Fake Reviews*

Jacob Glazer
University of Warwick
and Tel Aviv University

Helios Herrera
University of Warwick
and CEPR

Motty Perry
University of Warwick

October 5, 2020

Abstract

We propose a model of product reviews in which some are genuine and some are fake in order to shed light on the value of information provided on platforms like TripAdvisor, Yelp, etc. In every period, a review is posted which is either genuine or fake. We characterize the equilibrium of the dynamic model and prove that it is unique. In equilibrium, valuable learning takes place in every period. Fake reviews, however, do slow down the learning process. It is established that any attempt by the platform to manipulate the reviews is counterproductive. **JEL Classification:** C72, D82, D83.

Keywords: Fake Online Reviews, Fake News, Information Aggregation.

*We would like to thank the editor of this journal and the referees for very helpful suggestions and comments which led us to rethink and change the focus of this paper.

1 Introduction

The internet has created many new markets and industries that rely on the wisdom of the crowd, namely on information provided by market participants. Online reviews are a major feature of this trend. Yelp, TripAdvisor and Angie's List are billion-dollar platforms dedicated to offering online reviews of nearly every existing product and service. An extra star on a restaurant's Yelp rating can increase revenues by 5-9% (see Luca and Zervas (2016)). While online reviews are ubiquitous and have become an essential part of a consumer's everyday decision making, their credibility has been undermined by the incentives of reviewed businesses (or their competitors) to manipulate them. Cases of businesses caught hiring fake reviewers or individuals offering fake online review services abound in the popular press. The extent of review manipulation, while is 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. Fake reviews can also be negative, in the sense that businesses plant unfavorable fake reviews of competitors, especially in highly competitive markets.

Given that some reviews are written by benevolent agents who truthfully report their experience while the others are written by strategically interested parties whose objective is to falsely manipulate the readers' belief, a natural question arises: Should review platforms (such as Yelp) simply report all reviews (knowing that some of them may be fake) or could they apply a filtering mechanism to reduce the fake reviewers' influence?

In this paper, a surprisingly strong result is obtained which we refer to as *full transparency*: a review platform cannot do better than simply reporting all messages. More specifically, when a platform reports all messages as is, a learning process takes place and any attempt by the platform to manipulate the reviews (e.g. by blocking "extreme" reviews

or by pooling them), will make future users of reviews worse off in expectation.

In the model, receivers, namely potential future consumers of a particular product, obtain review information from multiple senders by means of a platform. Senders are either "honest" thus truthfully reveal the (noisy) signal they received while using the product, or are "fake" in which case they wish to persuade the receiver that the product is good (in the case of a "positive fake") or bad (in the case of a "negative fake"). The platform is uninformed about the state of the world and can commit to a reporting mechanism that maps the senders' reviews to a report to be sent to the receivers. The platform's objective is to maximize the receivers' welfare. We say that the platform is nonstrategic if it simply reports all reviews to the receivers.

We characterize the equilibrium of the dynamic setup and prove that it is unique. It is shown that the platform cannot do better than to simply report all the reviews. Note that the only way that manipulating reviews could possibly benefit the receiver is if the platform can somehow affect the fake sender's strategy in a way that will make the messages less harmful. However, we show that any attempt by the platform to do so will influence the fake sender's strategy in a way that makes the messages sent by an honest sender less informative. By the same reasoning, we also show that if the "honest" sender could behave strategically (in order to maximize the receivers' welfare) then multiple equilibria would exist; however, the best equilibrium for the receivers is achieved when the honest sender is nonstrategic.

The analysis proceeds as follows: Section 2 reviews the literature. Section 3 presents a one-period model with a nonstrategic platform, a nonstrategic honest sender and a fake sender. The one-period equilibrium is presented in Section 4. In Section 5, we relax the assumption that the platform is nonstrategic. Section 6 extends the model to the case where there are many periods, many senders and many receivers. Section 7 concludes.

2 Related Literature

Despite the pervasive use of online reviews and the extent of review manipulation, until recently there has not been any theoretical work explicitly studying optimal reporting mechanisms used by such platforms. Nonetheless, there is a vast related literature.

In a recent (and independently written) working paper, Lahr and Winkelman (2019) also study a model with multiple senders who can either share the same preferences as the receiver or prefer that the receiver always take a particular action. They show, as does our model, that in equilibrium the fake senders ("partisans" in their analysis) randomize over some messages and honest senders ("advisors" in their analysis) simply report the truth even if strategic. In contrast to the current paper, they do not consider the design of the optimal reporting mechanism.

The literature includes models of manipulation/elimination of *existing* reviews such as Aköz, Arbatl and Çelik (2018) and Smirnov and Strakov (2018), in which the firm does not produce fake reviews but rather alters 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 an interested party, but only by the platform itself.

There is an extensive theoretical literature that looks at static models of communication, in which the sender can be either strategic or honest (see, for example, Benabou and Laroque (1992), Morgan and Stocken (2003), Dziuda (2011), Chen (2011), Avery and Meyer (2012), Kim and Pogach (2014), Gratton, Holden, and Kolotilin (2017) and Lipnowski, Ravid and Shishkin (2019))), or in which the sender incurs some cost for cheating (see, for example, Kartik, Ottaviani and Squintani (2007) and Kartik (2009)), or in which the receiver also has private information and can therefore assess the honesty of the sender (Olszewski (2004)). The equilibria in these models share some of the properties of the one-

period model presented here. However, thanks to an independence result obtained here, we are able to derive novel results regarding the properties of the market's equilibrium in a multi-period model, as well as its implications.

A growing empirical literature examines the impact of reviews on consumers along various dimensions. Kim and Martin (2018) use online experiments to ascertain 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 result of our model which predicts that in expectation an additional review incrementally corrects mistaken 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.

Lastly, our model relates to the phenomenon of *fake news* in the sense that it explores the extent to which a *long-run anonymous player* with a political agenda can derail information aggregation.[1]

3 The One-Period Model

We start with the case of two players: a sender (S) and a receiver (R). Player S can be one of two types: with probability q he is honest, (S_h), and with probability $(1-q)$ he is fake (S_f). The sender's type is chosen by nature before the beginning of the game.

The "state of the world" (e.g., the quality of the product) is a random variable $\theta \in \{0,1\}$

[1] As we explain in the conclusions, this relates to models of *media bias* such as Prat (2018) and Perego and Yuksel (2018) which respectively analyze the media's ability to influence voters and the social value of the information provided by the media in equilibrium.

and is not known to either player. p is the common prior that $\theta = 1$. If and only if the sender is of type S_h, then conditional on the realization of the state of the world θ, the sender (but not the receiver) receives a signal $\tilde{x}$, which takes a value in $[0, 1]$ according to the density $t_\theta(x)$ and the cdf $T_\theta(x)$. We make the following assumptions:

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

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

Assumption A.2 (hereafter referred to as MLRP (monotone likelihood ratio property)) captures the idea that the larger the signal, the more likely it is that $\theta = 1$. Define $\bar{x}$ to be the (unique) signal for which:

$$t_1(\bar{x}) = t_0(\bar{x}). \tag{1}$$

That is, $\bar{x}$, referred to as the *neutral news signal,* does not change the sender's prior. In fact, by MLRP, signals above (below) neutral news imply positive (negative) updating.

After observing the signal x, the sender sends a message $m \in [0, 1]$ to the platform. Upon receiving a message m from S the receiver uses Bayes' rule to update her beliefs about the state of the world. We assume that the receiver does not know the sender's type and assigns a probability q to the event that S is honest. The honest sender, S_h, reports his signal truthfully (i.e., $m = x$) whereas the fake sender, S_f (to be referred as Fake), chooses m strategically. We assume first that Fake's payoff is increasing with $R's$ posterior that the state is 1, although later we explore the case in which Fake's payoff can either increase or decrease with the receiver's posterior.

Initially, we assume that the platform is not strategic and simply forwards the message to the receiver. We will later show that such behavior is optimal for the receiver even when the sender is potentially fake.

3.1 Preliminaries

We define $\hat{p}(m)$ as the posterior probability that the state is 1, given message m and if the sender were *known to be honest*:

$$\hat{p}(m) = \frac{pt_1(m)}{pt_1(m) + (1-p)t_0(m)}$$

Let $E_\theta[\hat{p}]$ denote the expected value of $\hat{p}(m)$ given that the true state is θ. By assumption A.2, $\hat{p}(m)$ is clearly increasing in m.

A strategy for S_f is a distribution function defined over the set of all messages $M = [0,1]$.[2]

Note that upon observing the neutral news message $m = \bar{x}$ the receiver will not update her prior *regardless of Fake's strategy*: the receiver's prior will not change whether $m = \bar{x}$ is known to be coming from a fake sender (because it never does), or from an honest sender (because it would not for $m = \bar{x}$).

The following lemma states simply that, in equilibrium, Fake never assigns a strictly positive probability to any message m.

Lemma 1 *Fake's equilibrium strategy is atomless.*

Proof. See the appendix. ■

The intuition for this result is that, since the true/honest signal distributions have no atoms, atoms cannot help the Fake sender since they would reveal his identity to the receiver. Having ruled out atoms in equilibrium, we now view Fake's strategy as a density $f(m)$. Let

[2] The assumption that Fake does not observe $\tilde{x}$ or the state seems intuitive in our context. If we allow the fake sender to observe the signal then, besides the equilibrium that we characterize, a special type of babbling equilibrium may also arise: the Fake nullifies the information of the honest sender by sending the singnals according to the honest distribution albeit in the *wrong* state. Thus, there is no updating for any signal so it is a akin to a babbling equilibrium: if the receiver (correctly) believes all signals are uninformative, then no deviation can be profitable. Evidently, a strategic platform has no scope for improving this zero information transmission equilibrium.

$\hat{p}(m \mid f)$ denote the receiver's posterior probability that the state is 1 upon receiving the message m and given that with probability q the sender is honest and with probability $(1-q)$ is fake. Let $E_\theta[\hat{p}(m \mid f)]$ denote the expected value of $\hat{p}(m \mid f)$ given that the true state is θ. Given (p, q, t_0, t_1), let:

$$\hat{P}(m \mid f) \equiv \frac{\hat{p}(m \mid f)}{1-\hat{p}(m \mid f)} = \frac{p}{(1-p)} \frac{qt_1(m) + (1-q) f(m)}{qt_0(m) + (1-q) f(m)} = P \frac{Qt_1(m) + f(m)}{Qt_0(m) + f(m)} \tag{2}$$

where $P = p/(1-p)$ and $Q = q/(1-q))$. Thus, $\hat{P}(m \mid f)$, hereafter referred to as the receiver's *likelihood ratio*, can be thought of, without loss of generality, as Fake's payoff from sending the message m when the receiver believes that Fake is playing the strategy f. For the range of values of m, note that:

- When $f(m) = 0$: $\hat{P}(m \mid f) = P\frac{t_1(m)}{t_0(m)}$.
- When $f(m) > 0$: $m \gtrless \bar{x} \implies \hat{P}(m \mid f) \lesseqgtr P\frac{t_1(m)}{t_0(m)}$.

4 One-Period Equilibrium

Let f^* denote Fake's equilibrium strategy. The following proposition provides a set of conditions that f^* must satisfy:

Proposition 1 *If f^* is an equilibrium strategy for Fake, then there exists a point $z \in (\bar{x}, 1)$ 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 any message) is $P\frac{t_1(z)}{t_0(z)}$.*

Proof. See the appendix. ■

The following theorem establishes existence and uniqueness:

Theorem 1 *(i) An equilibrium exists and is unique.*

(ii) Fake's equilibrium strategy is:

$$f^*(m) = \begin{cases} 0 & \text{for: } m \in [0, z] \\ Q\frac{t_0(z)t_1(m) - t_1(z)t_0(m)}{t_1(z) - t_0(z)} & \text{for: } m \in (z, 1] \end{cases} \tag{3}$$

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

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

(iii) Fake's cutoff z is increasing in q.

Proof. See appendix. ■

Intuitively, the equilibrium can be described as follows: Fake randomizes over an interval $(z, 1]$ in a way that generates the *same posterior* for the receiver at all $m \in (z, 1]$, which is equal to the receiver's posterior after the threshold message $m = z$. That is, the equilibrium likelihood ratio for all $m \in (z, 1]$ has the property:

$$\hat{P}(m \mid f^*) = \hat{P}(z \mid f^*) = P\frac{t_1(z)}{t_0(z)} > P.$$

To guarantee the constant posterior, $f^*(m)$ is strictly increasing in $m \in (z, 1]$, implying that higher values of m are also more likely to originate from Fake. Lastly, the more likely it is that the sender is fake (the lower is q), the lower z will be and consequently the lower will be the posterior/persuasion Fake can generate. Thus, even though Fake is able to "manipulate" the receiver's beliefs by generating a posterior $\hat{P}(z \mid f^*)$ that is higher than the prior P, he can only do so to a limited extent and his ability to manipulate decreases with q.

Example 1 *Consider the following linear example:*

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

For $q = 3/4$, we have $z = 2/3$. Figure 1 depicts Fake's strategy.[3]

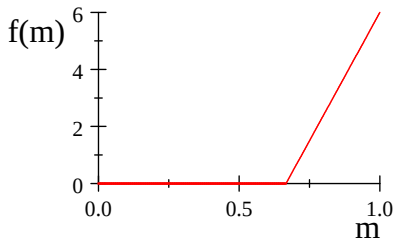

Figure 1

Figure 2 shows the receiver's posterior upon receiving a message m, starting from a prior $p = 1/2$ (dashed line). The vertical distance between the solid line and the dashed line depicts the equilibrium persuasion achieved by a fake sender who manages to induce a posterior above the prior.

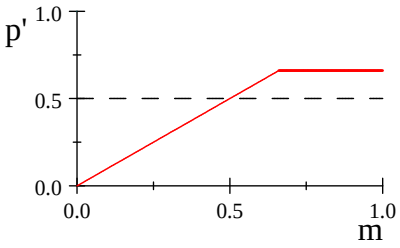

Figure 2

Figure 3 shows the expected message distribution $\tau_0(m) = qt_0(m) + (1-q)f^*(m)$ (the solid line) when the state is $\theta = 0$, as compared to the honest sender distribution $t_0(m)$ (the dashed line):

[3] This continuous review setup is used because it is "cleaner", although its discrete analog would achieve the same purpose. If reviews are extremely coarse, i.e. binary, as in the case of "likes," then the *pure strategy* with only "up-likes" is the equilibirum.

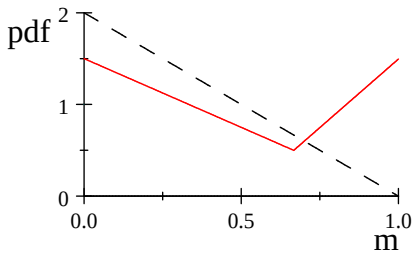

Figure 3

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, a result that will be particularly useful in the analysis of the multi-period multi-sender game. The independence result is intuitive since it emerges from the fact that if some piece of information is *better news* than another under some prior, then the same should be true under any other prior. As a consequence, information that maximizes the posterior (i.e. is *best news*) under some prior should also do so under any other prior.

We next present a *learning* result which states that: (i) as long as there is some strictly positive probability that the sender is honest, then the receiver benefits from paying attention to the sender's messages; and (ii) the more likely it is that the sender is honest, the higher is the receiver's benefit. While the first part of the proposition is somewhat obvious, the second part is more surprising given that the fake sender is strategic and becomes more aggressive as q increases (see (iii) in Theorem 1).

In order to prove this proposition, it will be more convenient to focus on the receiver's prior p rather than the likelihood ratio P and on her posterior probability $\hat{p}(m \mid f^*)$ rather than her posterior likelihood ratio $\hat{P}(m \mid f^*)$. Let $E_\theta[\hat{p}_{f^*}]$ denote the receiver's expected posterior probability that the state is 1, given that the true state is θ.

Proposition 2 ***Learning:***

$$\textit{(i)}\ E_0[\hat{p}_{f^*}] < p < E_1[\hat{p}_{f^*}].$$

$$\textit{(ii)}\ \frac{dE_1[\hat{p}_{f^*}]}{dq} > 0, \qquad \frac{dE_0[\hat{p}_{f^*}]}{dq} < 0.$$

Proof. *See the appendix* ■

Remark 1 *Suppose that Fake can be one of two types: Fake-1 (S_f^1) or Fake-0 (S_f^0), where Fake-1's payoff increases with the receiver's posterior while Fake-0's decreases with the receiver's posterior. The sender is honest with probability $q > 0$, Fake-1 with probability $q_1 > 0$ and 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 $q/(1-q_0)$. Fake-0's equilibrium strategy is the mirror image of Fake's 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 $q/(1-q_1)$.*

Example 2 *Consider the linear case discussed in Example 1. Figure 4 and 5 depict the Fake-0 strategy (the green line) and the Fake-1 strategy (the red line), as well as the receiver's posterior as a function of the message he receives, for the parameters ($p = \frac{1}{2}$ (dashed line), $q = \frac{15}{21}, q_1 = \frac{1}{21}, q_0 = \frac{5}{21}$).*

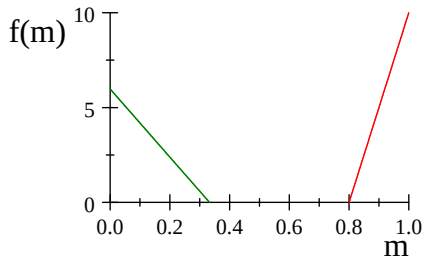

Figure 4

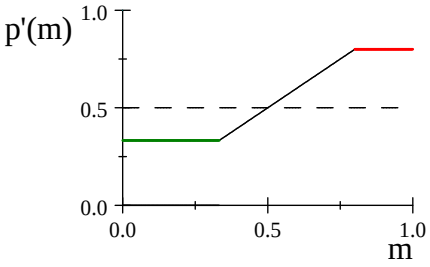

Figure 5

5 Strategic Platform

Up to this point, we have assumed that the platform is nonstrategic in the sense that it truthfully reports any message it receives. The question that arises is whether the platform can do better by somehow manipulating the messages it receives before sending them to the receiver. By doing so, the platform may be able to induce the fake sender to alter his strategy in a way that will make it less harmful to the receiver. An example of such manipulation might be to delete extreme messages or pool some messages (polices often used by platforms). In what follows, however, we show that any such manipulation by the platform can only make the receiver worse off.

We assume that the platform's objective is to be as informative (a la Blackwell) as possible. Such an assumption is intuitive in a market where platforms compete for users

who make choices (such as choosing the share of risky assets in their portfolio) based on the information they obtain from the platform. This will be the case, for example, if the receiver is choosing an action $x \in [0,1]$ and his VNM utility from choosing x in state $\theta \in \{0,1\}$ is $-(x-\theta)^2$. The notion of informativeness employed here is second-order stochastic dominance (see Blackwell and Girshick (1979)).

We adopt a mechanism design approach and assume that the platform commits in advance to a reporting policy g, known to both the sender and the receiver. Formally, the platform's strategy $g : [0,1] \to [0,1]$ assigns to each message it receives from the sender a message to be sent to the receiver. Since the platform does not have any information other than the message it receives, the only manipulation it can apply is to pool some messages.[4] Clearly, if not for the presence of a fake sender, such a strategy would certainly make the receiver worse off.

The driving force behind the result is the characterization of Fake's equilibrium strategy f_g^* when the platform applies the policy g. Essentially, Fake's optimal strategy in this case is similar to the one it used when the platform was not strategic, with the only modification that it is now facing a different distribution of messages. In other words, Fake will, as before, assign a positive weight to those messages that, in its absence, would yield the highest posterior, and it will do so in a way that equalizes the posteriors for all these messages. Notice, however, that since the posterior is not necessarily monotonic, in the presence of a strategic platform and in the absence of the fake sender, the support of Fake's strategy may consist of more than one interval.

Recall that $\hat{p}(m)$ is the receiver's posterior that the state is 1 in the absence of the fake sender and in the absence of manipulation by the platform. Similarly, let $\hat{p}_g(m)$ denote the receiver's posterior belief that the state is 1, given that the sender is honest and given

[4] Allowing the platform to randomize over messages does not change any of the results as will become clear from the analysis that follows. Similarly, allowing the platform to remain silent is equivalent to it choosing the same message in all those cases.

that the platform received the message m and applies the strategy g. Notice that m is not necessarily the message sent by the platform but rather the message sent by the sender before the manipulation by the platform. Thus, $\hat{p}_g(m)$ is the posterior that the sender induces when it sends the message m. That is,

$$\hat{p}_g(m) = \frac{\int_{m'|g(m')=g(m)} pt_1(m')dm'}{\int_{m'|g(m')=g(m)} [pt_1(m') + (1-p)t_0(m')]dm'}.$$

Let $\hat{p}_g(m \mid f_g)$ denote the receiver's posterior given the sender's (fake or honest) message m , Fake's strategy f_g and the platform's strategy g. That is

$$\hat{p}_g(m \mid f_g) = \frac{\int_{m'|g(m')=g(m)} p[qt_1(m') + (1-q)f_g(m')]dm'}{\int_{m'|g(m')=g(m)} [q[pt_1(m') + (1-p)t_0(m')] + (1-q)f_g(m')]]dm'}.$$

We can now state the following proposition (the proof of which is omitted since it is essentially a repetition of the arguments in the proof of Proposition 1).

Proposition 3 *If f_g^* is an equilibrium strategy for Fake, then:*

(i) if $f_g^*(m') > 0$ and $f_g^*(m") > 0$, then $\hat{p}_g(m' \mid f_g^*) = \hat{p}_g(m" \mid f_g^*) \equiv \bar{p}_g$.

(ii) if $\hat{p}_g(m) \leq \bar{p}_g$, then $f_g^*(m) = 0$.

Thus, in equilibrium, all messages to which Fake assigns a positive probability induce the same posterior $\bar{p}_g$, which is the highest induced posterior inferred by the receiver. Notice however that unlike the case in which the platform is not strategic, and since $\hat{p}_g(m)$ is not necessarily monotonic, Fake's strategy may not be monotonic in this case.

Recall that $\hat{p}(\cdot \mid f^*)$ is the distribution of posteriors in equilibrium when the platform is nonstrategic, and $\hat{p}(z \mid f^*) \equiv \bar{p}$ is the highest point in its support. The following proposition establishes that the highest point in the support of the distribution of posteriors when the platform is strategic, i.e. $\bar{p}_g$, cannot be above $\bar{p}$. Roughly speaking, the reason for this result

is that the introduction of a strategic platform shifts the distribution of posteriors to the left. Therefore, when the fake agent assigns a positive weight to these posteriors, they are shifted even further to the left.

Proposition 4 $\bar{p}_g \leq \bar{p}$.

Proof. *For every m, let $\gamma(m) = \{m' \mid g(m') = g(m)\}$. That is, $\gamma(m)$ is the set of all messages for which the platform sends the same message as in the case of m. Without loss of generality we can assume that if $f_g^*(m) > 0$ for some m, then $f_g^*(m') > 0$ for all $m' \in \gamma(m)$ (since the platform sends the same message for all messages in $\gamma(m)$). There are two cases to consider:*

(i) There exists a message m' such that $f_g^*(m') > 0$ and $\gamma(m')$ is contained in the interval $[0, z)$. In such a case, since $\hat{p}(m) < \bar{p}$ for all $m \in \gamma(m')$, it must be that $\hat{p}_g(m) < \bar{p}$ for all messages in $\gamma(m')$ and hence $\hat{p}_g(m \mid f_g^*) < \bar{p}$. By Proposition 3, $\hat{p}_g(m \mid f_g^*) = \bar{p}_G < \bar{p}$ for all messages in the support of $f_g^*(m)$.

(ii) For every m' for which $f_g^*(m') > 0$, $\gamma(m') \cap [z, 1] \neq \varnothing$. We first show that for any strategy g adopted by the platform, if Fake uses his original strategy f^*, then, for every message m he sends, the posterior is (weakly) lower than $\bar{p}$. Formally, let $\hat{p}_g(m \mid f^*)$ denote the receiver's posterior if the sender sends the message m, the fake sender plays the equilibrium strategy f^* in the game in which the platform is not strategic, and the strategic platform chooses the strategy g. Then, for every m',

$$\begin{aligned} \hat{p}_g(m' \mid f^*) &= \frac{p\int_{\{m|m\in\gamma(m')\}}(qt_1(m) + (1-q)f^*(m))dm}{\int_{\{m|m\in\gamma(m')\}}(q(pt_1(m) + (1-p)t_0(m)) + (1-q)f^*(m))dm} \qquad (6) \\ &\leq \frac{p\int_{\{m|m\in\gamma(m'),\ m\ \in[z,1]\}}(qt_1(m) + (1-q)f^*(m))dm}{\int_{\{m|m\in\gamma(m'),\ m\ \in[z,1]\}}(q(pt_1(m) + (1-p)t_0(m)) + (1-q)f^*(m))dm} = \bar{p}. \end{aligned}$$

The inequality above states that since the platform is simply garbling any given strategy used by Fake, then the posterior under strategy f^* can never be larger than $\bar{p}$.

Therefore, since $\bar{p} > p$, in order for the fake sender to induce a posterior higher than $\bar{p}$, it must be that under his new strategy f_g^*, there exists m' for which:[5]

$$\int_{\{m|m\in\gamma(m')\}} f_g^*(m)dm < \int_{\{m|m\in\gamma(m')\}} f^*(m)dm,$$

and $\hat{p}_g(m' \mid f_g^*) > \bar{p}$.

Since $\int_0^1 f_g^*(m)dm = 1$, it must also be that there exists a message m" for which

$$\int_{\{m|m\in\gamma(m")\}} f_g^*(m)dm > \int_{\{m|m\in\gamma(m")\}} f^*(m)dm.$$

Since by 6, $\hat{p}_g(m" \mid f^*) \leq \bar{p}$, it must be that

$$\hat{p}_g(m" \mid f_g^*) < \bar{p}.$$

But this is in contradiction to Proposition 3 in which it is shown that in equilibrium Fake assigns a strictly positive probability only to messages that induce the highest posteriors. ■

We will now prove that the receiver cannot be better off when the platform is strategic. That is, the equilibrium distribution of posteriors when the platform is strategic second-order stochastically dominates the distribution of posteriors when the platform is not strategic. Denote the equilibrium cumulative distribution of the posteriors by $\Pi_G\left(\hat{p}_G(\cdot \mid f_g^*)\right)$ when the platform is strategic and by $\Pi(\hat{p}(\cdot \mid f^*))$ when it is not.

Definition 1 *The posterior distribution $\hat{p}_G(m \mid f_g^*)$ is less informative than the posterior distribution $\hat{p}(m \mid f^*)$ if for all $x \in [0,1]$:*

$$\int_0^x \Pi_G\left(\hat{p}_G(\cdot \mid f_g^*)\right) d\hat{p}_G(\cdot \mid f_g^*) \leq \int_0^x \Pi(\hat{p}(\cdot \mid f^*))d\hat{p}(\cdot \mid f^*). \tag{7}$$

[5] The reason for this is that when the posterior is higher than the prior, a positive probability weight by Fake lowers the posterior whereas when the posterior is lower than the prior, a positive weight by Fake increases the posterior.

Theorem 3 *$\hat{p}(\cdot \mid f^*)$ is more informative than $\hat{p}_G(\cdot \mid f_g^*)$.*

Proof. *We need to show that (7) holds for all $x \in [0,1]$. First, observe that for all $x \in [0, \bar{p}_G]$ the above inequality holds since the only difference between $\hat{p}(\cdot \mid f^*)$ and $\hat{p}_G(\cdot \mid f_g^*)$ is the result of the platform's pooling strategy (since Fake does not operate in this region). Next, notice that since $\bar{p}_G \leq \bar{p}$, it follows that for every $x > \bar{p}_G$ we have $\Pi_G\left(\hat{p}_G(\cdot \mid f_g^*)\right) \geq \Pi(\hat{p}(\cdot \mid f^*)$. Since the two distributions have the same expected value (i.e., the prior) inequality (7) holds for all $x \in [0,1]$.* ■

Furthermore, using essentially the same proof as in the case of a strategic platform, we show that if the "honest" sender could behave strategically (in order to maximizes the receivers' welfare), then multiple equilibria would exist, however, the best of those from the receiver's viewpoint would be the one in which the honest sender is nonstrategic.

Remark 2 ***Strategic Honest Sender.*** *We have assumed throughout that the honest sender is not strategic and simply reports his signal. In the context of consumer reviews, this appears to be a realistic assumption. Given the above result, which showed that the receiver cannot benefit from a strategic platform, it is straightforward to show (using the same arguments) that if the honest sender is strategic, then the best equilibrium from the receiver's viewpoint would be the one in which the honest sender is not strategic.*[6]

6 The N-Period Model

In this section, we extend the model to N periods and allow for many senders, some of whom are honest and some of whom are fake, who send messages at different times.

[6] Formally, it can be shown that every equilibrium outcome in the case of a strategic honest sender while a fake sender can be obtained in the game in which the honest sender is not strategic and the platform is and commits in advance to a policy. Consider the following "direct" game in which for every m received by the platform, the platform reports the posterior obtained for that m in the equilibrium of the game between a strategic honest sender and a fake sender. Obviously, it is now optimal for the fake sender to use the same strategy as in the game with the strategic honest sender and hence the posterior reported by the platform is the correct one.

Senders can appear more than once, and fake senders are not necessarily myopic when choosing their strategy. We also allow for multiple receivers who form their beliefs after observing messages at various points in time. To characterize the equilibrium of the general model, we rely heavily on the independence result in Corollary 2 which implies that a fake sender's action in a given period is not affected by previous messages (sent either by himself or by other senders) and will not affect his actions or those of other senders in future periods. In what follows, we begin assuming the platform is not strategic and then show that this assumption is without loss of generality also valid in a multi-period model.

There is a pool of receivers and in every period $n \in (1, 2, ..., N)$ one of them (who may have already been a receiver in a previous period) is drawn from that pool and forms her posterior based on the history of messages up to period n.[7]

Let L_0 be the set of fake senders whose objective is to minimize all receivers' posterior that the state is 1. Likewise, L_1 is the set of fake senders whose objective is to maximize the receivers' posterior that the state is 1, and L_h is the set of honest senders. The set of all senders is denoted by L. In every period, sender l is selected with probability q^l, where $q^l \geq 0$ and $\sum_{l \in L} q^l = 1$, 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 $\bar{q}_0 + \bar{q}_1 + \bar{q}_h = 1$. If l is honest, then he truthfully reports his signal; otherwise he reports strategically.

A strategy for a fake sender l , denoted by σ_N^l, specifies his move, for every period n, given the history of previous messages, in the case that he is selected to move in that period. Formally, $\sigma_N^l = \{f_n^l\}_{n \in N}$ where $f_n^l : [0,1]^n \to R_+$. That is, σ_N^l assigns a weight to every message $m_n \in [0,1]$ for every possible history of messages $(m_1, ..., m_{n-1}) \in [0,1]^{n-1}$. We say that $\sigma_N^* = \{\sigma_N^{l*}\}_{l \in L_0 \cup L_1}$ is *an equilibrium* if for all $l \in L_0 \cup L_1$, σ_N^{l*} is a best response for player l, given the strategies of all the other players.

[7] The result can easily be extended to the case in which each receiver observes a subset of the messages up to period n, which are randomly drawn according to some commonly known distribution.

Let $f^*_{\bar{q}_o}$ and $f^*_{\bar{q}_1}$ be the equilibrium strategies of Fake-0 and Fake-1 respectively, in the two-sided one-period model where the sender is type Fake-θ with probability $\bar{q}_\theta$ and is honest with probability $1- \bar{q}_0 - \bar{q}_1$. The following proposition states that a fake agent's strategy is stationary in the sense that it is independent of n and of the history of messages up to that period. Furthermore, in every period a type Fake-θ's strategy chooses messages in the same way that a sender of his type would have done in the (two-sided) one-period model in which his type is chosen with probability $\bar{q}_\theta$.

Proposition 5 *Let l be a fake sender. Then, σ^*_N is unique and for all $n = 1, ..., N$, $f^{l*}_n \equiv f^*_{\bar{q}^\theta}$ if l is of type Fake-θ.*

Proof. Consider period N. Recall the independence result in Corollary 2 which states that a Fake-θ agent's strategy in a one-period game is a function of the probability that the receiver assigns to his type, which is independent of the prior. Therefore,

$$f^{l*}_N (m_1, ..., m_{N-1}, m_N) \equiv f^*_{\bar{q}_\theta}(m_N)$$

for all $(m_1, ..., m_{N-1})$ and $l \in L_\theta$.

Given that the last period's strategies are independent of the history, we can now move one step backwards and claim that:

$$f^{l*}_{N-1} (m_1, ..., m_{N-2}, m_{N-1}) \equiv f^*_{\bar{q}_\theta}(m_{N-1}).$$

A similar argument can be applied to all periods. ■

6.1 Strategic platform in the N-period model

In a multi-period model, a strategic platform can apply strategies that are not feasible in the one-period case. For example, the platform can condition the messages it forwards to

the receiver on the messages it received in previous periods. Since the platform can commit to a reporting mechanism, such a policy could potentially alter the fake sender's strategy in some periods so as to benefit receivers overall. However, and as in the case of the one-period model, such a strategy can only harm the receivers.

Theorem 4 *In the N-period model, the optimal strategy for a strategic platform is to truthfully reveal the signals it receives in every period.*

Proof. *Consider a fake sender in period $n < N$. Regardless of the platform's strategy, the expected posterior in all periods $n' > n$ is the posterior obtained at the end of period n. Thus, Fake's strategy in period n is not affected by the platform's strategy in any period $n' > n$, but only by the platform's strategy in period n. With this in mind, we can now use Theorem 3 to argue that the platform's optimal strategy is to truthfully reveal the message it receives in every period.* ■

7 Conclusion and Further Research

We have proposed a simple and parsimonious model of information aggregation in the presence of fake reviews. A major advantage of this model is nonetheless its several potential applications and extensions. Since the model is malleable and delivers a unique prediction, it can be used to answer a number of questions regarding the supply of fake reviews that are examined in the industrial organization (IO) domain. For instance, when would a business want to hire fake reviewers? What proportion of fake reviews is optimal for a (dishonest) business? When can negative fake reviews be used to sink new and competing, though as yet unreviewed, products? To answer these questions one needs to know the amount of persuasion that can be obtained from each fake review. Therefore, several IO (first-stage) questions can be addressed by recasting our simple model as the second stage in the game.

A natural extension of the model would be to add uncertainty, so that learning occurs on

the *proportion* of fake reviews. This would entail dynamic path dependence since long-term senders would be trying to also persuade receivers that the number of fake reviews is low, so as to better disguise their fake messages and achieve greater impact on beliefs. A special case of this, which applies more to the case of *fake news* (than to fake reviews where aliases are used), is that of *non-anonymous*, possibly honest, long-term senders. If a fake sender is recognized as a sender of multiple messages (i.e., possibly fake news articles), then he might want to occasionally send true news articles in order to conceal his objective. Along these lines, our minimal model can serve as a benchmark to analyze the effect of fake long-term senders on information aggregation and voting outcomes in a fake news world.

8 References

1. Aköz K.K., C.E. Arbatli and L.Çelik (2018): "Manipulation through Biased Product Reviews". Working paper.

2. Avery, C. and M. Meyer (2012): "Reputational Incentives for Biased Evaluators". Working paper.

3. Benabou R. and G. Laroque (1992): "Using privileged information to manipulate markets: insiders, gurus, and credibility". *Quarterly Journal of Economics* 107 (3), 921–958.

4. Blackwell A. David and M. A Girshick. (1979) Theory of Games and Statistical Decisions. Courier Corporation .

5. Chen Y. (2011): "Perturbed communication games with honest senders and naive receivers," *Journal of Economic Theory* 146 401–424.

6. Crawford V. and J. Sobel (1982): "Strategic information transmission". *Econometrica* 50 1431–1452.

7. Dziuda W. (2011): "Strategic Argumentation". *The Journal of Economic Theory*, vol. 146, issue 4, 1362-1397

8. Francetich A. and D. Kreps (2014): "Bayesian inference does not lead you astray. . . on average". *Economics Letters*, vol. 125, issue 3, 444-446.

9. Gratton G. and R. Holden . (2017): "When to Drop a Bombshell". *Review of Economic Studies,* 85, 2139–2172.

10. Kartik N. (2009): "Strategic communication with lying cost". *Review of Economic Studies,* 76, 1359–1395.

11. Kartik N., M. Ottaviani and F. Squintani (2007): "Credulity, lies and costly talk". *Journal of Economic Theory,* 134 93–116.

12. Kreps D., P. Milgrom, J. Roberts and R. Wilson (1982): "Rational cooperation in the finitely repeated prisoners' dilemma". *Journal of Economic Theory*, vol. 27, issue 2, 245-252.

13. Kim K. and J. Pogach (2014): "Honest vs. Advocacy". *Journal of Economic Behavior & Organizatio*n 105 51–74

14. Lahr P and J.Winkelmann (2018): "Fake Experts". Working paper.

15. Laouenan M., and R. Rathelot (2018): "Ethnic Discrimination on an Online Marketplace of Vacation Rentals". Working paper.

16. Luca M., and G. Zervas (2016): "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud". *Management Science*, 62, No 12.

17. Lipnowski E., D. Ravid and D. Shishkin (2018): "Persuasion via Weak Institutions". Working paper.

18. Kim, T. and D. Martin (2018): "Inference about Ratings: How Good Is a Good Rating? Working paper.

19. Mayzlin, D., Dover, Y., & Chevalier, J. (2014) "Promotional reviews: An empirical investigation of online review manipulation." *American Economic Review,* 104(8), 2421-55.

20. Morgan J. and P.C. Stocken (2003): "An analysis of stock recommendations". *Rand Journal of Economics* 34, 183–203.

21. Olszewski W. (2004): "Informal communication " *Journal of Economic Theory* 117(2), 180-200.

22. Ottaviani M. and F. Squintani (2006): "Naive audience and communication bias". *International Journal of Game Theory* 35 129–150.

23. Perego J. and S. Yuksel (2018): "Media Competition and Social Disagreement " working paper.

24. Prat A. (2018): "Media Power " *Journal of Political Economy* 126:4, 1747-1783.

25. Smirnov A. and E. Strakov (2018): "Bad News Turned Good: Reversal Under Censorship". Working paper.

26. The Economist (2015): "Five-star fakes: The evolving fight against sham reviews" Print Edition Oct 22.

9 Appendix

Proof of Lemma 1. Assume that there exists an equilibrium strategy f^* for Fake with an atom at m, namely containing a message $m \in [0,1]$ occurring with strictly positive probability in equilibrium. Given that the true signals $t_\theta(x)$ come from an atomless distribution,

when a receiver observes the message m, she must conclude that the sender of m is fake and therefore she will not update her prior. Since any such atoms are of no use to the Fake since they do not increase the receiver's posterior, a deviation is easy to find. 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^* has no atom at m'. Upon receiving this m' the receiver's posterior increases: there is positive updating as the receiver can no longer rule out that m' comes from an honest sender. Thus, deviating to the message m' is strictly better for Fake than sending the message m, a contradiction. ■

Proof of Proposition 1. The proof is established by proving a series of claims. ■

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

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

Claim 2 *If* $f^*(m) > 0$ *for some* $m \neq 1$, *then* $\hat{P}(m \mid f^*) = \hat{P}(1 \mid f^*)$.

Proof. *Follows from Claim 1 and the definition of equilibrium.* ■

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 1 that* $\hat{P}(m' \mid f^*) = \hat{P}(1 \mid f^*)$ *or*

$$P\frac{Qt_1(m') + f^*(m')}{Qt_0(m') + f^*(m')} = P\frac{Qt_1(1) + f^*(1)}{Qt_0(1) + f^*(1)} \tag{8}$$

contradicting that $t_1(m') \leq t_0(m')$ *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 $m' < m" \leq 1$, it must be that $f^*(m") > 0$.*

Proof. *Assume that there exists $\bar{x} < m' \leq 1$ and $m" \in (m', 1)$ such that $f^*(m') > 0$ and $f^*(m") = 0$. It follows that $\hat{P}(m" \mid f^*) = P\frac{t_1(m")}{t_0(m")} > P\frac{Qt_1(m')+f^*(m')}{Qt_0(m')+f^*(m')} = \hat{P}(m' \mid f^*)$, 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 $P\frac{t_1(z)}{t_0(z)}$.*

Proof. *We start by proving (i). From the four previous claims 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 $m \in [z, 1]$:*

$$\hat{P}(m \mid f^*) = \hat{P}(z \mid f^*) = P\frac{Qt_1(z) + f^*(z)}{Qt_0(z) + f^*(z)} < P\frac{t_1(z)}{t_0(z)}.$$

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

$$P\frac{t_1(m')}{t_0(m')} > \hat{P}(z \mid f^*) = P\frac{Qt_1(z) + f^*(z)}{Qt_0(z) + f^*(z)}$$

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

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

$$\hat{P}(m \mid f^*) = P\frac{Qt_1(m) + f(m)}{Qt_0(m) + f(m)} < P\frac{t_1(m)}{t_0(m)}$$

and since, by A.1, $\lim_{m\downarrow z} P\frac{t_1(m')}{t_0(m')} = P\frac{t_1(z)}{t_0(z)}$*, it follows that for* m *close enough to* z*,*

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

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

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 $\hat{P}(m \mid f^*) = P\frac{Qt_1(m)+f^*(m)}{Qt_0(m)+f^*(m)} = P\frac{t_1(z)}{t_0(z)}$ for all $m \in (z, 1]$, . We can therefore solve for f^* 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 $m \in [z, 1]$ 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 μ. Define $\Psi(\mu) = \int_\mu^1 \psi(m \mid \mu)\, dm$ and observe that $\Psi(\mu)$ is strictly decreasing with μ and is unbounded as μ approaches $\bar{x}$ from above. Since $\Psi(1) = 0$, we conclude that there exits a unique z such that $\Psi(z) = 1$. Lastly, z increases in q because from (3) and (4) we have:

$$\frac{1}{Q} = \frac{1}{Q}\int_z^1 f^*(m)dm = \int_z^1 \frac{l(m) - l(z)}{l(z) - 1} t_0(m)\, dm$$

The RHS is strictly decreasing in z, since by MLRP the integrand is positive and strictly decreasing in z. ■

Proof of Proposition 2. (i) If $E_0[\hat{p}_{f^*}] < E_1[\hat{p}_{f^*}]$, this follows trivially from the fact that

their *average* is p. Since the terms in f cancel out, we obtain:

$$E_1[\hat{p}_{f^*}] - E_0[\hat{p}_{f^*}] = q\int_0^1 \hat{p}_{f^*}(m)\,(t_1(m) - t_0(m))\,dm > 0.$$

The inequality follows because MLRP implies strict first order stochastic dominance and from the fact that $\hat{p}_{f^*}(m)$ is non-decreasing overall and (by MLRP) increasing below the neutral signal where $\hat{p}_{f^*}(m) = \hat{p}(m)$.[8]

(ii) We prove the theorem by showing that:

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

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

$$= q\int_0^1 \hat{p}_{f^*}(m)\,(t_1(m) - t_0(m))\,dm = q\left(\begin{array}{c}\displaystyle\int_0^z \hat{p}_{f^*}(m)\,(t_1(m) - t_0(m))\,dm \\ \displaystyle+\int_z^1 \hat{p}_{f^*}(m)\,(t_1(m) - t_0(m))\,dm\end{array}\right).$$

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

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

where the first term is positive due to MLRP and the last is positive since $\frac{d\hat{p}_{f^*}}{dq} > 0$ and $t_1(m) > t_0(m)$ for all $m > z$. ■

[8] See also Corollary 1 of Francetich and Kreps (2014).