to correct for classification errors (i.e., to correct for real reviews that were wrongly filtered by Yelp, as well as fake reviews that Yelp failed to filter). Moreover, the number of reviews filtered by Yelp may not be a very good measure as to the effort spent on review manipulation, since, according to Yelp’s website, they not only filter reviews that have high chances of being fake, but also reviews that are likely to be less relevant to consumers (say, because the reviews lack useful content, or they are too old, etc.). Finally, even though the authors find that having more previous 4 and 5 stars are usually associated with less fake reviews in the current period, they haven’t actually created a univariate measure of reputation to test their conjecture. One contribution from my paper is that it uses a different database that targets the detection of fake reviews exclusively. Moreover I use a logit specification that corrects for endogenous classification errors by using data at the individual level. And finally, in order to measure the impact that reputation has on the incentives to fake reviews, I create a univariate measure of sellers’ reputation, as opposed to using a vector of the number of previous 1,
2, 3, 4 and 5 stars received.

There is another branch of the literature that focuses in analyzing the impact that reviews have on sales. To cite a few papers, Chevalier and Mayzlin [2006], Luca [2016] and Shukla, Gao and Agarwal [2018] find that positive reviews have a positive and significant impact on sales (Chevalier and Mayzlin [2006] use data from book reviews at Amazon and bn.com, while Luca [2016] uses data from restaurant reviews on Yelp, and Shukla, Gao and Agarwal [2018] use review data from a doctor appointment platform in India). My paper, on the other hand, takes those results as given, and focuses instead in identifying the types of sellers that have most incentives to fake reviews.

This paper uses a combination of methodologies employed in the computer science literature to create training databases for the purpose of fake review detection. Essentially, to detect fake reviews through supervised machine learning techniques, researchers need to feed the machine with some examples of reviews that they know to be fake, and another subsample of reviews that they know to be real, so that the machine can learn to distinguish the patterns from each group. The challenge is that in practice researchers can not tell for sure whether a review is fake or not, after all, fake reviews are supposed to be convincing. As a matter of fact, some experimental evidence suggest that humans are in general poor judges when it comes to detecting fake reviews (see for instance Ott et al. [n.d.]). So in order to build a training sample, researches usually spot reviews that are “clearly fake” based on some some baseline criteria that rely on computational methods. That is, while it is virtually impossible to determine from the naked eye whether a review is fake or not, by using automated methods that process big amounts of data, one can find reviews that are almost certainly fake.

Kaghazgaran, Caverlee and Alfifi [n.d.], for instance, looked at Amazon products that were soliciting fake reviews on the crowdsourcing platform RapidWorkers, and then classified a review as fake if the reviewer in question had posted reviews to two or more different products from the list. The premise behind this criterion lies in the fact that fake reviews are usually mass produced, and that the probability that a customer happens to review two or more products that were crowdsourcing fake reviews on the same platform by pure chance is very small, given that Amazon sells millions of different products. Jindal and Liu [n.d.], on the other hand, classify a review as fake if its review text is a near duplicate to some other review from the sample. The intuition behind this criterion lies in the fact
that, since fake reviews are usually mass produced, fake reviewers have a tendency to copy and paste the same text to describe different products. In my paper I combine these two criteria to classify reviews as fake and real, while also adding a new criterion, which, to the best of my knowledge, has not been exploited in previous literature (though the method draws some resemblance to the one used by Mukherjee et al. (2012) [Mukherjee, Liu and Glance n.d.] which regards groups of reviewers that provide feedback to the same products as suspicious).

3 Theoretical Model

The outline of the theoretical model can be described as follows: nature initially determines the type of the seller as being high or low quality with an exogenous probability \( \mu_0 \). After learning its type, the seller chooses how much to invest on review manipulation. Then a random signal \( v_1 \) is generated that is observable to consumers. The signal \( v_1 \) is positively correlated with the firm’s quality and its investment on review manipulation. After observing the signal, potential buyers compute \( \mu_1 \), their updated beliefs regarding the probability that the seller is high type. After consumers update their beliefs, the seller chooses the optimum price \( p_1 \) from its product and then the heterogenous consumers decide whether or not to purchase the product. After the firm’s profits are realized for that period, the firm goes to the next period with a new reputation \( \mu_1 \) and is matched with a new set of customers, wherein the same process is repeated iteratively.

Figure 1 depicts the outline of the model, where \( \eta(q, \mu) \) corresponds to the effort of review manipulation chosen by the firm as a function of its type \( q \), and its current reputation \( \mu \).

Later on, section 3.6 adds the possibility that at any point a seller can pay a fixed cost to exit and reenter the market with a new name. Sellers will of course only choose to do so once their reputation (i.e., their \( \mu \)’s) have reached a sufficiently low point. Unsurprisingly, our simulations predict that the reputation of high quality sellers tend to improve over time, so that they usually find no need to resort to this tactic. Low quality sellers, on the other hand, tend to constantly exit and reenter the market with a new name, since their reputation tends to deteriorate over time as a result of honest reviews.
3.1 Firm’s profits as a function of its expected quality

A monopolist wishes to sell a product with quality $q$ that is unknown to potential customers. Time is discrete and finite, and indexed by $t \in \{0, 1, 2, \cdots \}$. At each period the firm is matched with a continuum of potential consumers uniformly distributed between $[0, 1]$ indexed by $i$. The utility that a consumer located at $i \in [0, 1]$ gets from purchasing a product with quality $q$ and price $p_t$ is given by:

$$u_i = q - p_t + i.$$ 

So implicitly this specification assumes that the firm is located at point 1 from the unit interval, and that customers located closer to the firm place a higher valuation for the product. As highlighted by Tirole [1988], the connotation of location does not have to be geographical: it can represent differences in tastes which causes consumers to have heterogeneous willingness to pay for a certain product.

Letting $\mu_t \equiv E_t(q)$ denote the expected quality from the firm given customers’ beliefs at time $t$, we have that a consumer $i$ will purchase the product if and only if

$$\mu_t - p_t + i \geq 0$$

$$\iff i \geq p_t - \mu_t.$$ 

This implies that if $p_t > 1 + \mu_t$ the demand is zero; if $p_t < \mu_t$ the demand is the entire unit interval; and finally, if $\mu_t \leq p_t \leq 1 + \mu_t$, the demand is given by $1 - (p_t - \mu_t)$. So in
the end the demand faced by the firm is given by

\[
D(\mu_t, p_t) = \begin{cases} 
1, & \text{if } p_t < \mu_t \\
1 - p_t + \mu_t, & \text{if } \mu_t \leq p_t \leq 1 + \mu_t \\
0, & \text{if } p_t > 1 + \mu_t 
\end{cases}
\]

Now assume that, after observing \(\mu_t\), the firm chooses the price \(p_t\) that maximizes its revenues \(D(\mu_t, p_t)p_t\) (i.e., the firm is assumed to face zero marginal cost of production, so that its profits equals to its total revenue). We also make the high level assumption that customers do not update their beliefs regarding \(q\) after observing the price \(p_t\). This is more a result than an assumption, since it can be shown that, in the event prices can be used as a signal, there exists a “pooling” Bayesian equilibrium in the sense that all types with the same reputation charge the exact same price, given by the price that maximizes their revenue. Intuitively, for games of incomplete information in which the costs from sending a signal is the same for all types (in the current situation, the signal being the price), one should expect all types to send the same signal. So if all firm types are to choose the same price, they might as well choose the price that maximizes their expected revenue.

If \(0 \leq \mu_t \leq 1\), then the optimal price chosen by the firm is given by \(p_t = (1 + \mu_t)/2\), which yields the firm an expected revenue of

\[\omega(\mu_t) = \frac{(1 + \mu_t)^2}{4}.\]

3.2 Allowing the firm to manipulate customers’ beliefs through fake reviews

Now assume that at each period the firm has the ability to influence \(\mu_t\) by exerting some effort \(\eta_t \geq 0\) in the fabrication of reviews that praise its own products. As an example, in Akerlof’s market of lemons (Akerlof [1970]), all sellers are assumed to charge the same price, irrespective of the quality of the cars being sold. Fake reviews disparaging the firm’s rivals should have a similar effect as to fake reviews that praise the firm’s products. The main reason as to why I restrict attention on fake positive reviews is because empirically it is hard to detect the culprits from negative fake reviews (it could be anyone of the firm’s rivals), whereas fake positives are usually orchestrated by the firm that is having its products praised.
We assume the cost from choosing \( \eta_t \geq 0 \) is given by \( \lambda \eta_t^2 \). At each period \( t \) consumers get a noisy signal about the quality from the firm, given by

\[
v_t = q + \eta_t + \varepsilon_t,
\]

where \( q \) and \( \varepsilon_t \) are independent random variables that are not observable by customers. The term \( q + \varepsilon_t \) from this expression can be interpreted as the part from the signal generated from honest reviews, while \( \eta_t \) is the fraction from the signal attributed to review manipulation financed by the seller. After observing \( v_t \), customers update their beliefs regarding the distribution of \( q \) to form their expectation \( \mu_t \) of \( q \), and then decide whether or not to purchase the product.

We now closely examine how customers update their beliefs for a specific data generating process (DGP) for \( q \) and \( \varepsilon_t \).

### 3.2.1 Customers’ Bayesian updates

Assume that \( q \in \{0, 1\} \), and let \( \mu_0 = \mathbb{E}(q) = \text{Prob}(q = 1) \). Also 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 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 \( \eta : \{0, 1\} \times [0, 1] \to \mathbb{R}^+ \) as the amount of effort dedicated in faking reviews 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, starting at initial beliefs \( \mu_0 \), customers’ beliefs and the seller’s choices obey the following Markov process: for all \( t = 0, 1, 2, \cdots \),

\[
\mu_t = \frac{\mu_{t-1} e^{-\frac{(v_t - 1 - \eta(q, \mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1} e^{-\frac{(v_t - 1 - \eta(q, \mu_{t-1}))^2}{2\sigma^2}} + (1 - \mu_{t-1}) e^{-\frac{(v_t - \eta(0, \mu_{t-1}))^2}{2\sigma^2}}},
\]

where

\[
v_t = q + \eta(q, \mu_{t-1}) + \varepsilon_t.
\]

---

5For tractability, this framework does not model customers’ incentives to leave reviews; instead, it just assumes that the signal generated from sellers’ honest reviews are stochastic and positively correlated with their quality. For a theoretical framework that examines buyers’ incentives to post reviews, see Campbell, Mayzlin and Shin 2017.
(ε_t)_{t=1}^{∞} is iid with ε_t ∼ N(0, σ^2), and 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 − μ_0.

Implicitly we have made the high level assumption that consumers expect the strategy chosen by the seller, η: {0, 1} × [0, 1] → R^+, to only depend on the seller’s type and on the previous beliefs μ_t held by customers. As shown in the next section, given such beliefs pertaining the strategy chosen by the seller, q and μ_t will indeed be sufficient statistics for the seller’s optimal policy in period t.

3.2.2 Seller’s optimal choice of review manipulation

Once consumers update their beliefs μ_{t+1} in period t + 1, the firm’s expected quality is given by μ_{t+1}, which yields the firm a profit of ω(μ_{t+1}) = (1 + μ_{t+1})^2/4 (see section 3.1).

So if customers expect the firm to adopt strategy η(q, μ), we have that, starting at the initial μ_0, a firm with quality q ∈ {0, 1} chooses a sequence of (˜η(q, μ^t))_{t=1}^{∞} that solves:

\[
\max_{\tilde{\eta}(q, \mu^{t-1})} \sum_{t=1}^{\infty} \delta^{t-1} \left[ \mathbb{E}_{t-1}[\omega(\mu_t)] - \lambda \tilde{\eta}(q, \mu^{t-1})^2 \right]
\]

s.t. \[ \mu_t = \frac{\mu_{t-1}e^{-(v_t-\eta(1,\mu_{t-1}))^2/2\sigma^2}}{\mu_{t-1}e^{-(v_t-\eta(1,\mu_{t-1}))^2/2\sigma^2} + (1 - \mu_{t-1})e^{-(v_t-\eta(0,\mu_{t-1}))^2/2\sigma^2}}, \]

\[ v_t = q + \tilde{\eta}(q, \mu^{t-1}) + \epsilon_t, \]

where (ε_t)_{t=0}^{∞} is iid with ε_t ∼ N(0, σ^2), and \( μ^{t-1} = (μ_1, μ_2, \cdots, μ_{t-1}) \) is the entire history of beliefs up to time \( t - 1 \), and \( \delta \in [0, 1) \) is the firm’s discount factor. At this point it is important to emphasize the distinction between \( \tilde{\eta}(q, \mu^{t-1}) \) and \( \eta(q, μ_{t-1}) \). \( \tilde{\eta}(q, \mu^{t-1}) \) is the strategy adopted by the firm, while \( \eta(q, μ_{t-1}) \) is what customers think what the strategy from the firm will be, which the firm takes as given. This is actually one of the main distinctions between this model and standard advertising models: in a standard advertising model, the amount of advertising is observed by customers, so the firm takes into account the direct impact that advertising has on customers’ beliefs pertaining the strategy adopted by the firm; but in the present model customers do not update their beliefs regarding the strategy taken by the firm once they observe a signal, since customers can not observe the effort undertaken by the firm in review fraud (Mayzlin 2006).

Because the expected payoff from the firm at period t only depends on the choice of \( \tilde{\eta} \) made by the firm at that period and on the variables q and \( \mu_t \), we can write the above...
sequential problem as a functional equation, where the state variables are $\mu_t$ and $q$ (notice that $q$ is determined in period 0 and does not change over time). Therefore, from the principle of optimality, one can find the solution to the sequential problem by solving the following Bellman equation:

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

(4)

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

(5)

**Proposition 3.1** Given $\eta(q, \mu)$, and imposing the constraint that the amount of fake reviews chosen by the firm, $\tilde{\eta}$, can not exceed a certain upper limit $\tilde{\eta} > 0$, i.e, $\tilde{\eta} \in [0, \tilde{\eta}]$, we have that the Bellman equation has a unique solution.

So given that customers believe that the seller adopts strategy $\eta(q, \mu)$, there is a unique solution to the seller’s problem of choosing the optimal expenditure on review manipulation.

One can also easily show that, for the extreme points in which $\mu = 0$ or $\mu = 1$, the seller’s optimal strategy consists on choosing $\tilde{\eta} = 0$, regardless of consumers’ guess regarding the strategy taken by the seller, $\eta(q, \mu)$. Indeed, at those points the signal generated from reviews can not affect customers’ beliefs, so that the seller has no incentives to try to influence the signal. This extreme result can be relaxed by allowing the seller’s type to change over time according to a certain Markovian process. But qualitatively, adding that additional friction does not affect the main predictions of the model. So we now proceed to describe the equilibrium from this economy.

### 3.3 Equilibrium

Informally, the perfect Bayesian equilibrium (PBE) equilibrium from this economy is given by a policy function from the seller, and a policy function for customers such that: 1) Customers’ maximize their expected utility when deciding whether or not to purchase a product, given their beliefs regarding the seller’s type; 2) the seller maximizes its expected profits given customers’ beliefs and customers’ strategy; 3) Customers’ beliefs regarding the seller’s type are correctly updated through Bayes’ rule.
Definition 3.1 (Equilibrium) Given the initial probability of a firm being of high type, \( \mu_0 \), and the firm’s quality \( q \), a PBE from this economy is characterized by a strategy \( \eta \) dictating the effort chosen by the firm at the beginning of each period as a function of its quality \( q \in \{0,1\} \) and its current reputation \( \mu \in [0,1] \), and consumers’ beliefs, such that, for every \( (q,\mu) \in \{0,1\} \times [0,1] \),

\[
\eta(q,\mu) \in \arg \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\eta)^2}{2\sigma^2}} \left[ \omega(\mu') - \lambda \tilde{\eta}^2 + \delta V(q,\mu') \right] dv
\]

s.t. \( \mu' = \frac{\mu e^{-\frac{(v_1-\eta(q,\mu-1))^2}{2\sigma^2}}}{\mu e^{-\frac{(v_1-\eta(0,\mu-1))^2}{2\sigma^2}} + (1-\mu)e^{-\frac{(v_1-\eta(0,\mu-1))^2}{2\sigma^2}}} \),

and customers update their beliefs through Bayes’ rule, so that \( (\mu_t)_{t=1}^{\infty} \) obey the following stochastic process:

\[
\mu_t = \frac{\mu_{t-1}e^{-\frac{(v_t-\eta(q,\mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1}e^{-\frac{(v_t-\eta(q,\mu_{t-1}))^2}{2\sigma^2}} + (1-\mu_{t-1})e^{-\frac{(v_t-\eta(q,\mu_{t-1}))^2}{2\sigma^2}}} \quad \forall t \geq 1,
\]

where \( v_t \sim N(q-\eta(q,\mu_{t-1}),\sigma^2) \).

3.4 Finding the equilibrium numerically

An alternative way of interpreting the PBE concept is to think of the seller as choosing a policy function \( \tilde{\eta}(q,\mu) \), and then having customers guessing a policy function \( \eta(q,\mu) \) chosen by the seller, and then requiring that:

I) Given \( \eta(q,\mu) \), the strategy adopted by the seller, \( \tilde{\eta}(q,\mu) \), is a solution to the Bellman equation 4 (i.e., the seller chooses the optimal amount of fake reviews given consumers’ expectations regarding the strategy chosen by the seller);

II) \( \tilde{\eta}(q,\mu) = \eta(q,\mu) \) (i.e., customers correctly guess the strategy adopted by the seller).

This way of thinking about the PBE motivates the usage of the following algorithm for finding the equilibrium:

Algorithm 1 (Finding the PBE numerically)

i) Guess a strategy \( \eta(\cdot,\cdot) \).

ii) Given this strategy, solve the Bellman equation 4 (say, by iterating the value function) to obtain a new guess \( \tilde{\eta}(\cdot,\cdot) \) for the policy function, then go to the next step.
iii) Compare $\tilde{\eta}(\cdot, \cdot)$ obtained in the previous step with $\eta(\cdot, \cdot)$. If these two policy functions are sufficiently close to each other (i.e., if $\sup_{(q, \mu)} |\tilde{\eta}(q, \mu) - \eta(q, \mu)|$ is sufficiently small), an approximation to the fixed point representing the seller’s equilibrium strategy has been found, so stop the algorithm; else redefine $\eta(\cdot, \cdot) = \tilde{\eta}(\cdot, \cdot)$ and repeat step ii.

Applying this algorithm, we obtain the equilibrium strategy from the seller as depicted in figure 2. As it is clear from this figure, regardless of its type, a seller optimally chooses to exert more effort on review manipulation for intermediate values of $\mu$, the intuition being that, for very low levels of reputation the seller finds it too costly to signal that it is of high quality, whereas a seller that has already accumulated a very good reputation does not need to prove that it sells a high quality product. In mathematical terms, reputation levels of $\mu = 0$ or $\mu = 1$ are absorbing states: once a seller achieves those reputation levels, they can not be altered.

Figure 2: Equilibrium as a function of $\mu$, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$.

Another interesting feature from this equilibrium is that low quality sellers do not necessarily exert more effort on review manipulation. Which type spends most effort on review manipulation depends on the current level of reputation held by the seller. Indeed, given a very low level of reputation, a high quality seller should spend more effort
on review manipulation, and the opposite should hold when the seller has a very high reputation.

The intuition being this result can be explained as follows. Imagine that a seller currently possesses a very low reputation. In that case, the disutility from getting a bad signal is not so great, since the seller is already close to “rock bottom”. The gain in utility, however, from getting a good signal is more promising, as it can help the seller to separate itself from low types. In that case, because it is too costly for low types to pretend that they are high types, a high type should spend proportionally more on review manipulation. Analogously, if the seller has already accumulated a good reputation, then the marginal benefit from getting a good signal is relatively small, as compared to the disutility from getting a very bad signal. Because high quality sellers are very unlikely to get a very bad signal, it should be the low types, on that case, that will put more effort on the fabrication of fake reviews.

This result has implications on the dynamic choices made by the seller. Indeed, if the initial $\mu_0$ is very low (i.e., if a firm in the market is most likely to be of low quality), then high quality types should be expected to be the ones spending most effort on review manipulation throughout the periods, as depicted in figure 18a. If, however, $\mu_0$ is large, then it is the low quality sellers that should be expected to spend most effort on review manipulation over the periods, as depicted in figure 3c.

From figure 3 one can also see that both types spend most effort on review manipulation at the initial stages after they have opened their businesses, so as to maximize the impact from each fake review.

Now turning the attention to the evolution of reputation held by the seller, one can see from figure 4 that the reputation from a high quality seller tends to improve over time, while the reputation from a low quality seller gradually deteriorates. This happens because, as high quality sellers systematically receive positive reviews praising their products, their reputation tends to improve over time. For low quality sellers, however, it is too costly to maintain a high reputation in the long run due to the fact that they systematically receive negative reviews from honest consumers. So the model essentially predicts that in the long run customers learn the truth regarding the quality from sellers.

One can also compare the rate at which buyers learn the truth about the seller in this model with the case in which sellers are not allowed to fake reviews, say, because
Figure 3: Average simulated effort of review manipulation chosen throughout the periods, starting at different $\mu_0$’s, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. $\eta_H$ corresponds to the average effort of review manipulation chosen by a high type seller (i.e., a seller with $q = 1$), while $\eta_L$ corresponds to the average effort chosen by a low type seller (i.e., a seller with $q = 0$).

Figure 5 compares the evolution of reputation for a high and low quality seller, when fake reviews is allowed to take place, and when it is not. As it is clear from these plots, the rate at which buyers learn the truth about the seller depends on the initial probability that the seller is of high type. Indeed, as discussed previously, high types will have incentives to fake more reviews than low types for low levels of reputation (see figure 2). In that case

the monitoring of fake reviews is very intense or because the punishment applied to those caught faking reviews is very harsh, so that sellers have no incentives to fake reviews.
Figure 4: The average simulated evolution of reputation, starting at different $\mu_0$’s, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. $\mu_H$ corresponds to the average simulated reputation from a high type seller (i.e., a seller with $q = 1$), while $\mu_L$ corresponds to the average simulated reputation from a low type seller (i.e., a seller with $q = 0$).

reviews are actually more informative when sellers are allowed to fake reviews, as the gap between the signal generated by high quality sellers and low quality sellers is greater. So in that case, buyers learn the truth about the type from the seller at a faster rate when fake reviews are allowed to take place, as depicted in figure 5a. On the contrary, for high levels of reputation, it is the low type seller that has more incentives to fake reviews. So in that case allowing sellers to fake reviews closes the gap between the signal generated from high and low types, thus making reviews less informative, which in turn decreases
the rate at which buyers learn the truth about the sellers’ types as depicted in figure 5b.

![Graph](image)

(a) $\mu_0 = .2$

(b) $\mu_0 = .8$

Figure 5: The average simulated evolution of reputation, when the seller is allowed and not allowed to fake reviews, starting at different $\mu_0$’s, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$.

### 3.5 When consumers do not anticipate review manipulation

The results form the previous section were built with the assumption that consumers know the strategy taken by the seller in equilibrium, i.e., that consumers correctly anticipate that some reviews may be fake, and that high and low quality sellers manipulate reviews in different proportions. But one can also imagine scenarios in which customers are unaware of the existence of fraudulent reviews, or at least greatly underestimate how prevalent they are. So this section presents the results from the model in a scenario in which consumers incorrectly believe that the effort on review manipulation is zero for both high and low quality sellers.

Figure 6 displays the equilibrium strategy from both high and low quality sellers in this new environment with naive consumers, together with the equilibrium strategy from the standard version of the model presented in the previous section. At least for the set of parameters under consideration ($\delta = .8$ and $\lambda = \sigma^2 = 1$), when consumers are unaware of the existence of fake reviews, high quality sellers tend to engage in more review manipulation, while low quality sellers end up faking less reviews. As this diminishes the gap between the signals from high and low quality sellers, consumers take longer to learn
the true type from the seller, as displayed in figure 7.

Figure 6: Equilibrium as a function of $\mu$, 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 $q, \mu$).

But qualitatively speaking, the results from either version of the model are very similar. Figure 8 displays the seller’s average simulated effort on review manipulation through time for both scenarios. In both scenarios the seller tends to fake more reviews at the beginning, and as its reputation gradually improves (for a high type seller) or deteriorate (for a low type seller) it gradually reduces its effort on review manipulation as time goes by.

3.6 Allowing the firm to exit and reenter the market with a new name

Now assume that the market starts with $N$ sellers. As before, at each period a seller acts as a monopolist on its own market and they are each matched with a continuum of consumers with mass 1, and a seller’s optimal profits given that customers believe that its expected quality is $\mu$ is given by $\omega(\mu) = (1 + \mu)^2 / 4$ (see section 3.1). At the beginning of each period, each seller retires with an exogenous probability $p_e$. At every period, $[p_e N]$
Gustavo Saraiva

Figure 7: The average simulated evolution of reputation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different $\mu_0$'s, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. $\mu_H$ corresponds to the average reputation from high quality sellers, while $\mu_L$ corresponds to the average reputation from low quality sellers.

Figure 8: The average simulated evolution of effort on review manipulation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different $\mu_0$'s, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. $\mu_H$ corresponds to the average reputation from high quality sellers, while $\mu_L$ corresponds to the average reputation from low quality sellers.
new sellers enter the market, where they are each high type \( q = 1 \) with probability \( \mu_0 \) and low type \( q = 0 \) with probability \( 1 - \mu_0 \). But now sellers can “pretend” to retire and reenter the market with a new name, after paying a fixed cost \( C > 0 \). If consumers were oblivious of this scheme, they would believe a newcomer to be of high quality with ability \( \mu_0 \). But it is assumed that customers correctly anticipate that some firms may try to exit and reenter the market with a new name in order to hide a potential bad reputation obtained from previous reviews.

For technical reasons, it is assumed that at each period there is a small probability \( \rho_s \) that the seller is not allowed to exit and reenter the market. Without this assumption buyers would be allowed to have any beliefs whatsoever regarding the effort spent on review manipulation from a seller with a sufficiently low level of reputation. Indeed, once a seller’s reputation goes below a certain threshold, the seller optimally chooses to exit and reenter the market with a new name. So if sellers were always allowed to do that, in equilibrium one would never observe a seller with a very low level of reputation choosing some effort of review manipulation, which would then allow buyers to have any arbitrary beliefs regarding how much a seller would spend on fake reviews on those hypothetical scenarios. Because arbitrary beliefs pertaining decisions taken outside the equilibrium path can potentially affect the decisions made on the equilibrium path, it is important to impose some discipline on customers’ beliefs on contingencies that are never reached in equilibrium. One way to accomplish that is by using equilibrium refinements. For example, instead of assuming that agents play a PBE, one could assume that they play a sequential equilibrium. But for the current model, imposing discipline on customers’ beliefs can be more easily achieved by simply assuming that there is a small probability \( \rho_s \) that the seller is not allowed to exit and reenter the market. This way, all feasible contingencies can be reached with positive probability, which implies that customers’ beliefs regarding the actions taken by the firm in each contingency have to coincide with the firm’s actual actions taken on those contingencies.

Henceforth a seller is defined as an *apparent newcomer* for period \( t \) if the seller has either started selling its product in period \( t \), or if the seller was already selling its product before period \( t \) but changed its name in period \( t \), so as to erase his past reputation. In that case, deriving the reputation from apparent newcomers can be difficult, since the reputation form those sellers should depend on the strategies chosen by the firms in
equilibrium. So in order to compute the reputation from those firms I rely on Monte Carlo simulations. More precisely, I add an outer loop to algorithm 1 and iteratively compute the realized expected quality from newcomers in the long run to then use that statistic as a new guess for the reputation from newcomers, and keep repeating this process until a convergence criterion is reached.

Algorithm 2 (When firms can exit and reenter the market)

i) Guess $\tilde{\mu}_0$, the probability that a *an apparent newcomer* is of high type in the long run (notice that because the term “an apparent newcomer” also encompasses old sellers that pretend to be new ones, $\tilde{\mu}_0 \neq \mu_0$).

ii) Given $\tilde{\mu}_0$, compute the strategy taken by the firm by implementing a procedure similar to algorithm 1, then go to the next step.

iii) Given the sellers’ strategy conduct Monte Carlo simulations to compute $\tilde{\mu}_0'$, the long run expected probability that an apparent newcomer is of high type. If $|\tilde{\mu}_0' - \tilde{\mu}_0|$ is sufficiently small, stop the algorithm, else redefine $\tilde{\mu}_0 = \tilde{\mu}_0'$ and repeat step ii.

Applying this algorithm to a given set of parameters, we find a similar optimal strategy for the seller, as the one obtained previously, as depicted in figure 9.

But because now sellers can exit and reenter the market with a new name once their reputation has been squandered, the dynamics from their effort on review manipulation changes. In particular, the reputation from low quality sellers tend to reach low levels constantly, which leads them to periodically exit and reenter the market with a new name, always concentrating review manipulation at the time they reenter the market, so as to maximize the impact from each fake review. This phenomenon can be visualized in figure 10a, which depicts the simulated effort of review manipulation from a single high and low type seller, while figure 10b depicts their corresponding level of reputation, assuming they are never selected to retire (recall that there is an exogenous probability $p_e$ that at each given period the seller retires).

Figure 11 averages the strategy from sellers and their respective reputation over several simulations. As it is clear from those plots, consumers on average do not learn the type from low quality sellers in the long run, as they systematically exit and reenter the market.
Figure 9: Equilibrium as a function of $\mu$, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$ and $C = .01$.

![Figure 9: Equilibrium as a function of $\mu$](image)

Figure 10: The evolution of the effort of review manipulation and reputation chosen by a high and a low quality seller, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$.

![Figure 10: The evolution of the effort of review manipulation and reputation](image)

with a new name once their reputation has been squandered. This result suggests that building obstacles that prevent sellers from anonymously selling their products under different account names should be pursued by online platforms such as Amazon, so as to guarantee that reviews are informative.
Figure 11: The evolution of the average effort of review manipulation and reputation from high and low quality firms, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$.

4 Empirical Strategy

The theoretical model from the previous section provides a series of predictions regarding the pattern form review manipulation observed in online rating platforms. One of them is that the process of review manipulation is usually more concentrated during the initial periods following a seller’s entrance (or reentrance with a new brandname) into the market, as sellers wish to maximize the impact from each fake review (see figures 3 and 11a). Another prediction is that the amount of effort dedicated in review manipulation does not vary monotonically with the seller’s reputation. Indeed, from figures 2 and 9, sellers with very high or very low reputation levels will usually spend less effort manipulating reviews. For sellers with low reputation, that happens because they find it too costly to pretend that they are high quality types, in which case they optimally choose to either give up trying to convince buyers that they are high quality sellers, or to exit and reenter the market with a new brand name in order to reenter the game with a fresh reputation. Analogously, for sellers with a very good reputation, the marginal benefit from faking reviews is small, since at that point buyers are already pretty much convinced that sellers are of high quality.

In order to verify to which extent those two predictions are observed empirically, I first scraped data from reviews posted on Amazon from a set of products that I classified as
suspicious on the basis that they were soliciting positive fake reviews for their products on
the internet (namely, on Facebook and Rapidworkers). Using that data I then apply a set
of criteria to identify fake reviews from those sellers. Once fake reviews are identified, I
then run a Logit regression to estimate the probability that a review is fake as a function
of the product’s reputation and the time since the product was first introduced in the
market.

Because this dataset is comprised exclusively of suspicious products, it allows the
researcher to detect fake reviews more easily. Indeed, if the researcher knows that two
different products were involved in review solicitation, and he also observes that the same
customer posted reviews on these two suspicious products, then the researcher can safely
assume that those reviews are almost certainly fake, given that Amazon has a million
of other products from which that customer could have chosen from, thus making the
observed event very unlikely to have happened merely by chance. Clearly this detection
criterion wouldn’t be effective at all if applied to products for which one has no prior
knowledge about their involvement in review solicitation.

Therefore, an advantage about using this dataset to study fake reviews is that it allows
the researcher to detect fake reviews more easily, which in turn diminishes the occurrences
of classification errors in the sample (i.e., it diminishes the number of instances in which
the researcher incorrectly classifies a fake review as real and vice versa).

But using this dataset also has a few caveats, one of them being that it may be sus-
ceptible to selection bias. Indeed, by focusing the analysis on sellers who are known to
solicit fake reviews on the internet, the resulting sample may end up with an overrep-
resentation of fake reviews. For that and other reasons, on section 4.4 I then repeat a
similar estimation analysis using a different set of products from Amazon, one in which
sellers were not targeted in the sampling process.

Consistent with the theoretical model, the results from both regressions indicate that
the probability of a review being fake decreases with time, and it varies non-monotonically
with the seller’s reputation, where very high or very low reputation levels are associated
with a lower probability of the review being fake.
4.1 Database

This study uses a brand new dataset of reviews scraped from Amazon. The dataset contains information from 16,935 reviews from 206 different products. Information from each review includes the review text, the review title, the date the review was posted, the number of stars given by the reviewer, whether the review came from a verified purchase, whether the review contained pictures or videos, whether the review received positive feedback, etc.

As to the 206 products for which reviews were collected, they were individually selected based on the fact that their sellers were soliciting fake review on online platforms such as RapidWorkers and Facebook. Knowing that these sellers were soliciting fake reviews online allows one to more easily detect which of the reviews posted were fake (see section 4.1.1), which then allows the researcher to identify the characteristics from fake reviews.

Figure 12 provides an example of an Amazon seller soliciting fake reviews through RapidWorkers.

![Figure 12: An example of a seller soliciting fake reviews through RapidWorkers.com.](image)

Figure 12 helps to illustrate two aspects about fake reviews solicited through RapidWorkers: 1) when sellers use this website to solicit fake reviews, they provide a link specifying the url from the product in question, which can then be used by the researcher
to identify the seller responsible for the fake review solicitation. 2) Moreover, notice that the amount paid per fake review is usually very small, given that this website focuses on the solicitation of non-verified purchase reviews, which are virtually costless to fabricate. This small amount paid for unverified purchase reviews is consistent with anecdotal evidence provided in the News (see for instance A Rave, a Pan, or Just a Fake? (New York Times, 2011) and other scientific research that also relies on RapidWorkers to spot fake reviews, such as Kaghazgaran et al. (2017) Kaghazgaran, Caverlee and Alfifi n.d.).

As to verified purchase reviews, since they are perceived as being more reliable, and since they are more costly to fabricate, sellers are usually willing to pay a higher price for those type of fraudulent reviews, oftentimes by completely reimbursing reviewers after they have purchased the product and left their positive reviews. Figure 13 shows an example of a seller soliciting verified purchase fake reviews through a community on Facebook dedicated solely to facilitate the transaction of Amazon fake reviews.

![Figure 13: An example of a seller soliciting fake reviews through a Facebook community.](image-url)
Unlike Rapidworkers, on Facebook sellers tend to be less brazen when it comes to fake review solicitation. For starters, on Facebook sellers usually talk using code language. For the example in figure 13, “PP” stands for PayPal, which means that the buyer will be reimbursed through PayPal after purchasing the product through Amazon and leaving a positive review. “US” means that the purchase must be purchased in the US. And “PM” stands for private messaging, meaning that whoever is interested in the gig should contact the seller through Facebook’s private message system.

On top of that, sellers that solicit fake reviews through Facebook usually do not provide the url from their product. Instead they only display a picture and sometimes a small product description. This makes it harder for researchers and Amazon staff to detect the sellers responsible for the reviews solicitation. However, by doing a Google image search, one can identify some of these products. Indeed, for the product in question, a Google image search leads to the page depicted in figure 14. From the figure, one can see that, surprisingly, the product in question has an Amazon best-seller stamp.

So by doing a google image search on some of the products listed in Facebook communities, I was able to identify sellers that were probably soliciting fake reviews for their products.

Finally, another set of suspicious sellers were targeted using information posted from Amazon’s seller forum, where users would complain about the credibility of some of the reviews from certain products. As an example, a post from the forum would read:
“...This item gets over 5 reviews a day verified reviews and they are all 5 stars how is that even possible. Link below:

https://www.amazon.com/Ashwagandha-EnhancerArtichokeEnhanced-Supplement/dp/B06XC9CZWN/

Is this legit or fake reviews. Am I missing out on some method to have this many reviews or are they all paid for...”

Edited by: Wholesale promo on Apr 30, 2017 1:03 PM

So I searched for similar posts and targeted the products being flagged as suspicious by concerned users as well as some other products that were being reviewed by the same set of customers.

In the end I ended up with a database of reviews 26,971 from 247 different products. Some of those products were targeted based on the fact that they were either soliciting fake reviews through RapidWorkers or Facebook, or they were being flagged as suspicious by concerned users on Amazon’s seller forum. Table 2 provides some summary statistics about the dataset.

<table>
<thead>
<tr>
<th>Source</th>
<th>Number of products</th>
<th>Number of reviews</th>
</tr>
</thead>
<tbody>
<tr>
<td>RapidWorkers</td>
<td>61</td>
<td>4,724</td>
</tr>
<tr>
<td>Facebook</td>
<td>165</td>
<td>20,685</td>
</tr>
<tr>
<td>Amazon Seller Forum</td>
<td>21</td>
<td>1,562</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>247</strong></td>
<td><strong>26,971</strong></td>
</tr>
</tbody>
</table>

Table 2: Number of products/reviews collected from each source

4.1.1 Fake review detection

As mentioned earlier, this study focuses in analyzing the patterns from positive fake reviews only. While fraudulent negative reviews aimed at a seller’s rivals should have a similar effect on sales as compared to fabricating positive reviews praising the seller’s own products; for the first tactic it is usually difficult to determine the agent(s) responsible for the review fabrication: they could have originated from any of the seller’s rivals, or even
some disgruntled consumer. For positive fake reviews, on the other hand, it is invariably
the seller receiving the positive fake reviews that is behind their solicitation.

Now, given Amazon 1 to 5 star rating metric, classifying a review as positive or negative
can be subjective: if the average rating given on Amazon was around 2 stars, then a 3
star would actually be considered a good review. But given that the great majority of
reviews posted on Amazon are 5 star reviews (see figures ?? and [17]), I classify a review
as positive if and only if it has more than 4 stars. So reviews with 4 or 5 stars will be
our candidates for positive fake reviews.

Depending on the source that led us to include a certain product in our list, a different
set of criteria was used to determine whether a positive review from that product was
fake or not. The combination of the two criteria listed below was used to categorize
reviews from products in which fake review solicitation was happened on RapidWorkers
or Facebook:

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 5 words, and
they were both from products in which fake review solicitation happened in the same
online platform, and they both gave the seller at least 3 stars, then those reviews
were classified as fake.

II) If a reviewer id was linked to two or more reviews from two or more different products
that were soliciting fake reviews on the same online platform, and if the correspond-
ing reviews gave the sellers a grade of at least 3 stars, then they were classified as
fake.

All other reviews were classified as real. Of course, this process inevitably lead us to
incorrectly classify some actual fake reviews as real. But one should keep in mind that,
when it comes to fake review detection, no classification is perfect. Certain methods,

\[6\] Jabr and Zheng [Jabr and Zheng 2014] adopt the same criteria for classifying an Amazon book review
as positive.

\[7\] To compute the Jaccard similarity index I used shingles containing 4 consecutive words, and I em-
ployed a hashing algorithm that addresses the computational burden of computing the actual Jaccard
Similarity index by computing an unbiased estimator of the index. For details, check section \[A.2\] from
the appendix.
however, can be implemented to correct for potential misclassification errors, as discussed later in section 4.3.

As to reviews from products that were being discussed on Amazon seller forums, a more conservative approach had to be implemented in order to classify them as fake or real. Indeed, from the way in which products were targeted using Amazon seller forums (see section 4.1), intersections among the products reviewed by customers are to be expected even when the reviews in question are real. So if one finds a buyer reviewing two different products from the list of products targeted using Amazon seller forum, that does not constitute a strong indication that the review is fake, so that criterion II listed above does not effectively detect fake reviews for that sample. So for products targeted using Amazon seller forum, while still preserving criterion I, I replaced criterion II for the more conservative classification rule:

II*) Consider an undirected graph where each reviewer is linked to the products that they review. If a cycle containing 2 or 3 reviewers is formed, then the corresponding reviews responsible for that cycle are classified as fake.

Applying the criteria described above to our dataset, 3,834 of the total of 26,971 reviews are classified as fake, while the remaining 23,137 reviews are classified as real. That is, approximately 14% of the reviews from the sample were classified as fake. Though that is a lot of fake reviews, one should keep in mind that, when building our sample we deliberately targeted suspicious products, so as to simplify the task of detecting fake reviews. Therefore one should not interpret this percentage as an accurate depiction as to how prevalent fake reviews are in online platforms such as Amazon.

4.1.2 Measuring reputation

The theoretical model from section 3 defined a seller’s reputation as customer’s beliefs that the seller is of high type. Of course, such probability is not observable in practice, which motivates the usage of some statistic that captures this reputation. One possibility would be to use the average number of stars received by a seller at any given time. The problem with this statistic, however, is that it would imply that if a product had a single review, and that review happened to give the seller 5 stars, then the seller would have the highest reputation score that one could possibly get.
So instead I use the following statistic as a proxy for reputation: 1) each seller starts with a score of zero. 2) For each review received by the seller, the seller’s score is added or subtracted by a certain number depending on whether the review gave him 1, 2, 3, 4 or 5 stars. A 1-star review reduces the score from the seller by 2 points, a 2-star review reduces the raw score by 1 point, a 3-star review does not affect the score, a 4-star review adds 1 point to the score, and a 5-star review adds 2 points to the score. 3) After computing the raw score from seller $s$ at time $t$, $r_{s,t}$, I then normalize it to a 0 to 1 scale by computing

$$\tilde{\mu}_{s,t} = \frac{1}{1 + \exp\left(-r_{s,t}/\sigma_r\right)},$$

the final statistic used to measure the seller’s reputation, where $\sigma_r$ is the standard deviation of the raw score $r_{s,t}$.

As a final observation, notice that this measure of reputation does not take into account that some reviews may be fake. So on that regard it is more aligned with the version of the model in which consumers are naive and do not expect some reviews to be fake (see section 3.5).

4.1.3 Covariates

Some covariates were added in our regressions in order to control for endogenous shocks. Two of these covariates use text analysis to estimate the probability that a review is real conditional on its content. One of these two variables uses content from the review title, while the other uses content from the review text. I call these variables “reliability from review title” and “reliability from review text”. Each of these variables are dummies that assume value 1 if the contents of the review text (review title, resp.) are more likely to have been generated by legitimate reviews. For both of these measures I also imposed the restriction that a review with 1 to 3 stars is real with probability 1 (recall that this paper focuses on the detection of positive fake reviews, since for those cases it is clear who is responsible for the solicitation of fake reviews: the seller who is getting its product praised by dishonest reviewers). The method used for computing these dummies is the naïve Bayes’ estimate. In spite of its name, this statistic has proven to perform surprisingly well as compared to more sophisticated methods. This, added to its simplicity, has lead this statistic to be commonly employed in the computer science literature.
Other covariates include whether or not a review contained a picture or a photograph. Controlling for these variables is important since, from several examples that I found on Facebook and RapidWorkers, buyers who were soliciting fake reviews would occasionally pay extra if a photograph or video was added to the review, supposedly to make the reviews more convincing.

Another important variable consists on the amount of feedback received by a review.\textsuperscript{8} The impact that this variable should have on the probability of a review being fake is ambiguous. Indeed, on the one hand, one would expect that more positive feedback would imply that the others found the review to be useful, and therefore less likely to be fake. But on the other hand, one can see cases in which sellers solicit positive feedback on positive reviews, as depicted in figure [15]. Because of this, it could actually be the case that more positive feedback are associated with a higher probability of a review being fake. To their surprise, Jindal and Liu (2008)\textsuperscript{9} find that the latter occurs in their Amazon dataset, i.e., they find that more positive feedback actually increases the probability of a given review being fake, probably due to the fact that some sellers solicit fake feedback on fake reviews.

An additional variable consists on the number of words from a review text. Again, the effect of this variable on the probability of a review being fake is ambiguous. On the one hand one could argue that, since fake reviewers are usually mass produced, the length from fake reviews are expected to be shorter. Some of the examples discussed earlier, however, suggest that the opposite can occur. Indeed, given that sellers occasionally ask for pictures or videos to be added to the reviews, or ask for positive feedback to be given to favorable reviews, it seems that sellers are not only interested in a acquiring a high volume of positive fake reviews, but they also want those reviews to be convincing. So it may actually be the case that the text from fake reviews are on average longer than real reviews, due to the extra effort put by fake reviewers so as to make their reviews look more convincing.

Last, but not least, a dummy that determines whether or not a review came from a verified purchase is added to the vector of covariates. This variable is potentially a good predictor as to whether or not a review is fake. Indeed, since unverified purchase

\textsuperscript{8}As I write this in late 2018, Amazon only displays positive feedback, not negative ones.
fake reviews are usually a lot cheaper to acquire as compared to verified purchase ones, one would expect unverified purchases to be more likely to be fake. One force, however, that can potentially change the effect from this variable is the selection bias present in our sample. Indeed, most of our sample consists on products for which fake reviews were being solicited on Facebook. Given that the reviews solicited on Facebook were mostly (if not exclusively) verified purchase ones, this can cause our data to have an overrepresentation of verified purchase fake reviews.

Figure 15: An example of a seller soliciting positive feedback to reviews praising its products.
4.2 Logit model

Let $y_{i,s,t}$ be a binary variable that indicates whether review $i$ from product $s$ posted at time $t$ is fake or not, i.e., $y_{i,s,t} = 1$ if the review is fake, and $y_{i,s,t} = 0$ if the review is real.

Next, define the latent variable

$$y_{i,s,t}^* = \tau_{i,s,t} \beta_0 + \beta_1 \tilde{\mu}_{s,t} + \beta_2 \tilde{\mu}_{s,t}^2 + z_{i,s,t} \gamma + v_{i,s,t},$$

(9)

where $\tau_{i,s,t}$ is the time it took for review $i$ from product $s$ to be posted since the first review received by seller $s$, $\tilde{\mu}_{i,s,t}$ is the statistic derived from expression 8, which measures the reputation from seller $s$ up to time $t$; $z_{i,s,t}$ is the vector of covariates described in section [4.1.3] and $v_{i,s,t}$ is an idiosyncratic error term with cdf $F(\cdot)$.

Assume that

$$y_{i,s,t} = \begin{cases} 1, & \text{if } y_{i,s,t}^* \geq 0 \\ 0, & \text{if } y_{i,s,t}^* < 0 \end{cases}.$$  

Given the predictions from the theoretical model, one would want to test the following hypothesis: 1) $\beta_0 < 0$, so that older reviews are more likely to be fake, and 2) that $\beta_1$ is positive, while $\beta_2$ is negative, in such a way that sellers with very low or very high reputation have less incentives to fake reviews. Therefore, denoting $x_{i,s}$ as the vector containing the variables of interest, i.e.,

$$x_{i,s,t} = [\tau_{i,s,t}, \tilde{\mu}_{i,s,t}, \tilde{\mu}_{i,s,t}^2],$$

the loglikelihood function from this model can then be written as:

$$l(\beta, \gamma) = \sum_{i,s,t} [y_{i,s,t} \log(F(x_{i,s,t} \beta + z_{i,s,t} \gamma)) + (1 - y_{i,s,t}) \log(1 - F(x_{i,s,t} \beta + z_{i,s,t} \gamma))].$$

Table 3 displays the results from Logit regressions (i.e., assuming that $F(\cdot)$ is the cdf from a logit distribution). Since the estimates form all specifications are very similar we will analyze the results from a single specification, namely column (1) from the table. Probit regressions also yield similar results and are therefore omitted.

Consistent with the predictions from the theoretical model, the time coefficient is negative and statistically significant, so that the longer it takes for a review to be posted since the seller entered the market, the less likely the review is to be classified as fraudulent. This result is also consistent with the plot depicted in figure 16 which displays the average
Table 3: Regression Results

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$y = 1\text{ (review is fake)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>$-0.795^{***}$</td>
<td>$-5.715^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.607)$</td>
<td>$(0.531)$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>$9.591^{***}$</td>
<td>$16.095^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(1.612)$</td>
<td>$(1.414)$</td>
</tr>
<tr>
<td>$\mu^2$</td>
<td>$-7.199^{***}$</td>
<td>$-11.888^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(1.043)$</td>
<td>$(0.910)$</td>
</tr>
<tr>
<td>time</td>
<td>$-0.008^{***}$</td>
<td>$-0.015^{***}$</td>
</tr>
<tr>
<td></td>
<td>$(0.001)$</td>
<td>$(0.001)$</td>
</tr>
<tr>
<td>Dummy for text reliability</td>
<td>$-2.607^{***}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.059)$</td>
<td></td>
</tr>
<tr>
<td>Dummy for title reliability</td>
<td>$-1.488^{***}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.065)$</td>
<td></td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>$-0.019^{***}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.006)$</td>
<td></td>
</tr>
<tr>
<td>Verified purchase</td>
<td>$-0.209^{***}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.065)$</td>
<td></td>
</tr>
<tr>
<td>Has images or videos</td>
<td>$0.607^{***}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$(0.086)$</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>18,440</td>
<td>18,440</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>$-5,241.875$</td>
<td>$-7,552.646$</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>10,501.750</td>
<td>15,113.290</td>
</tr>
</tbody>
</table>

*Note:* $^*$p$<0.1; ^{**}$p$<0.05; ^{***}$p$<0.01$
proportion of fake reviews chosen by sellers as a function of time. As it is clear from the graph, the bulk of fake reviews is mostly concentrated around the few weeks following a seller’s entrance (or reentrance) into the market. The small increase on fake reviews at the right tail of the graph can be attributed to the fact that this measure becomes increasingly less precise as time advances, since the number of observations used to compute this statistic decreases as time increases (i.e., the longer the time span, the less sellers there are in the sample that have lived long enough to be included in the average).

Figure 16: Average proportion and absolute number of fake reviews chosen by sellers as a function of the time since sellers’ first review. Time was discretized into biweekly intervals. The gray area corresponds to bootstrapped 95% confidence intervals for the average proportion of fake reviews.

Turning back to the results from the regression, the coefficients for reputation and reputation squared suggest an inverted U-shape relationship between reputation and effort on review manipulation. This is consistent with the model’s prediction that sellers that have either accumulated a very high or very low reputation have less incentives to fake reviews.

As to coefficients from covariates, they mostly have the sign that one would normally expect. For instance, the coefficients for the reliability index dummies, which capture the probability that a review is real based on their text content, have negative signs, which
means that reviews with more reliable review texts and review titles are unsurprisingly less likely to be fake (see section A.3.1 for a detailed explanation as to how these dummies were constructed).

A review coming from a verified purchase decreases the probability of a review being fake, most likely due to the fact that verified purchase reviews are more costly to fake. And finally, a review having images or videos increases the probability of a review being fake. That is probably due to the fact that sellers occasionally solicit pictures or videos to be added to fake reviews, so as to make them more convincing.

4.3 Logit model correcting for classification error

The logit model presented in section 4.2 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 real, 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 4.1.1), it is very likely that some of the 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 \( y_{o,i,s,t} \), 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 \( y_{o,i,s,t} \neq y_{i,s,t} \).

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 Tennekoon and Rosenman [2016] that corrects for endogenous misclassifications. Formally, let \( z_{i,t,s} \) be a vector of covariates that can predict whether or not a review is fake, such as the length from the review, whether or not the review was from a verified purchase, whether or not the review 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 \( z_{i,s,t} \) is given by:

\[
\text{Prob}(y_{o,i,s,t} = 1 | y_{i,s,t} = 1, z_{i,s,t}) = F_o(z_{i,s,t} \gamma),
\]

where \( F_o(\cdot) \) is a cdf. Because reviews from our sample were only classified as fake when
strong evidence supported that those reviews were indeed fraudulent (see section 4.1.1).
I assume that a real review from our sample is never incorrectly classified as fake, i.e.,
\[
Prob(y_{i,s,t}^o = 0|y_{i,s,t} = 0, z_{i,s,t}) = 1.
\]
Therefore, letting \( x_{i,s,t} \) denote the vector of explanatory variables of interest (namely, the time it took for the review was posted, and the categorical dummies indicating the cohort of reputation from the product at the time the review was posted), we have that the overall probability of observing \( y_{i,s,t}^o = 1 \) given the covariates from the model is given by
\[
Prob(y_{i,s,t}^o = 1| x_{i,s,t}, z_{i,s,t}) = Prob(y_{i,s,t} = 1|x_{i,s,t})Prob(y_{i,s,t}^o = 1|y_{i,s,t} = 1, z_{i,s,t})
= F(x_{i,s,t}\beta) F_o(z_{i,s,t}\gamma)
\]
With these probabilities, we can then build the loglikelihood function
\[
l(\beta, \gamma) = \sum_{i,s,t} \left[ y_{i,s,t}^o \log(F(x_{i,s,t}\beta) F_o(z_{i,s,t}\gamma)) + (1 - y_{i,s,t}^o) \log(1 - F(x_{i,s,t}\beta) F_o(z_{i,s,t}\gamma)) \right],
\]
and maximize it to obtain estimates of \( \beta \) and \( \gamma \).

The results from this regression are depicted in table 4. Again, the results from the regression are very similar to the ones obtained earlier in section 4.2, 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 and very high levels of reputation \( \mu_{i,s,t} \).

4.4 Alternative database

As mentioned at the beginning of section 4, the database collected from sellers who were either caught soliciting fake reviews or were flagged by users for being involved 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 Logit 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 heterogeneous, 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.
<table>
<thead>
<tr>
<th>variable</th>
<th>estimates</th>
<th>p-values</th>
<th>std errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-164.6***</td>
<td>3.58e-14</td>
<td>2.17e+01</td>
</tr>
<tr>
<td>$\mu_{i,s,t}$</td>
<td>828.9***</td>
<td>2.25e-12</td>
<td>1.18e+02</td>
</tr>
<tr>
<td>$\mu_{t,i,s,t}$</td>
<td>-663.9***</td>
<td>3.37e-12</td>
<td>9.54e+01</td>
</tr>
<tr>
<td>time</td>
<td>-0.016***</td>
<td>3.506e-18</td>
<td>1.82e-03</td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>2.07***</td>
<td>0.0000</td>
<td>7.02e-02</td>
</tr>
<tr>
<td>Dummy for text reliability</td>
<td>-2.605***</td>
<td>0.0000</td>
<td>6.29e-02</td>
</tr>
<tr>
<td>Dummy for title reliability</td>
<td>-1.53***</td>
<td>0.0000</td>
<td>6.86e-02</td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>-0.015</td>
<td>1.00014e-02</td>
<td>5.804e-03</td>
</tr>
<tr>
<td>Verified Purchase</td>
<td>-0.285***</td>
<td>1.019621e-05</td>
<td>6.468e-02</td>
</tr>
<tr>
<td>Has images or videos</td>
<td>0.613**</td>
<td>7.7998e-12</td>
<td>8.95e-02</td>
</tr>
<tr>
<td>Observations: 18,440</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pseudo $R^2$: 0.3953554</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Logit regression after correcting for endogenous classification errors.

To address these issues, I collected a separate dataset comprised exclusively of wireless headsets sold at Amazon, not targeting any seller from such category in the sampling process. The reason I chose wireless headsets is because one can find evidence in the news that fake reviews for these products are very prolific on Amazon, thus making the analysis for this market economically relevant (see for instance [How merchants use Facebook to flood Amazon with fake reviews (April 23, 2018)]). The dataset was then used to estimate a model similar to the one presented in section 4.3.

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 than the average number of reviews from the previous sample. But regarding the distribution of stars, they are very similar for both samples as depicted by figure 17.

4.4.1 Fake review detection

Because there is no prior evidence to suggest that a particular seller from this new dataset solicited fake reviews, I no longer employ criterion number I presented in section 4.1.1 in the fake review detection process. Instead, I rely solely criterion I, i.e., I classify a review
Figure 17: Histogram of the number of stars per sample. The bars in blue correspond to the sample of wireless earphone products, while the one in orange corresponds to the sample described in section 4.1 generated by targeting suspicious products that were either soliciting reviews in online platforms, or were flagged as suspicious on Amazon forums.

Applying this unique classification rule to the sample results in 37,921 of the total of 278,829 reviews being classified as fake, which amounts to approximately 19% of reviews. While this percentage may already seem alarmingly high, it greatly underestimates the actual proportion of fake reviews for this market. Indeed, just as an illustration, consider a couple of products in the sample for which the Jaccard similarity index accused less than 15% of their reviews from being fake. By inspecting those two products more closely, one can find that: 1) more than 99% of their reviews were 5 star and unverified purchase reviews, and 2) they were mostly concentrated around a few days during the time period reviews were posted for these products, as depicted in figure 18. So it is safe to assume that for these two products the Jaccard similarity index alone was not capable to capture all suspicious activity. For this reason we add in our Logit regression variables aimed at detecting potential classification errors, such as adding dummies that assume value 1

Moreover, the computational burden from criterion II*) increases, while its efficacy decreases as the sample size increases, which is another reason not to use this criterion on the new dataset.
when the review was posted during a spike of positive reviews. The criteria for detecting those spikes is explained in more detail on the appendix session A.3.2.

Figure 18: 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.

Now to see how the detection of fake reviews using Jaccard similarity compares with other detection methods, I compute the average number of fake reviews obtained through this method for each different grade category received on Fakespot.com, a platform dedicated to detecting fraudulent reviews on online rating platforms such as Amazon and Yelp. Figure 19 displays the average number of fake reviews per product for each grade category, where “A” corresponds to the best possible grade (i.e., to the lowest level of review manipulation) that a product can get, and “F” corresponds to the worst grade possible. The plot depicts a negative relationship between grades and the number of reviews detected as fake using the Jaccard similarity index. If the relationship is not perfectly decreasing, that is probably due to the combination of two factors: 1) We are using a single criterion to detect fake reviews, namely, the level of text similarity among the reviews, whereas Fakespot seems to use a machine learning algorithm that computes the probability of a review being fake using a combination of several different criteria; 2) Moreover, Amazon has its own fake review detector, and it excludes fraudulent reviews
from its platform on a regular basis. This implies that on several occasions we would encounter a product that engaged in a lot of suspicious activity, and yet had a high score on Fakespot, solely because their fraudulent reviews were removed by Amazon by the time we checked its grade. To mitigate the selection bias caused by having Amazon removing some of the suspicious reviews from the platform, for many products in our sample we scraped their corresponding reviews at several different days, so that our database contains many reviews that were filtered by Amazon, in addition to reviews that were posted after Amazon’s filtering.

Figure 19: Average number of fake reviews detected using text similarity, compared to the product’s grade on Fakespot.com. According to the website, a grade of “A” indicates low level of review manipulation, whereas a grade of “F” indicates a high number of fraudulent reviews. The 95% confidence intervals displayed in the figure were built using 100,000 bootstrap simulations.
4.4.2 Logit Regressions

Table 5 reports the results from standard Logit regressions. The results are mostly similar to the ones obtained before: for the fully specified model, the coefficients of interest follow the right direction predicted by the theoretical model, namely, that as times goes by, reviews are less likely to be fake; and that the relationship between reputation and the probability of a review has a downward parabola shape.

Table 5: Simple Logit regressions using the earphone dataset.

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y = 1 (review is fake)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>−0.457***</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td>(0.156)</td>
<td>(0.129)</td>
</tr>
<tr>
<td>µ</td>
<td>6.006***</td>
<td>−4.615***</td>
</tr>
<tr>
<td></td>
<td>(0.459)</td>
<td>(0.378)</td>
</tr>
<tr>
<td>µ²</td>
<td>−3.851***</td>
<td>4.218***</td>
</tr>
<tr>
<td></td>
<td>(0.320)</td>
<td>(0.263)</td>
</tr>
<tr>
<td>time</td>
<td>−0.001***</td>
<td>−0.006***</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
<td>(0.0001)</td>
</tr>
<tr>
<td>Peak dummy</td>
<td>0.512***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td></td>
</tr>
<tr>
<td>Dummy for text reliability</td>
<td>−2.147***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>Dummy for title reliability</td>
<td>−1.175***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
<td></td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>0.002***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0003)</td>
<td></td>
</tr>
<tr>
<td>Verified Purchase</td>
<td>−1.313***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.019)</td>
<td></td>
</tr>
<tr>
<td>Has images or videos</td>
<td>0.166***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.036)</td>
<td></td>
</tr>
</tbody>
</table>

Observations 232,176 232,176
Log Likelihood −88,368.930 −122,434.000
Akaike Inf. Crit. 176,757.900 244,876.100

Note: *p<0.1; **p<0.05; ***p<0.01

Also, once we add the dummy coefficient that assumes value 1 when the review was posted during an abnormal peak of 5 stars, we see that reviews are more likely to be
classified as fake during those periods. That is not surprising given that in our sample 45% of the reviews posted during abnormal peaks were classified as fake, while only 17% of the reviews outside those peaks were classified as fake.

Moving to the model specification that addresses classification error, we get coefficients similar to the ones obtained using the previous dataset, as displayed in table 6. The only main difference between the two regressions are the signs from the coefficients for reputation, which now display signs not consistent with the inverted U shape relationship between reputation and effort on review manipulation. But other than that, all other coefficients exhibit the same signs as in the previous regressions.

<table>
<thead>
<tr>
<th>variable</th>
<th>estimates</th>
<th>p-values</th>
<th>std errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>70.9***</td>
<td>4.36e-20</td>
<td>7.72</td>
</tr>
<tr>
<td>$\tilde{\mu}_{i,s,t}$</td>
<td>-2.49e+02***</td>
<td>6.33e-19</td>
<td>28</td>
</tr>
<tr>
<td>$\tilde{\mu}_{i,s,t}^2$</td>
<td>2.26e+02***</td>
<td>5.52e-19</td>
<td>25.352</td>
</tr>
<tr>
<td>time</td>
<td>-2.87e-03***</td>
<td>2.40e-14</td>
<td>3.76e-04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable</th>
<th>estimates</th>
<th>p-values</th>
<th>std errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>1.86***</td>
<td>0.000</td>
<td>2.765e-02</td>
</tr>
<tr>
<td>peak dummy</td>
<td>0.523***</td>
<td>0.000</td>
<td>1.867e-02</td>
</tr>
<tr>
<td>Dummy for text reliability</td>
<td>-2.22***</td>
<td>0.000</td>
<td>1.41e-02</td>
</tr>
<tr>
<td>Dummy for title reliability</td>
<td>-1.24***</td>
<td>0.000</td>
<td>1.824e-02</td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>2.29e-03***</td>
<td>1.80e-10</td>
<td>3.596e-04</td>
</tr>
<tr>
<td>Verified Purchase</td>
<td>-1.35***</td>
<td>0.000</td>
<td>1.954e-02</td>
</tr>
<tr>
<td>has images or videos</td>
<td>0.16***</td>
<td>1.29e-05</td>
<td>3.6724e-02</td>
</tr>
</tbody>
</table>

Observations: 232,176  pseudo $R^2$: 0.2933641

Table 6: Logit regression after correcting for endogenous classification errors.

4.5 Placebo test

A placebo test was conducted in order to certify that the correlations obtained in the previous sections were not spurious. To perform the test, we first randomly classify reviews as fake or real according to the empirical distribution from the sample. Since roughly 19% of the reviews in the sample were classified as fake, we randomly choose a review to be assigned as fake with probability .19. Then we run the same regressions
as before with this new random assignment. The results from those regressions are displayed in tables 7 and 8 from which one can see that, with the exception of intercepts, all coefficients are statistically insignificant.

Table 7: Placebo Tests

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y = 1(review is fake)</td>
</tr>
<tr>
<td></td>
<td>(1) (2)</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.142***</td>
</tr>
<tr>
<td></td>
<td>(0.131)</td>
</tr>
<tr>
<td>Time</td>
<td>−0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.0001)</td>
</tr>
<tr>
<td>( \tilde{\mu} )</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>(0.386)</td>
</tr>
<tr>
<td>( \tilde{\mu}^2 )</td>
<td>−0.127</td>
</tr>
<tr>
<td></td>
<td>(0.269)</td>
</tr>
<tr>
<td>peak dummy</td>
<td>−0.010</td>
</tr>
<tr>
<td></td>
<td>(0.017)</td>
</tr>
<tr>
<td>Reliability index from review title</td>
<td>−0.001**</td>
</tr>
<tr>
<td>Reliability index from review text</td>
<td>−0.00002</td>
</tr>
<tr>
<td>Numb. of words</td>
<td>−0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.0003)</td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
</tr>
<tr>
<td>Has Images</td>
<td>−0.020</td>
</tr>
<tr>
<td></td>
<td>(0.027)</td>
</tr>
<tr>
<td>Has videos</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>(0.124)</td>
</tr>
<tr>
<td>Observations</td>
<td>232,176</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>−125,382.700</td>
</tr>
<tr>
<td>Akaike Inf. Crit.</td>
<td>250,787.400</td>
</tr>
</tbody>
</table>

Note: *p<0.1; **p<0.05; ***p<0.01

10The variables “Reliability index from review title” and “Reliability index from review text” were also recomputed based on the new assignment.
On Incentives to Manipulate Online Ratings

<table>
<thead>
<tr>
<th>variable</th>
<th>estimates</th>
<th>p-values</th>
<th>std errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>0.882</td>
<td>0.919</td>
<td>8.655</td>
</tr>
<tr>
<td>$\tilde{\mu}_{i,s,t}$</td>
<td>14.539</td>
<td>0.555</td>
<td>24.639</td>
</tr>
<tr>
<td>$\tilde{\mu}^2_{i,s,t}$</td>
<td>-11.553</td>
<td>0.494</td>
<td>16.908</td>
</tr>
<tr>
<td>time</td>
<td>0.093</td>
<td>0.422</td>
<td>0.116</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable</th>
<th>estimates</th>
<th>p-values</th>
<th>std errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>1.204***</td>
<td>0.000</td>
<td>0.020</td>
</tr>
<tr>
<td>peak dummy</td>
<td>-0.013</td>
<td>0.449</td>
<td>0.018</td>
</tr>
<tr>
<td>Reliability index from review title</td>
<td>-0.063</td>
<td>0.029</td>
<td>0.029</td>
</tr>
<tr>
<td>Reliability index from review text</td>
<td>-0.002</td>
<td>0.218</td>
<td>0.002</td>
</tr>
<tr>
<td>Numb. helpful feedback</td>
<td>-0.000</td>
<td>0.870</td>
<td>0.000</td>
</tr>
<tr>
<td>Verified Purchase</td>
<td>-0.012</td>
<td>0.512</td>
<td>0.019</td>
</tr>
<tr>
<td>has images</td>
<td>-0.016</td>
<td>0.560</td>
<td>0.028</td>
</tr>
<tr>
<td>has videos</td>
<td>0.052</td>
<td>0.674</td>
<td>0.125</td>
</tr>
</tbody>
</table>

Observations: 204,219  
 pseudo $R^2$: 4.95e-05

Table 8: Placebo test for the Logit regression, correcting for classification error.

5 Conclusion

This paper develops a theoretical model in which sellers dynamically choose the effort spent on review manipulation. One of the predictions from the model is that the effort spent on review manipulation tends to be concentrated during the initial periods following a seller’s entrance (or reentrance with a new brandname) into the market, since sellers wish to maximize the impact from each fake review. Another prediction from the model is that the amount of effort dedicated in review fraud does not vary monotonically with the firm’s reputation. Indeed, sellers that currently possess a very good or very bad reputation will usually spend less effort in review manipulation, the intuition being that, for very low levels of reputation, the seller finds it too costly to signal that it is of high quality, whereas a seller that has already accumulated a very good reputation does not need to prove that it is a high quality type. In mathematical terms, very high or very low reputation levels are absorbing states: the closer a seller is to those states, the harder it is to depart from them, which in turn defeats the purpose of trying to influence signals generated from reviews.
Another interesting prediction from the theoretical model is that low quality sellers do not necessarily exert more effort on the fabrication of fake reviews as compared to high quality sellers. Indeed, which type spends most effort on review manipulation depends on the current level of reputation held by the seller. For very low levels of reputation, it is the high quality seller that spends most effort in review manipulation, while the opposite holds for low levels of reputation.

While the benchmark version from the model predicts that in the long run buyers eventually learn the true type form the seller; by allowing sellers to exit and reenter the market with a new name, we observe that low quality sellers tend to resort to this tactic very frequently, thus preventing customers from learning the true type from low quality sellers in the long run. So this result suggests that one way of making reviews more informative in online platforms such as Amazon and TripAdvisor, is by building obstacles that prevent sellers from anonymously selling their products under different account names.

In order to verify some of the model’s predictions empirically, I collected data from products for which sellers were soliciting fake reviews on Amazon. After classifying the reviews posted on those products as fake or real, I then ran a Logit regression to estimate the probability of a review being fake as a function of the seller’s outstanding level of reputation and the time it took for the review to be posted. Consistent with the predictions from the theoretical model, I find that the probability of a review being fake decreases with time, and it varies non-monotonically with the seller’s reputation, where very high or very low reputation levels are associated with a lower probability of the review being fake.

These results have potential practical applications when it comes to fake review detection. Indeed, the performance from machine learning algorithms can potentially be improved with the inclusion of a measure of sellers’ reputation as well as the time it took for the review to be posted since the seller entered into the market, as predictor variables. In research currently underway, I plan to compare the prediction power form neural network algorithms that include and exclude these variables as predictors.
Bibliography


Jindal, Nitin, and Bing Liu. n.d.. “Opinion Spam and Analysis.”

Kaghazgaran, Parisa, James Caverlee, and Majid Alfifi. n.d.. “Behavioral Analysis of Review Fraud: Linking Malicious Crowdsourcing to Amazon and Beyond.”


Mukherjee, Arjun, Bing Liu, and Natalie Glance. n.d.. “Spotting Fake Reviewer Groups in Consumer Reviews.”
Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. n.d.. “Finding Deceptive Opinion Spam by Any Stretch of the Imagination.”


Wang, Jingdong, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen. n.d.. “A Survey on Learning to Hash.”

A Appendix

A.1 Proofs

Proof of proposition 3.1: Let $C(X)$ be the set of real bounded continuous functions with the sup norm defined over $[0, \eta]$. If $V(q, \cdot) \in C(X)$, then applying the following transformation $T$ to $V(q, \cdot)$:

$$T(V(q, \mu)) \equiv \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} \left[ \omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu') \right] dv$$  \hspace{1cm} (10)

s.t. $\mu' = \frac{\mu e^{-\frac{(v-\mu\eta(1,\mu))}{2\sigma^2}}}{\mu e^{-\frac{(v-\mu\eta(1,\mu))}{2\sigma^2}} + (1-\mu)e^{-\frac{(v-\mu\eta(0,\mu))}{2\sigma^2}}}$,  \hspace{1cm} (11)

we have that $T(V(q, \cdot))$ also belongs to $C(X)$. Indeed, because $\tilde{\eta} \in [0, \eta]$, the expression $\lambda\eta^2$ 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 $[0, 1] \times \mathbb{R} \times [0, \eta]$ (where $\mu'$ is obtained by constraint (11)), we have that $X$ is bounded. Therefore, there exists $\underline{x}, \overline{x} \in \mathbb{R}$ such that $\underline{x} \leq X(\mu, v, \eta) \leq \overline{x}$ for any $(\mu, v, \eta) \in [0, 1] \times \mathbb{R} \times [0, \eta]$. This implies that

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

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

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

is continuous\(^{11}\) and the set of feasible choices for \( \eta \), \([0, \eta]\), is compact so that, from the maximum theorem,

\[
T(V(q, \mu)) = \max_{\eta \in [0, \eta]} f(\mu, \eta)
\]
is continuous with respect to \( \mu \).

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

\[\Box\]

### A.2 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!”

\(^{11}\)To show that \( f(\mu, \eta) \) is continuous, define

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

and let \((\mu_n, \eta_n)_{n=1}^{\infty}\) be a generic sequence defined on \([0, 1] \times [0, \eta]\) such that \((\mu_n, \eta_n) \rightarrow (\mu, \eta)\). 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

\[
h_n(v) = g(\mu_n, \eta_n, v), \quad \forall n \in \mathbb{N} \text{ and } \forall v \in \mathbb{R},
\]

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

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

Moreover, because \(|h_n(v)| \leq l(v) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\eta)^2}{2\sigma^2}} \max\{\eta, -\eta\}\) 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

\[
\lim_{n \rightarrow \infty} f(\mu_n, \eta_n, v) = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g(\mu_n, \eta_n, v) \, dv = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} h_n(v) \, dv = \int_{-\infty}^{\infty} h(v) \, dv = f(\mu, \eta, v).
\]
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!
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. Fortunately, computer scientists have devised clever hashing algorithms that allows one to consistently estimate the actual Jaccard index in a way that is computationally feasible. For the purposes of this research I used the widely popular MinHash algorithm. For more details about how the algorithm works, see Wang et al. [n.d.].

A.3 Variables

A.3.1 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

\[12\text{For one of my samples, I would need to make } 4.17 + 10 \text{ of those computations.}\]
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 \( \text{text} = (w_1, w_2, \cdots, w_n) \) 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

\[
P(\text{fake}|\text{text}) > P(\text{real}|\text{text})
\]

\[
\iff P(\text{text}|\text{fake})P(\text{fake}) > P(\text{text}|\text{real})P(\text{real})
\]

\[
\iff \log(P(\text{text}|\text{fake})) + \log(P(\text{fake})) > \log(P(\text{text}|\text{real})) + \log(P(\text{real})). \tag{12}
\]

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 unless one is willing to make restrictions regarding the data generating process (DGP) from review texts, one can not hope to obtain an unbiased and consistent estimates of \( P(\text{text}|\text{fake}) \) and \( P(\text{text}|\text{real}) \).

The Naive Bayes classifier approach simplifies the DGP from review texts by assuming that words are generated randomly and independently. Though this assumption is not very realistic since words need to be put in a logical order in order to convey meaning, it greatly simplifies the process of finding a reliable estimate of \( P(\text{text}|\text{fake}) \). Indeed, letting \( \text{text} = (w_1, w_2, \cdots, w_n) \) 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.\textsuperscript{13}

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 \textsuperscript{12}. 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. The same procedure was used to compute “Reliability index from review title”, but using the contents from the review title as opposed to the review text.

A.3.2 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 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:

$$X_{i,t,p,s} = \beta_0 X_{i-1,t-1,p,s} + \beta_1 t + \sum_{j=1}^{7} \gamma_j D_{j,t} + \alpha_s + \varepsilon_{i,t,p,s},$$

where \(\{D_{j,t}\}_{j=1}^{7}\) are the dummies for the corresponding days of week, $\alpha_s$ is the seller’s fixed effect, 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 4 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 5 star reviews that the seller received on that day would be flagged as anomalous.

\textsuperscript{13} As a standard approach, stop words, such as “I”, “there”, “but”, etc., were removed from the reviews before conducting the Naïve Bayes estimation.