THE MARKET FOR PRODUCT REVIEWS*

Jacob Glazer    Helios Herrera    Motty Perry
University of Warwick    University of Warwick    University of Warwick and Tel Aviv University and CEPR

February 22, 2019

Abstract

We propose a model of product reviews with honest and fake reviews in order to study the value of information provided on platforms like TripAdvisor, Yelp, etc. In every period, a review is posted which is either honest, i.e., it reveals the reviewer’s true experience with the product/service, or fake, i.e., it is fabricated in order to manipulate the public’s beliefs. We establish that the equilibrium is unique and derive a number of robust and interesting results. While fake agents are able to manipulate the public’s beliefs, aggregation of information takes place as long as some of the reviews are honest. We demonstrate that some of the mechanisms currently used to filter out fake reviews can be harmful.

JEL Classification: C72, D82, D83.

Keywords: Fake reviews, Social learning.

*We thank Wioletta Dziuda, Hila Goldenberg, Marina Halac, Sergiu Hart, Navin Kartik, Ilan Kremer, Antonio Nicoló, Debraj Ray and Phil Reny for their insights. An earlier version of this paper circulated under the name "Fake Persuasion" previously.
1 Introduction

The advent of the internet has created many new markets and industries, some of which rely on information provided by market participants. Examples include TripAdvisor, Amazon, Yelp, Airbnb, Booking.com and many other sites where users write reviews on the basis of their experience with the site’s services/products. This type of rating mechanism is growing rapidly and many experts believe that it will soon become the main source of information about a product’s quality in many markets.\footnote{Recently Amazon announced the opening of its new store called Amazon 4-Star. This new concept will stock items that customers have rated as four stars or above, on average. See https://www.businessinsider.com/amazon-opens-new-store-4-star-2018-9}

While many of the reviews on this type of platform are written by benevolent agents who truthfully report their experience, some are written by strategically interested parties whose objective is to influence the readers’ belief (either in favor of or against the product). Assuming the market is rational and aware that some of the reviewers are strategic, three questions arise: \textit{Is complete aggregation of information still possible and if so under what conditions? What mechanisms can be employed to mitigate the effect of fake reviews? What are the consequences of mechanisms currently in use by platforms to achieve this?}

While the internet has been the driving force behind the creation of many new markets that rely on information sharing and the wisdom of the crowd, it is also partly responsible for the decline of other markets, such as that for news, as a result of the public’s loss of trust in news providers. Luca and Zervas (2016) describe the growing importance of rating mechanisms in various markets and the amount of resources devoted to both generating fake reviews and mitigating them.\footnote{Mayzlin, Dover and Chevalier (2014) show that in the case of hotels, there is a consistent}
somewhat surprising that in spite of the growing interest in the reviews industry, to the best of our knowledge there has not been any theoretical model that rigorously analyzes these phenomena.

We propose a formal setup which, in our view, faithfully represents many of the review platforms. In this model, some reviewers are honest, i.e. they truthfully report their experience with the product, while others are fake, i.e., they fabricate reviews in order to manipulate consumers.

In our model, a receiver (a potential future consumer of a certain product) obtains information from multiple senders. Each sender (possibly a past consumer) receives an informative but noisy signal about the state of the world (which can be either high or low) and posts a message that future users - for whom information about the state of the world is valuable - have the possibility of viewing. Each sender is one of three types: an "honest" (non-strategic) sender who truthfully reveals the signal he received while using the product; a "positive fake" sender whose goal is to persuade the receiver that the product is good (the state is high); or a "negative fake" sender whose goal is to persuade the receiver that the product is bad (the state is low). The receiver who observes past messages (or a subset of them) does not know the senders’ types but is rational and understands that some of the messages may have been posted by fake senders (some of whom might have sent more than one message) who choose what information to convey, with the intent of manipulating the receiver’s beliefs for their own benefit.

We first characterize the equilibrium of the dynamic setup and prove that it is 

\textit{the difference in the number of fake reviews between Tripadvisor-type sites where anyone can, in principle, post a review and Expedia-type sites where only customers with a booking can post a review.}
unique and robust. In order to characterize the equilibrium, we derive a technical though very useful result which we refer to as the independence result, according to which each sender’s strategy is independent of the receivers’ (prior) beliefs about the state of the world. The independence result leads to another key observation that, in every period, each fake sender’s message is completely independent of the messages sent in previous periods (by him or other senders). This property is used to show that the strategy of a fake sender is stationary, namely it remains unchanged regardless of changes in the size of the market (i.e. the number of senders and receivers).

One implication of the analysis is a learning result, according to which, in equilibrium valuable learning takes place in every period and the more reviews an agent observes, the more informed she will be.\(^3\) In other words, despite the presence of sophisticated and strategic fake reviewers, belief updating in the correct direction takes place in expectation after every individual review. Our independence/stationarity equilibrium property guarantees that learning does not taper out thus full information aggregation is always achieved in the limit. In sum, while fake agents are able to steer the opinions of all customers that read their reviews in their favored direction, the beliefs of every customer move in the correct direction on average as long as a proportion of reviews are honest, moreover eventually information is fully aggregated and agents perfectly learn the state. However, learning and information aggregation take place at a slower pace: the more fake reviews there are, the slower the learning will be.

The model enables us to derive important implications regarding efforts by a platform to detect fake reviews. We show that while decreasing the proportion of

\(^3\)We will refer throughout to the sender as "he" and to the receiver as "she".
fake reviews improves information aggregation, it comes at a cost: fake reviews, that are now less likely, will in equilibrium become more extreme and therefore more successful. Thus, in calculating how much effort to exert, platforms should take into account this additional cost. Finally, the model offers the additional insight that the strategy often used by platforms to eliminate extreme reviews is never optimal and often harmful. Not only do such policies hide some useful information, they also affect the fake strategies in a way that makes the honest reviews even less informative.

There is a broad theoretical literature that studies models of communication, in which the sender can be either strategic or honest (see for example Chen (2011), Benabou and Laroque (1992), Morgan and Stocken (2003), Dziuda (2011), Kim and Pogach (2014), Lipnowski, Ravid and Shishkin (2018)), Lahr and Winkelmann (2018) and Gratton, Holden, and Kolotilin (2017)), or games in which the sender incurs some cost for cheating (see, for example, Kartik, Ottaviani and Squintani (2007) or Kartik (2009)). Although all of these models are closely related to the current model, our goal is to understand the market for reviews, which has unique characteristics. Therefore, we are able to derive results regarding the properties of the market’s equilibrium, as well as its implications, which do not appear in the literature. Furthermore, we believe that our "many-periods, many-senders and many-receivers" model can serve as a useful platform to further study this industry and to draw conclusions about its performance. There are some models of manipulation/elimination of existing reviews such as Aköz, Arbatl and Çelik (2017) and Smirnov and Strakov (2018), in which the firm does not produce fake reviews but rather changes or eliminates existing ones. Such setups apply only to reviews on a business’ own site, while we focus on mass review platforms such as Yelp or TripAdvisor where existing reviews cannot be altered
by the interested party, but only by the platform itself.

There is also a growing empirical literature that examines the impact of reviews on consumers along several dimensions. Kim and Martin (2018) use online experiments to assess how individuals interpret ratings. Laouenan and Rathelot (2018) use data from an online marketplace of vacation rentals (Airbnb) to measure discrimination against ethnic minority hosts and find that an additional review helps to close the gap in price between minority and majority hosts. This is consistent with the results of our model which predicts that, in expectation, an additional review incrementally corrects for erroneous beliefs. Finally, Mayzlin, Dover and Chevalier (2014) compare fake reviews of hotels on platforms where only consumers can post reviews (such as Expedia) and platforms where anyone can (such as TripAdvisor) and show that fake reviews, whether positive or negative, are much more frequent on the latter platforms and when competition is stronger.

Finally, our analysis is relevant in many other contexts in which the cost of posting and reading reviews is low. Consider, for example, the comments made by individuals after reading an article on one of the many digital media sites. Clearly, while many of these comments are “honest” in the sense that they express the individual’s true opinion on an issue, some of them are made by interested parties. Similarly, the opinions voiced by participants in numerous “forums” that discuss or share information about a particular topic or experience can be “honest” or “strategic”. In spite of the fact that readers of such comments or opinions know that some of them may be “fake,” these platforms are nonetheless thriving. Our findings show that if the participants in such forums are rational, they can still benefit from reading such comments, as long as there are some honest people around.
2 The One-Period Model

We start with the case of two players: a sender ($S$) and a receiver ($R$). The “state of the world” is a random variable, $\theta \in \{0, 1\}$, and is not known to either player. The common prior that $\theta = 1$ is $p$. Conditional on the realization of the state of the world, $\theta$, player $S$ (but not the receiver $R$) receives a signal, $\bar{x}$, which takes a value in $[0, 1]$ according to the density $t_\theta(x)$ and the CDF $T_\theta(x)$. Define $\bar{x}$ to be the unique (neutral) signal for which

$$t_1(\bar{x}) = t_0(\bar{x}).$$

(1)

That is, a signal of $\bar{x}$ does not change the sender’s prior. We make the following assumptions:

A.1 $t_\theta(x)$ is continuous and differentiable, with support $[0, 1]$.

A.2 $\frac{\partial\left[\frac{t_1(x)}{t_0(x)}\right]}{\partial x} > 0$ for all $x \in [0, 1]$.

Assumption A.2 (hereafter referred to as MLRP) captures the idea that the higher the signal is, the more likely it is that $\theta = 1$.

After observing the signal $x$, the sender sends a message $m \in [0, 1]$ to the receiver. The information about the state is valuable only to $R$. Upon receiving a message $m$ from $S$ she uses Bayes’ rule to update her beliefs about the state of the world. Player $S$ is one of two types: Honest ($S_h$) who reports his signal truthfully (i.e., $m = x$), or Fake ($S_f$) who chooses $m$ strategically. Initially, we assume that Fake’s payoff increases with $R'$s posterior, $\hat{p}(m)$, that the state is 1 and that he chooses $m$ to maximize this posterior. A mixed strategy $f$ for Fake is a probability measure over
the set of messages $M = [0, 1]$. Finally, we assume that the receiver does not know the sender’s type and assigns a probability $q$ to the event that $S$ is honest.

## 3 Preliminaries

In this section, we assume that $(p, q, t_0, t_1)$ is given and obtain some preliminary results. Assume a strategy $f$ for Fake such that, for some message $m$, the strategy $f$ is atomless at $m$. Let $\hat{p}(m | f)$ denote the receiver’s posterior given that she observes the message $m$ and believes that Fake uses the strategy $f$ and let:

$$
\hat{P}(m | f) \equiv \frac{\hat{p}(m | f)}{1 - \hat{p}(m | f)} = \frac{p}{(1 - p)} \frac{q t_1(m) + (1 - q) f(m)}{q t_0(m) + (1 - q) f(m)} = \frac{P Q t_1(m) + f(m)}{Q t_0(m) + f(m)}
$$

where $P = p/(1 - p)$ and $Q = q/(1 - q)$.

Thus, $\hat{P}(m | f)$, hereafter referred to as the receiver’s likelihood ratio, can be thought of as Fake’s payoff from sending the message $m$ when the receiver believes that Fake is playing the strategy $f$.

For different values of $m$, note that:

- Outside the support of $f$, i.e., when $f(m) = 0$:

$$
\hat{P}(m | f) = \frac{P t_1(m)}{t_0(m)}
$$

---

4Formally, a mixed strategy $\tilde{f}$ is a mapping from $X = [0, 1]$, the set of signals, to $\Delta$, the set of probability measures over the set of messages $M = [0, 1]$. However, it is easy to see that in equilibrium Fake’s strategy must be independent of the signal that he observes.
• Neutral news always implies no updating:

\[ \hat{P}(\bar{x} | f) = P \]

• Inside the support of \( f \), i.e., when \( f(m) > 0 \):

\[ m \geq \bar{x} \implies \hat{P}(m | f) \leq P \frac{t_1(m)}{t_0(m)} \]

4 One-Period Equilibrium

In this section, we establish that in the one-period game there exists a unique equilibrium and characterize it. Let \( f^* \) denote Fake’s equilibrium strategy. In the following proposition, we provide a set of necessary conditions that \( f^* \) must satisfy:

**Proposition 1** If \( f^* \) is an equilibrium strategy for Fake, then \( f^* \) is atomless and there exists a message \( z \in (\bar{x}, 1) \) such that (i) \( f^*(m) = 0 \) for all \( m \in [0, z] \), (ii) \( f^*(m) > 0 \) for all \( m \in (z, 1] \), and (iii) Fake’s payoff (conditional on sending a message) is \( P \frac{t_2(z)}{t_0(z)} \).

**Proof.** See appendix. ■

Based on the above proposition, we can prove the following result:

**Theorem 1** An equilibrium exists and is unique. Fake’s equilibrium strategy is:

\[ f^*(m) = \begin{cases} 
0 & \text{for: } m \in [0, z] \\
\frac{Q_{t_0(z)}t_1(m) - t_1(z)t_0(m)}{t_1(z) - t_0(z)} & \text{for: } m \in (z, 1] 
\end{cases} \]  

(3)
where \( z \) (hereafter referred to as Fake’s cutoff) is the unique solution to:

\[
\int_z^1 f^*(m) dm = 1. \tag{4}
\]

**Proof.** See appendix. ■

Intuitively, in equilibrium Fake randomizes over an interval \((z, 1]\) in a way that generates the same posterior for the receiver at all \( m \in (z, 1] \), and this posterior is equal to the receiver’s posterior after a message \( m = z \). That is, \( \hat{P}(z \mid f^*) = P_{t_1(z)/t_0(z)} > P \). It is also apparent from (3) that the more likely it is that the sender is Fake (the lower is \( Q \)), the lower is \( z \) and consequently the lower is the posterior Fake can generate. Thus, even though the fake sender is able to "manipulate" the receiver’s beliefs by generating a posterior \( \hat{P}(z \mid f^*) \) which is higher than the prior \( P \), he can only do so to a limited extent and this ability to manipulate decreases with \( Q \). A careful look at \( f^* \) also reveals that \( f^*(m) \) is strictly increasing for \( m \in (z, 1] \).

**Example 1** Consider the following linear example:

\[
t_1(x) = 2x, \quad t_0(x) = 2(1 - x), \quad x \in [0, 1] \tag{5}
\]

Plugging into (3) and solving, the Fake’s equilibrium strategy becomes

\[
f^*(m) = \begin{cases} 
0 & \text{for: } m \in [0, z] \\
Q \frac{m - z}{z - 1/2} & \text{for: } m \in (z, 1] 
\end{cases}
\]

where

\[
z = \frac{1 - \sqrt{1 - q}}{q}
\]
For $q = 3/4$, we have $z = 2/3$. Figure 1 depicts Fake’s strategy:

Moreover, for $p = 1/2$ the posterior for the receiver conditional on receiving the message $m$ is depicted in Figure 2:

Thus, Fake manages to manipulate the receiver by generating a posterior above the prior. When the state is $\theta = 0$, Figure 3 shows the expected message distribution $\tau_0(m) = qt_0(m) + (1 - q)f^*(m)$ (the solid line) compared to the honest-sender distribution $t_0(m)$ (the dashed line):
The following result, hereafter referred to as the *independence* result, follows immediately from (3) and will be essential in the subsequent development of the model.

**Corollary 2 Independence:** Fake’s equilibrium strategy does not depend on the prior $p$.

Thus, Fake’s equilibrium strategy is not affected by the receiver’s prior beliefs about the state of the world. An immediate and interesting implication of Corollary 2 is that even if Fake is uncertain about the receiver’s prior or if he faces a distribution of many receivers with possibly different priors, his equilibrium strategy will still be the one presented in Theorem 1. This result will be particularly useful in the analysis of the multi-period multi-sender game.

We next present a *learning* result which states that as long as there is some strictly positive probability that the sender is honest, the receiver benefits from paying attention to the sender’s messages. In other words, the receiver’s posterior moves in the "correct" direction in expectation. In order to prove this proposition, it will be more convenient to focus on the receiver’s prior $p$ instead of the likelihood ratio $P$ and her posterior probability $\hat{p}(m \mid f^*)$ rather than her posterior likelihood ratio $\hat{P}(m \mid f^*)$. Let $E_0[\hat{p}_{f^*}]$ denote the receiver’s expected posterior probability that the state is 1, given that the true state is $\theta$.

**Proposition 2 Learning:** $E_0[\hat{p}_{f^*}] < p < E_1[\hat{p}_{f^*}]$.

**Proof.** This follows from $E_0[\hat{p}_{f^*}] < E_1[\hat{p}_{f^*}]$ and the fact that their average is $p$. In
fact, the current belief must equal the unconditional expected belief:

\[ p = E[\hat{p}_{f^*}] = (1 - p) E_0[\hat{p}_{f^*}] + p E_1[\hat{p}_{f^*}] \]

Thus, even when the probability of the sender being honest is very small, the receiver’s posterior moves towards the true state in expectation.\(^5\)

At this point, it will be useful to consider the case of two-sided faking.

### 4.1 Two-Sided Faking

Assume now that Fake can be one of two types: Fake-1 \((S_f^1)\) and Fake-0 \((S_f^0)\), where the Fake-1 payoff increases with the receiver’s posterior while that of Fake-0 decreases with the receiver’s posterior. Assume that the sender is honest with probability \(q > 0\), that he is Fake-1 with probability \(q_1 > 0\) and that he is Fake-0 with probability \(q_0 > 0\) where \(q + q_1 + q_0 = 1\).

An analysis similar to the one above shows that the (unique) equilibrium is characterized by two cutoffs: \(z^1 \in (\bar{x}, 1)\) for Fake-1, and \(z^0 \in (0, \bar{x})\) for Fake-0, such that Fake-1’s equilibrium strategy coincides with Fake’s when he is the only fake sender and the probability of the sender being honest is \(1 - q_1\). Fake-0’s equilibrium strategy is the mirror image of Fake’s equilibrium strategy when he is the only fake sender (whose objective is to increase the receiver’s posterior that the state is 1) and the probability of the sender being honest is \(1 - q_0\).

\(^5\)Note that if receivers are not rational and misperceive (e.g. underestimate) \(q\), then the strategy of the fake sender is unchanged (as it just adjusts to the misperceived/wrong \(q\)), but this expected learning result fails as well as the information aggregation that will follow later.
Example 2 Consider the linear case discussed in Example 1. Figure 4 and 5 depict the Fake-0 (the green line) and Fake-1 (the red line) strategies, as well as the receiver’s posterior as a function of the message he receives, for the parameters \((q = \frac{9}{19}, q_1 = \frac{9}{19}, q_0 = \frac{1}{19})\).

Remark 1 Strategic Honest Sender. We now show that the equilibrium characterization described so far survives even if the honest sender is strategic. In other words, consider the case where the “naive” honest sender is replaced by a strategic player whose payoff decreases with the distance of the receiver’s posterior from the sender’s privately-assessed belief, given the signal he received and the prior. Assume, for simplicity, that there is only one sender and that he is either the “honest” strategic sender or the Fake-1 sender. An equilibrium in this modified game is a pair of
strategies, one for the fake sender and one for the “honest” sender.

**Proposition 3** There exists an equilibrium in which the “honest” sender truthfully reports his signal while the fake one uses the strategy $f^*$ described in Theorem 3

**Proof.** Assume that these are indeed the two strategies and hence the receiver’s posterior likelihood ratio is given by $\hat{P}(m \mid f^*)$. We only need to show that the "honest" sender has no incentive to deviate. If the signal received by the sender is below $z$, then by reporting truthfully he reveals that he is the "honest" sender and the receiver’s posterior is equal to that of the sender. If, on the other hand, the signal observed is above $z$, then no matter what message is sent, the receiver’s likelihood ratio will not exceed $\hat{P}(z \mid f^*)$. Thus, the sender cannot benefit from any strategy other than truthfully reporting his signal.

Of course, when the honest sender is strategic there are several other equilibria, as in any cheap talk game, such as the babbling equilibrium. The uniqueness of the equilibrium that characterizes our basic setup is a result of pinning down the honest sender strategy to be fully revealing.$^6$

5 The N-Period Model

In this section, we use the above results to analyze the more general case in which there are many different senders moving at different times and where the fake senders can move more than once. We also allow for multiple receivers who form their beliefs after observing different sets of messages at various points in time.

$^6$It is realistic to assume that the vast majority of disinterested reviewers report their actual experience.
In characterizing the equilibrium of the general model, we rely heavily on the independence result presented in Corollary 2. One implication of this result is that a fake sender’s action in a given period is not affected by previous messages (sent either by himself or by other players) and will not affect his or other senders’ actions in future periods. Before considering the general N-period model with many types of senders and many receivers, it will be helpful to start with the case of one sender.

5.1 Special Case: one receiver, one fake sender and one honest sender

There are $N$ periods. The state of the world (which is the same throughout the game) is $\theta \in \{0, 1\}$. There is one honest agent and one Fake-1 agent, but in every period $n \in N$ only one of them is chosen to send a message. In particular, in every period $n \in N$, and independently of the history, the honest sender is chosen with probability $q$ to receive and truthfully report a signal $\tilde{x}$, which takes values in $[0, 1]$ according to the density $t_{\theta}(x)$, $\theta \in \{0, 1\}$ while the fake sender is chosen with probability $(1 - q)$ to report a signal. Assume (again for now) that there is only one receiver who forms her posterior about the state after observing the $N$ messages. In that case, fake sender’s objective is to maximize the receiver’s posterior belief, after $N$ periods, that $\theta = 1$. Assume for now that the history of messages is fully observed by the senders and the receiver. Then, a strategy for the fake sender is a function $\sigma_N = \{f_n\}_{n \in N}$ where $f_n$ is a density function and $f_n : [0, 1]^n \to R_+$. Thus, $\sigma_N$ specifies the mixed action of the fake sender for each period $n \in N$ if he is chosen to send a message.
Let $f^*$ be the (unique) fake sender’s equilibrium strategy when $N = 1$, as defined in Theorem 1. Let $\sigma^*_N = \{f^*_n\}_{n \in N}$ denote an equilibrium strategy for the fake sender in the $N$-period game. The following proposition states that the unique equilibrium strategy for the fake sender is to employ his one-period equilibrium strategy in every period in which he is chosen to send a message.

**Proposition 4** $\sigma^*_N$ is unique and $f^*_n \equiv f^*$ for all $n = 1, \ldots, N$.

**Proof.** See appendix. ■

### 5.2 General Case

Consider the following extension of the special $N$-period model analyzed above. In every period $n$, a new receiver forms her posterior about the state of the world on the basis of a subset of the history of messages up to period $n$, where the subset is randomly drawn according to some commonly known distribution. There is a set of $L = L_0 + L_1 + L_h$ senders where $L_0$ is the number of fake senders whose objective is to minimize the receiver’s posterior that the state is one, $L_1$ is the number of fake senders whose objective is to maximize the receiver’s posterior that the state is one, and $L_h$ is the number of honest senders who in every period observe a signal according to $t_\theta$.

In every period, sender $l$ is selected with probability $q_l$, $l = 1, \ldots, L$ where $q_l \geq 0$ and $\sum_{l=1}^L q_l = 1$, to be the one to send a message in that period. For $i \in \{0, 1, h\}$ let $\bar{q}^i = \sum_{l \in L_i} q_l$ and observe that $q_0 + q_1 + q_h = 1$. If $l$ is honest, he truthfully reports his signal; otherwise he reports strategically. A strategy for a fake sender $l$ specifies his move in every period $n$ if he is selected to move in that period and given the history
of previous messages he can observe. That is, $\sigma_{N}^{l} = \{f_{n}^{l}\}_{n \in N}$ where $f_{n}^{l}$ is a density function and $f_{n}^{l} : S_{n-1} \times [0,1] \to R_{+}$ where $S_{n-1}$ is the set of all possible partial histories up to (and including) period $n - 1$. Let $f_{q_{0}}^{*}$ and $f_{q_{1}}^{*}$ be, respectively, Fake-0 and Fake-1's equilibrium strategies in the two-sided one-period model, where the sender is of type Fake-$\theta$ with probability $q_{\theta}$ and honest with probability $1 - q_{0} - q_{1}$. Let $\sigma_{N}^{ls}$ be $l$'s equilibrium strategy in the $N$-period model. We can now state the following proposition, which can be proved with a straightforward application of the argument in the proof of Proposition 4 (and therefore the proof is omitted).

**Proposition 5** Let $l$ be a fake sender. Then $\sigma_{N}^{ls}$ is unique and for all $n = 1,...,N$,

$$f_{n}^{ls} \equiv f_{q_{\theta}}^{*} \text{ if } l \text{ is type Fake-$\theta$.}$$

In order to put Proposition 5 into context, consider TripAdvisor and a particular hotel owner. Sitting down to send a fake review, he is aware of some of the history (though he might not be aware of messages that are being written "simultaneously" with his own and are about to be posted) and understands that his message is going to be read by future receivers who will be influenced by other senders, whether honest or fake. Proposition 5 fully characterizes his strategy.

### 5.3 How much do we learn from a review?

It follows from Proposition 2 and 5 that in the general $N$-period model, the receivers’ beliefs shift over time (in expectation) towards the true state of the world. Fake reviews, however, slow down the learning process. In other words, fake senders manage to steer beliefs their way to some extent every time they are chosen to send a message, i.e., the smaller $q$ is, the less the expected belief $E_{\theta}[\hat{p}_{f_{l}}]$ will shift in the right
direction. While the proportion of fake reviews slow learning on average, asymptotic convergence to the truth is preserved as long as the proportion of true reviews is positive. Both these statements are summarized in the following result:

**Proposition 6** *Fake reviews slow learning:*

\[
\frac{dE_1[\hat{p}_{f^*}]}{dq} > 0, \quad \frac{dE_0[\hat{p}_{f^*}]}{dq} < 0.
\]

For any \( q > 0 \), the belief converges to the true state \( \theta_T \):

\[
\lim_{N \to \infty} \hat{p}_{f^*}(m_1, ..., m_N) = \theta_T, \quad \text{almost surely.}
\]

**Proof.** See appendix.\(^7\)

Note that positive learning in every period does not guarantee per se full learning in the limit, i.e. full information aggregation, because learning can potentially taper out gradually as the belief becomes more certain.\(^8\) The latter does not happen with fake reviews however: full information aggregation is guaranteed by the stationarity of the fake agents’ persuasion strategies, namely by the independence result.

### 6 Implications for Platforms

As discussed in the introduction, review platforms invest a large amount of resources and employ various mechanisms in order to fight fake reviews. Our simple model can be used to evaluate such strategies. In what follows, we discuss two of

---

\(^7\)Ignoring the abuse of notation, we refer to \( f^* \) as the \( N \)-period fake strategy.

\(^8\)This is typical for instance in models of observational learning, such as herding models.
them and point out some of their unexpected consequences.

6.1 Reducing the frequency of fake reviews

Clearly, if a platform invests more resources in detecting fake reviews and is successful in doing so, the chance that any given review is honest, \( q \), will increase. As shown above, the benefit from increasing \( q \) (net of the direct cost of doing so) is positive since the belief, after any given review, will in expectation move faster toward the correct state. However the benefit should be looked at more closely since this seemingly obvious comparative statics actually involves a key trade-off: on the one hand increasing \( q \) lowers the probability that a review is fake (the direct effect), while on the other hand, it also leads to more aggressive fake reviews in equilibrium (the indirect effect). The larger \( q \) is, the higher will be \( z \) (Fake’s cutoff) and the induced posterior. Thus, efforts to detect fake reviews come with a price, namely that the fake reviews that do manage to sneak in will be more extreme and more persuasive. When choosing a mechanism to increase \( q \), platforms should take this indirect cost into account.

6.2 Purging Extreme Reviews

The model predicts that in equilibrium fake messages will be concentrated at either extremes. This perhaps explains the policy - often used by platforms - of purging extreme reviews.\(^9\) We now show that if fake reviewers are sophisticated and anticipate this, then such a policy is counterproductive. Namely, not only will

\(^9\)In practice, what the platforms often do is erase reviews that use extreme language. In our model, this translates into extreme messages, i.e., close to zero or one.
the proportion of honest reviews that a given reviewer can see be reduced, but also the distribution of fake messages will be shifted more toward the center, making informative messages less so.

To see this, notice first that even though, in equilibrium, extreme messages are more likely to be fake, the receiver - when interpreting such messages- already internalizes this fact. Therefore, even in the unlikely case that players are not aware of the platform’s policy (of eliminating extreme messages) and follow their original strategy, this purging tactic backfires since it leads to less learning. Moreover, since fake reviewers are strategic and aware of the purging policy they will make sure to send less extreme messages, leaving only honest messages to be purged.

Formally, consider first the general $N$-period model discussed in Section 5.2 with the modification that if the message $m$ is above a certain cutoff $L$, then the message is purged. One of the ways to model this modification (the "no-replacement model") is to assume that if a message is purged then no message appears in that same period. In such a case, a rational receiver in period $n$ counts the number of reviews he observes and infers from it the number of extreme messages that were purged because they were above the cutoff $L$. We discuss this case below and establish that purging extreme reviews is never optimal. An alternative method (the "with-replacement model") to model the modification is to assume that whenever a message is purged a new sender is selected to send a message in that period, where with probability $q$ the new sender is honest and with probability $1 - q$ he is fake. In this alternative model, the number of reviews a receiver observes does not reveal anything about the number of messages that were purged, and one could argue that it is a better fit to the real world. We discuss this case as well and, as above, show that the policy of purging
extreme reviews is counterproductive.

It follows from Corollary 1 that in order to analyze the effect of the purging policy it is sufficient to study its effect in the one-period model. We can use no-replacement and one-period one-sided fake sender model analyzed in Section 4 with the modification that if the message \( m \) is above a certain cutoff \( L \), then it is purged and no message is revealed. Let \( z \) denote Fake’s cutoff in the original model when \( L \) was not imposed, and assume first that \( L \) is above \( z \). It is easy to see that in such a case, Fake’s original strategy \( f^* \) is still optimal. Indeed any strategy that coincides with the original strategy for all messages below \( L \) and randomizes arbitrarily above \( L \) is also optimal and yields the same outcome. The receiver’s payoff in all of these equilibria is the same as that in the game without the imposed \( L \). Thus, imposing \( L > z \) does not have any effect on the receiver’s payoff. Next, consider the case in which \( L < z \). Fake’s strategy will be an equilibrium strategy if and only if it assigns all the weight to messages above \( L \). The receiver’s payoff is the same in all of these equilibria and, since it involves less learning, the receiver is worse off in comparison to the equilibrium without the imposed \( L \).\(^{10}\)

Now consider the one-period two-sided model presented in Section 4.1. Recall that \( z^1 \) and \( z^0 \) are the two strategic cutoffs when extreme reviews are not eliminated. Let \( L^1 \) and \( L^0 \) denote the two imposed cutoffs. For simplicity, assume that \( z^1 < L^1 \) and \( z^0 > L^0 \). In such a case, when a message is not revealed the receiver must conclude that either \( m > L^1 \) or \( m < L^0 \). In the generic case, when the model is not fully symmetric

---

\(^{10}\)If the receiver is not aware of the platform policy of erasing extreme reviews and keeps his prior in the case that no review is revealed, the situation is even worse. In such a case, Fake will shift the support of the distribution of his messages to be within the "allowed" interval and the share of fake messages will increase since only honest messages will be purged.
and no message appears, the receiver’s posterior must move in one direction or the other and assume without loss that it increases. In such a case, Fake-1’s strategy is still an equilibrium strategy, but Fake-0’s strategy will move to the right with a support $[L^0, z']$ where $z' > z^0$ and payoff $P = pt^0(z')/t^1(z')$. Thus, messages in $[z, z']$ are now less informative than previously.

Consider now the model with replacement in which whenever a message is purged a new sender is selected from the same distribution to send a message. Once again, it suffices to study the one-period two-sided model. Assume that $L_0 \leq z_0$ and $L_1 \geq z_1$. It can immediately be seen that in this model a fake sender will never send a message that will be purged (below $L_0$ or above $L_1$). Consequently, only messages that were sent by honest senders will be purged, and therefore the resulting share of honest reviews, $q$, will now be lower. Moreover, it is now easy to verify that Fake’s strategy (the interval on which he mixes) is shifted towards the center, thus contaminating even the more moderate reviews, which were known to be truthful. In sum, purging extreme reviews is a counterproductive policy: while it might mitigate the persuasion of any single fake review, it totally backfires by increasing the share of fake reviews, which, in turn, crucially slows down the learning and thus reduces welfare.

In summary, the key drawback of any systematic review-purging policy is that as long as honest reviewers continue to be disinterested and non-strategic and as long as fake reviewers continue to adapt their persuasion strategy to outguess the review-purging algorithm, then any attempt to eliminate or reduce fake reviews will backfire since proportionally more honest reviews than fake reviews will be purged in the process. Purging algorithms can work only if (some) fake reviewers are also naive and/or cannot adapt their strategy.
7 Conclusion

Online reviews are ubiquitous and have become an essential part of a consumer’s everyday decisions. However, their credibility has been undermined by the incentives of the reviewed businesses (or their competitors) to manipulate them. An extra star on a restaurant’s Yelp rating can increase revenues by 5-9% (see Luca and Zervas (2016)). Cases of businesses caught hiring reviewers or individuals offering fake online review services abound in the popular press. Fake reviews are not only positive: businesses also plant unfavorable fake reviews of competitors, especially in highly competitive markets. Yelp, TripAdvisor and Angie’s List are billion-dollar businesses dedicated to offering online reviews for nearly every existing product and service. They combat fake reviews by filtering them with word algorithms, although the rate of false positives and negatives remains high since fakes are improving their methods. The extent of review manipulation, while hard to measure precisely, can be inferred indirectly. For instance, Yelp, which alone contains over 80 million reviews, filters out 16% of restaurant reviews, and has even created a special list of “recommended reviews” by removing the 30% of reviews that look suspicious (see Luca and Zervas (2016) and the Economist, Oct. 22 2015).

The aim of the proposed model is to accurately represent the product review market described above. In this market, there is a plethora of reviews for any given product. Any customer can read only some of them and is aware that they may be true or fake (whether positive or negative). Fake reviewers know they can write more than one review and that some or none of them might be read by any given consumer. Consumers may have different initial beliefs or develop them after reading reviews.
Despite the generality of the setup, the characterization of the fake reviewer strategy is unique and does not depend on the aforementioned parameters, but only on the true/fake review shares. This allows us to deliver general and robust predictions of information transmission in the presence of true and fake reviews. Thus, fake reviewers manage to persuade and shift beliefs individually, but only enough to slow down, but not derail, correct long-run information aggregation. Reading reviews is obviously costly and time-consuming and therefore the question arises of whether it is worthwhile. The answer depends on, among other things, the amount of information obtained from reading a finite number of reviews, which in turn depends only on how easy and cheap it is to fabricate fake reviews. In this sense, efforts to reduce or limit the proliferation of fake reviews are worthwhile.

Finally, we believe that the model can also serve as a platform to study other scenarios or interactions in which fake information transmission may take place. Consider, for example, a news outlet (such as a newspaper or radio station) or a blog, whose audience is unsure whether it is biased. In such a case, the receiver may not be sure whether the sender is honest or fake but she is sure that it is the same sender in all periods. Therefore, when the fake sender is considering which message to send, he knows it will affect not only the receiver’s beliefs about the state of the world, but also about the sender’s type, which in turn influences the receiver’s interpretation of future messages (since she now knows with certainty that all messages originate from the same sender). This model was not dealt with here since it is less able to capture the real-world situations we have in mind. Nonetheless, some preliminary analysis indicates that many of the important features of our model’s equilibrium are preserved. In particular, the equilibrium will be unique and in it the fake sender
randomizes over an interval. However, the fake sender’s strategy (i.e. his cutoff) will depend on \( N \), the number of messages he can send. This modification and others are left for future research.

8 References


   Working paper.


9 Appendix

**Proof of Proposition 1.** The proof is established by proving a series of claims. The first simply states that in equilibrium Fake never assigns a strictly positive probability to any message \( m \).

**Claim 1** Fake’s equilibrium strategy is atomless.

**Proof.** Assume that there exists an \( m \) to which \( f^* \) assigns a strictly positive probability. Given our assumption that \( t_\theta(x) \) is atomless, the receiver’s likelihood ratio at \( m \), \( \hat{P}(m \mid f^*) \), must be \( P \). This is because when the receiver observes the message \( m \), she must conclude that the sender is fake and therefore must stick to her prior. Since the number of messages with a strictly positive mass in any probability distribution is countable, there must be a message \( m' > \bar{x} \) such that \( f^*(m') \) is atomless and \( \hat{P}(m' \mid f^*) > P \). Thus, sending the message \( m' \) is strictly better for Fake than sending the message \( m \), a contradiction.
We can assume hereafter that if $f^*$ is an equilibrium strategy then it is atomless and we let $\hat{P}(m \mid f^*)$ denote Fake’s equilibrium payoff from sending the message $m$.

**Claim 2** $f^*(1) > 0$.

**Proof.** Assume, by contradiction, that $f^*(1) = 0$ and thus $\hat{P}(1 \mid f^*) = P_{t_1(1)}^{t_0(1)} > \hat{P}(m \mid f^*)$ for all $m < 1$. The last inequality follows from the assumed MLRP of $t_0(\cdot)$. Thus, deviating to $m = 1$ is a profitable deviation for Fake. ■

It follows from Claim 2 that if $f^*(m) > 0$ for some $m \neq 1$, then $\hat{P}(m \mid f^*) = \hat{P}(1 \mid f^*)$. We will use this fact in the following claims.

**Claim 3** For all $m \leq \bar{x}$ (i.e., $\frac{t_1(m)}{t_0(m)} \leq 1$), $f^*(m) = 0$.

**Proof.** Assume, to the contrary, that for some $m' \leq \bar{x}$, $f^*(m') > 0$. Then, it follows from Claim 2 that $\hat{P}(m' \mid f^*) = \hat{P}(1 \mid f^*)$ or

$$P_{t_1(m')}f^*(m') + \frac{Qt_1(m') + f^*(m')}{Qt_0(m') + f^*(m')} = P_{t_0(1)}f^*(1)$$

which is in contradiction to $t_1(m') \leq t_0(m')$ and $t_1(1) > t_0(1)$. ■

In the following claim, we show that in equilibrium Fake mixes over an interval $(z, 1]$ for some $z > \bar{x}$. That is, there exists a $z > \bar{x}$ such that $f^*(m) > 0$ if $m \in (z, 1]$ and $f^*(m) = 0$ if $m \in [0, z)$. We later establish that $f^*(z) = 0$.

**Claim 4** If for some $m' < 1$, $f^*(m') > 0$, then for every $m''$ such that $m' < m'' < 1$, it must be that $f^*(m'') > 0$.

**Proof.** Assume that there exist $\bar{x} < m' < 1$ and $m'' \in (m', 1)$ such that $f^*(m') > 0$ and $f^*(m'') = 0$. It follows that $\hat{P}(m'' \mid f^*) = P_{t_1(m'')}^{t_0(m'')} > P_{Q_{t_1(m')}f^*(m') + f^*(m')}^{Q_{t_0(m')}f^*(m')} = \hat{P}(m' \mid f^*)$, 29
where the inequality is implied by \( m'' > m' > \bar{x} \) and MLRP. Thus, deviating from \( m' \) to \( m'' \) is profitable for Fake. ■

**Claim 5**

(i) There exists a message \( z \in (\bar{x}, 1) \) such that \( f^*(m) = 0 \) for all \( m \leq z \) and \( f^*(m) > 0 \) for all \( m > z \) and (ii) Fake’s equilibrium payoff is \( P_{t_0(z)}^{t_1(z)} \).

**Proof.** We start by proving (i). From the four claims above we know that there exists a message \( z \in (\bar{x}, 1) \) such that \( f(m) = 0 \) for all \( m < z \) and \( f(m) > 0 \) for all \( m > z \). We now establish that \( f^*(z) = 0 \). Assume that \( f^*(z) > 0 \) and therefore by the equilibrium condition, for all \( m \in [z, 1] \),

\[
\hat{P}(m | f^*) = \hat{P}(z | f^*) = P_{t_0(z)}^{Qt_1(z) + f^*(z)} < P_{t_0(z)}^{Qt_1(z)}.
\]

From Claim 4, it follows that for all \( m \in [0, z) \), \( \hat{P}(m | f^*) = P_{t_0(m)}^{t_1(m)} \). By Assumption A.1, there exists \( \epsilon > 0 \) such that for \( m' \in [z - \epsilon, z) \),

\[
P_{t_0(m')}^{t_1(m')} > \hat{P}(z | f^*) = P_{t_0(z)}^{Qt_1(z) + f^*(z)}
\]

and a deviation from the message \( z \) to \( m' \) is profitable, a contradiction.

(ii) First observe that, in equilibrium, \( \hat{P}(m | f^*) \geq P_{t_0(z)}^{t_1(z)} \), for all \( m > z \), since by part (i) above \( P_{t_0(z)}^{t_1(z)} \) is the payoff Fake could obtain by sending the message \( z \). We will now show that for all \( m > z \), \( \hat{P}(m | f^*) \) cannot be strictly greater than \( P_{t_0(z)}^{t_1(z)} \). Since for all \( m > z \), \( f^*(m) > 0 \) it must be that

\[
\hat{P}(m | f^*) = P_{t_0(m)}^{Qt_1(m) + f(m)} < P_{t_0(m)}^{t_1(m)}
\]
and since, by A.1, \( \lim_{m \to z} \frac{P_{t_1}^{t_1}(m')}{t_0(m')} = P_{t_1}^{t_1}(z) \), it follows that for \( m \) close enough to \( z \),

\[
\hat{P}(m \mid f^*) \leq \hat{P}(z \mid f^*) = \frac{P_{t_1}(z)}{t_0(z)}.
\]

We conclude that for all \( m \geq z \), \( \hat{P}(m \mid f^*) = \frac{P_{t_1}(z)}{t_0(z)} \). ■

**Proof of Theorem 1.** Recall that if \( f^* \) is an equilibrium strategy then by Proposition 1 it must be that \( f(m) = 0 \) for \( m \in [0, z] \) and for all \( m \in (z, 1] \),

\[
\hat{P}(m \mid f^*) = P_{Qt_0(m)+f^*(m)}^{Qt_1(m)+f^*(m)} = P_{t_0(z)}^{t_1(z)}. \quad \text{We can therefore solve for } f^* \text{ in order to derive the functional form presented in the Theorem.}
\]

To prove existence and uniqueness, it is left to show that there exists a unique \( z \in (\bar{x}, 1] \) for which \( f^*(m) \geq 0 \) for all \( m \in [0, 1] \) and \( \int_z^1 f^*(m) dm = 1 \). Choose some \( \mu \in (\bar{x}, 1) \) and for any \( m \in [\mu, 1] \) let

\[
\psi(m \mid \mu) = \frac{Q[t_0(\mu) t_1(m) - t_1(\mu) t_0(m)]}{t_1(\mu) - t_0(\mu)}.
\]

By A.2 it must be that for any \( m \in (\mu, 1] \), \( \psi(m \mid \mu) \) is positive and strictly decreasing with \( \mu \). Define \( \Psi(\mu) = \int_{\mu}^1 \psi(m \mid \mu) dm \) and observe that \( \Psi(\mu) \) is strictly decreasing with \( \mu \) and is unbounded as \( \mu \) approaches \( \bar{x} \) from above. Since \( \Psi(1) = 0 \), we conclude that there exits a unique \( z \) such that \( \Psi(z) = 1 \). ■

**Proof of Proposition 4.** Define \( \hat{P}(m_1, \ldots, m_n, \sigma_N^*) \) to be the likelihood ratio of the receiver’s posterior given a vector of messages \((m_1, \ldots, m_n)\) and the fake sender’s strategy \( \sigma_N^* \). That is

\[
\hat{P}(m_1, \ldots, m_n, \sigma_N^*) =
\]

\[
= \frac{Pr(\theta = 1|(m_1, \ldots, m_n), \sigma_N^*)}{Pr(\theta = 0|(m_1, \ldots, m_n), \sigma_N^*)} = \frac{P_{\theta = 1}(m_1, \ldots, m_n)}{P_{\theta = 0}(m_1, \ldots, m_n)}.
\]

(7)
Consider the last period $n = N$. Using our assumption that in every period the type of the sender is independently drawn, we can rewrite 7 as

$$\hat{P}(m_1, \ldots, m_N, \sigma_N^*) =$$

$$= \frac{\Pr(m_N | \theta = 1, \sigma_N^*, (m_1, \ldots, m_{N-1})) \Pr(m_1, \ldots, m_{N-1} | \theta = 1, \sigma_N^*)}{\Pr(m_N | \theta = 0, \sigma_N^*, (m_1, \ldots, m_{N-1})) \Pr(m_1, \ldots, m_{N-1} | \theta = 0, \sigma_N^*)}$$

$$= \frac{qt_1(m_N) + (1 - q)f_N^* (m_1, \ldots, m_N) \Pr(m_1, \ldots, m_{N-1} | \theta = 1, \sigma_N^*)}{qt_0(m_N) + (1 - q)f_N^* (m_1, \ldots, m_N) \Pr(m_1, \ldots, m_{N-1} | \theta = 0, \sigma_N^*)}.$$  

(8)

With (8) in mind, we can essentially repeat the logic of the one-period proof. First, observe that in every equilibrium it must be that, for all $(m_1, \ldots, m_{N-1}), f_N^* (m_1, \ldots, m_{N-1}, 1) > 0$. Also for any message $\tilde{m}_N \neq 1$ such that $f_N^* (m_1, \ldots, \tilde{m}_N) > 0$ it must be that

$$\hat{P}(m_1, \ldots, m_{N-1}, \tilde{m}_N, \sigma_N^*) = \hat{P}(m_1, \ldots, m_{N-1}, 1, \sigma_N^*)$$

which implies that

$$\frac{qt_1(\tilde{m}_N) + (1 - q)f_N^* (m_1, \ldots, m_{N-1}, \tilde{m}_N)}{qt_0(\tilde{m}_N) + (1 - q)f_N^* (m_1, \ldots, m_{N-1}, \tilde{m}_N)} = \frac{qt_1(1) + (1 - q)f_N^* (m_1, \ldots, m_{N-1}, 1)}{qt_0(1) + (1 - q)f_N^* (m_1, \ldots, m_{N-1}, 1)}.$$  

(9)

Note the similarity between equation (9) in Section 2 and equation (6) above. By following step by step the arguments in the proof of Theorem 3 one can easily establish that

$$f_N^* (m_1, \ldots, m_{N-1}, m_N) \equiv f^*(m_N)$$

for all $(m_1, \ldots, m_{N-1})$. That is, the fake sender’s equilibrium strategy in the last period is independent of the history and coincides with his equilibrium strategy in the one-period model. One can then proceed by moving one step backward: Given
that the strategy in period $N$ is independent of the history it is easy to show that the fake sender’s strategy in period $N-1$ also coincides with his one-period equilibrium strategy $f^*$. A similar argument can be made for every period. ■

**Proof of Proposition 6.**

We will prove the first by showing that

$$\frac{dE_1[\hat{p}_{f*}]}{dq} > \frac{dE_0[\hat{p}_{f*}]}{dq}$$

It is easy to see, since the terms in $f$ cancel out, that $(E_1[\hat{p}_{f*}] - E_0[\hat{p}_{f*}]) =$

$$= q \int_0^1 \hat{p}_{f*}(m) (t_1(m) - t_0(m)) dm = q \left( \int_0^z \hat{p}_{f*}(m) (t_1(m) - t_0(m)) dm + \int_z^1 \hat{p}_{f*}(m) (t_1(m) - t_0(m)) dm \right)$$

Since $\frac{d\hat{p}_{f*}}{dq} = 0$ for $m \leq z$, using Leibnitz rule, we obtain:

$$\frac{d}{dq} (E_1[\hat{p}_{f*}] - E_0[\hat{p}_{f*}]) = (E[\hat{p}_{f*} | t_1] - E[\hat{p}_{f*} | t_0]) + q \int_z^1 \frac{d\hat{p}_{f*}}{dq} (t_1(m) - t_0(m)) dm > 0$$

where the first term is positive due to MLRP and the last term is positive since $\frac{d\hat{p}_{f*}}{dq} > 0$ and $t_1(m) > t_0(m)$ for all $m > z$. As for the claim on asymptotic learning,\footnote{We thank Phil Reny for providing key insights on this part of the proof.} we can write the likelihood ratio $\hat{P}_{f*} = \frac{\hat{p}_{f*}}{1-\hat{p}_{f*}}$ of the belief after $N$ messages as the
following product:

\[ \hat{P}_{f^*} (m_1, ... m_N) = P (\rho (m_1) \rho (m_2) ... \rho (m_N)) \]

where we defined:

\[ \rho (m_k) := \frac{\tau_1 (m_k)}{\tau_0 (m_k)} = \frac{q t_1 (m_k) + (1 - q) f^* (m_k)}{q t_0 (m_k) + (1 - q) f^* (m_k)} \]

It suffices to show that if \( \theta_T = 0 \), then:

\[ \lim_{N \to \infty} \hat{P}_{f^*} (m_1, ... m_N) = +\infty, \text{ almost surely.} \]

or, equivalently that:

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \log \rho (m_k) = +\infty, \text{ almost surely.} \]

By the strong law of large numbers, and given that the true state is \( \theta_T = 0 \),

\[ \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \log \rho (m_k) = E_0 (\log \rho (m)) \]

Thus, we just need to show that the above quantity is strictly negative. This follows from Jensen’s inequality, as:

\[ E_0 (\log \rho (m)) < \log (E_0 (\rho (m))) = \log \left( \int \frac{\tau_1 (m)}{\tau_0 (m)} \tau_0 (m) dm \right) = \log 1 = 0 \]