DISCUSSION PAPER SERIES

DP13244

FAKE PERSUASION

Jacob Glazer, Helios Herrera and Motty Perry

INDUSTRIAL ORGANIZATION

ISSN 0265-8003

FAKE PERSUASION

Jacob Glazer, Helios Herrera and Motty Perry

Discussion Paper DP13244
Published 13 October 2018
Submitted 10 October 2018

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

FAKE PERSUASION

Abstract

We propose a model of product reviews with honest and fake reviews to study the value of information provided on platforms like TripAdvisor, Yelp, etc. In every period, a review is posted which is either honest, namely reveals the reviewer's true experience with the product/service, or fake, namely entirely fabricated in order to manipulate the public's beliefs. We establish that the equilibrium is unique and derive robust and interesting results about these markets. While fake agents are able to affect the public's beliefs in their preferred direction, aggregation of information takes place as long as some of the reviews are honest.

JEL Classification: C72, D82, D83

Keywords: Sender-Receiver Games

Jacob Glazer - glazer@post.tau.ac.il
University of Warwick and Tel Aviv University

Helios Herrera - h.herrera@warwick.ac.uk
University of Warwick and CEPR

Motty Perry - m.m.perry@warwick.ac.uk
University of Warwick

Acknowledgements
We thank Hila Goldenberg, Marina Halac, Sergiu Hart, Ilan Kremer, Antonio Nicoló, Debraj Ray and Phil Reny for helpful discussions.

Fake Persuasion*

Jacob Glazer
University of Warwick
and Tel Aviv University

Helios Herrera
University of Warwick
and CEPR

Motty Perry
University of Warwick

October 10, 2018

Abstract

We propose a model of product reviews with honest and fake reviews to study the value of information provided on platforms like TripAdvisor, Yelp, etc. In every period, a review is posted which is either honest, namely reveals the reviewer's true experience with the product/service, or fake, namely entirely fabricated in order to manipulate the public's beliefs. We establish that the equilibrium is unique and derive robust and interesting results about these markets. While fake agents are able to affect the public's beliefs in their preferred direction, aggregation of information takes place as long as some of the reviews are honest.

JEL Classification: C72, D82, D83.

Keywords: Sender-Receiver Games

*We thank Hila Goldenberg, Marina Halac, Sergiu Hart, Ilan Kremer, Antonio Nicoló, Debraj Ray and Phil Reny for helpful discussions.

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, AirB&B, Booking.com and many more sites where users write reviews on the basis of their experience with the site's services/products. This type of rating mechanism is rapidly growing and many experts believe that it will soon become the main source of information about a product's quality in many markets.[1] While many of the reviews in this type of platforms 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, *is aggregation of information still possible and if so under what conditions*? Indeed, 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 some other markets, such as that for news, as a result of the public's loss of trust in them. To see the growing importance of rating mechanisms in different markets and the amount of resources devoted to both generating fake reviews and mitigating them, the reader is referred to the interesting paper by Luca and Zervas (2016).

In this paper, we characterize the equilibria of markets where dissemination of information (whether fake or real) is cheap and convenient. We establish that the equilibrium is unique and derive robust and interesting results regarding the performance of these markets. We show that, despite the presence of sophisticated and strategic fake reviewers, valuable information aggregation takes place. While fake agents are able to affect the public's beliefs in their preferred direction, in expectation, the public's beliefs move in the correct direction as long as some of the reviews are written by honest reviewers. Had the fake agents merely sent random messages, the result that there is learning would not be surprising: it is the fact that the fake

[1] 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

senders are sophisticated, and that there may be many of them, that makes the result non-trivial.

We introduce a dynamic setup which, in our view, is a faithful description of many review platforms. In the model, a receiver (potential user) obtains information from multiple senders. Each sender (a possible past user), 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 (a possible subset of) past messages does not know the senders' types but is rational and understands that some of the messages were posted by fake senders (some of whom might send more than one message) who choose what information to convey with the intent of manipulating the receiver's beliefs for their own benefit.

We characterize the equilibrium and prove uniqueness. (Notice that in order to characterize the equilibrium of the interaction above we need only to focus on the fake senders' strategies.) We first prove an *independence* result: each sender's strategy is independent of the receivers' (prior) beliefs about the state of the world. This strong result leads to another interesting and useful observation that, in every period, each fake sender's message is completely independent of the messages sent (by him or other senders) in previous periods. This property is used to show that the strategy of a fake sender remains unchanged even when the size of the market changes (i.e. number of senders and receivers). Furthermore, we prove an *aggregation* result: in equilibrium valuable information aggregation takes place in *every period* and the more reviews an agent observes, the more informed she will be[2]. In the limit, information is fully aggregated. Thus, our analysis uncovers not only the strategy of a strategic fake senders in the face of rational receivers, but also the process of rational learning in situations like the one described in the model.

[2] We will refer throughout to the sender as "he" and to the receiver as "she".

In order to better understand these results, assume that the set of possible signals, that the sender observes, is an interval on the real line, say $[0,1]$. In equilibrium, a "positive fake" sender randomizes over some interval $[\bar{z},1]$, in a way that yields the same posterior for the receiver, who knows that the sender might be fake. In other words, the fake sender randomizes in a way that, given that the receiver knows the senders' equilibrium strategies, he (i.e., the fake sender) is indifferent among all messages along the interval $[\bar{z},1]$. The posterior induced by the "positive fake" sender's strategy is strictly higher than the receiver's prior. Similarly, a "negative fake" sender randomizes over an interval $[0,\underline{z}]$ where $\underline{z}<\bar{z}$, in such a way that the receiver's posterior is constant for all messages in that interval and lower than the prior. The *independence* result implies that each of the fake type's "cutoffs" ($\underline{z}$ and $\bar{z}$) is only a function of that type's share in the population of reviewers and *not* a function of the receiver's prior belief about the nature (quality) of the product. The *aggregation* result states that, in expectation, there is learning in every period as long as the fraction of fake senders is smaller than 1. The reason for this last result is that as long as the probability of the sender being honest is not zero, the (expected) distribution of messages when the state is high will not coincide with the (expected) distribution of messages when the state is low. Given that these two distributions are not the same the posterior moves in the right direction.

In sum, the internet has given rise to new information platforms which exploit the wisdom of the crowd. However, the increase in the number of reviews has led to an increase in the number of fake reviewers which compromise this wisdom. The current paper shows why these information platforms remain useful even when the presence of fake reviews is paramount. Our stark conclusion is that *fake reviewers succeed individually but fail on aggregate*: they affect welfare by slowing down this learning process but are not able overall to overcome the wisdom of the honest crowd and undo the learning process. Last but not least, the appeal and strength of this result relies on the uniqueness of the prediction. Uniqueness is quite uncommon in the communication literature where, for instance, the “bubbling” equilibrium typically coexists with other more informative ones.

Our analysis is relevant in many other contexts where 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 the issue at question, some of the comments are made strategically by interested parties. Similarly, the opinions voiced by participants in the very 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 and the participants in such forums know that some of them may be "fake," these platforms are 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 Related Literature

There is a large literature on sender-receiver communication games that incorporate a non-strategic player. Following the seminal paper by Kreps, Milgrom, Roberts and Wilson (1982), this assumption has become quite common in the literature and therefore we will mention only a small selection of papers that are close to ours in spirit. Our paper differs from the recent papers on this topic in that we allow for multiple senders and multiple receivers in a dynamic context where each message can influences future players' actions. The one closest to ours is Chen (2011). Building on Crawford and Sobel (1982)'s general communication game, she introduces the possibility of an honest sender and a naive receiver and solves for the equilibrium. Unlike in Chen's paper where the sender can be of infinitely many types, we have only allowed for two. Thus, the one-period one-sender version of our paper can be viewed as a special case of Chen (2011). Chen's model is an extension of Ottaviani and Squintani (2006) in which only the receiver may be naive. Their model is also extended in Kartik, Ottaviani and Squintani (2007) where it is assumed that the sender incurs a cost of lying. Kartik (2009) introduces the assumption that the sender incurs a convex cost for lying.

Unlike the aforementioned papers in which the sender is allowed to send only one message, Levy, de Barreda Moreno and Razin (2018) study a receiver-sender model in which the sender sends multiple messages through different channels in order to manipulate the receiver who fails to understand that the different messages are coming from the same sender. They characterize the strategic sender's optimal strategy. In our model, the receiver is fully rational and aware of the information structure from which the messages are originating.

Lipnowski, Ravid and Shishkin (2018) extend Kamenica and Gentzkow (2011)'s Bayesian persuasion game by assuming that once the outcome of the experiment is privately revealed to the sender, it is then decided by nature whether the sender must reveal the experiment's result as is, or is free to report any message he chooses. Assuming all of these messages are common knowledge, they study the properties of the optimal experiment.

Finally, Laouenan and Rathelot (2018) use data from an online marketplace of vacation rentals (Airbnb) to measure discrimination against ethnic minority hosts. They find that an additional review helps to close the gap in price between minority hosts relative to majority hosts. This is consistent with the results of our model which predict that, in expectation, an additional review incrementally corrects for the erroneous beliefs.

3 One-Period Model

We start with the case of two players: a sender (S) and a receiver (R). (The results obtained in this case will be useful in the more general case). 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, θ, player S (but not the receiver R) receives a signal, $\tilde{x}$, which takes a value in $[0, 1]$ according to the density $t_\theta(x)$ and the CDF $T_\theta(x)$. Define $\bar{x}$ to be the unique (neutral) signal for which

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

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

Assumption A.2 (hereafter referred to as MLRP) captures the idea that the higher the signal is the more likely 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.[3] A mixed strategy f for Fake is a probability measure over the set of messages $M = [0,1]$.[4] 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.

4 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 $\hat{p}(m \mid f)$ denote the receiver's posterior given that she observes the message m and believes that Fake uses the strategy f and let

$$\hat{P}(m \mid f) \equiv \frac{\hat{p}(m \mid f)}{1 - \hat{p}(m \mid f)}.$$

[3] Below we will allow for various types of fake senders, i.e. some whose objective is to increase $R's$ posterior and some whose objective is to decrease it.

[4] Formally, a mixed strategy $\tilde{f}$ is a mapping from $X = [0,1]$, the set of signals, to Δ, the set of probability measures over the set of messages $M = [0,1]$. However, it can be shown that in equilibrium Fake's strategy must be independent of the signal that he observes. To see this, observe that if Fake's payoff from $\tilde{f}(x')$ is higher than his payoff from $\tilde{f}(x'')$ then when the signal is x'' he can benefit from deviating to $\tilde{f}(x')$. Thus, without loss of generality we can assume that Fake's strategy is independent of the signal and is simply a probability measure over the set M.

Then:

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

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 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 $f(m) = 0$

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

- Neutral news always implies no updating

$$\hat{P}(\bar{x} \mid f) = P$$

- Inside the support of f, namely, when $f(m) > 0$

$$\begin{aligned} m > \bar{x} &\implies \hat{P}(m \mid f) < P \frac{t_1(m)}{t_0(m)} \\ m < \bar{x} &\implies \hat{P}(m \mid f) > P \frac{t_1(m)}{t_0(m)}. \end{aligned}$$

5 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 is $P \frac{t_1(z)}{t_0(z)}$.*

Proof. *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.

In the following claim, we argue that if $f^*$is an equilibrium, then $f^*(1) > 0$.

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

Proof. *Assume, by contradiction, that* $f^*(1) = 0$ *and thus* $\hat{P}(1 \mid f^*) = P\frac{t_1(1)}{t_0(1)} > \hat{P}(m \mid f^*)$ *for all* $m < 1$. *The last inequality follows from the assumed MLRP of* $t_\theta(\cdot)$. *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 and start by showing that in equilibrium $f^*(m) > 0$ only if $m > \bar{x}$.

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*

$$P\frac{Qt_1(m') + f^*(m')}{Qt_0(m') + f^*(m')} = P\frac{Qt_1(1) + f^*(1)}{Qt_0(1) + f^*(1)} \tag{3}$$

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 $\hat{P}(m" \mid f^*) = P\frac{t_1(m")}{t_0(m")} > P\frac{Qt_1(m')+f^*(m')}{Qt_0(m')+f^*(m')} = \hat{P}(m' \mid f^*)$, 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 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 $m \in [z, 1]$,*

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

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 part (i) above $P\frac{t_1(z)}{t_0(z)}$ 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 for all $m > z$, $f^*(m) > 0$ 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 $m \geq z$*,* $\hat{P}(m \mid f^*) = P\frac{t_1(z)}{t_0(z)}$*.* ■

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

Theorem 1 *An equilibrium exists and is unique. Fake's equilibrium strategy is given by:*

$$\begin{aligned} f^*(m) &= 0 & \text{for: } m \in [0, z] \\ f^*(m) &= Q\frac{t_0(z)\, t_1(m) - t_1(z)\, t_0(m)}{t_1(z) - t_0(z)} & \text{for: } m \in (z, 1] \end{aligned} \tag{4}$$

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

$$\int_z^1 f^*(m)dm = 1 \tag{5}$$

Proof. 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 $m \in (z, 1]$, $\hat{P}(m \mid f^*) = P\frac{Qt_1(m)+f^*(m)}{Qt_0(m)+f^*(m)} = P\frac{t_1(z)}{t_0(z)}$. 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 $m \in [0, 1]$ 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 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 $\Psi(z) = 1$. ■

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, $\hat{P}(z \mid f^*) = P\frac{t_1(z)}{t_0(z)} > P$. It is also apparent from 4 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.*

$$t_1(x) = 2x, \;\; t_0(x) = 2(1-x), \;\; x \in [0,1] \tag{6}$$

Plugging into 4 and solving, the Fake's equilibrium strategy becomes

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

where

$$z = \frac{1-\sqrt{1-q}}{q}.$$

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

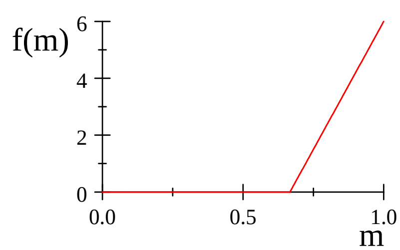

Figure 1

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

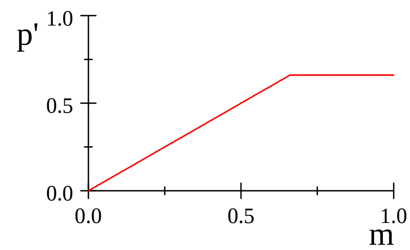

Figure 2

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 $\tau_0\,(m) = qt_0(m) + (1-q)f^*(m)$ (red) compared to the honest-sender distribution $t_0\,(m)$ (dashed):

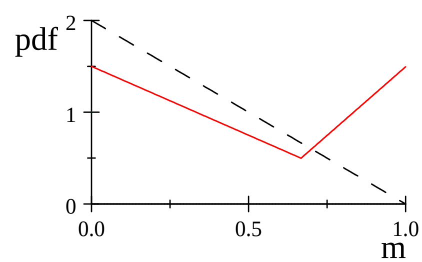

Figure 3

The following result, hereafter referred to as the *independence* result, follows immediately from 4 and it will be essential in the subsequent development of the model.

Corollary 2 ***Independence:*** *Fake's equilibrium strategy is independent of 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 muti-period multi-sender game.

We next present an *aggregation* 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 "right" direction in expectation. In order to prove the 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 $E_\theta[\hat{p}_{f^*}]$ denote the receiver's expected posterior probability that the state is 1, given that the true state is θ.

Proposition 2 ***Aggregation:*** *$E_1[\hat{p}_{f^*}] > p$ and $E_0[\hat{p}_{f^*}] < p$.*[5]

Proof. *It suffices to show that $E_1[\hat{p}_{f^*}] > p$. Let*

$$\tau_\theta(m) = qt_\theta(m) + (1-q)f^*(m)$$

That is, $\tau_\theta(m)$ is the (expected) density of receiving the message m, given that the state is θ and given Fake's equilibrium strategy. Using Bayes' rule we know that:

$$\hat{p}_{f^*}(m) = \frac{\tau_1(m)p}{\tau_1(m)p + \tau_0(m)(1-p)} \tag{7}$$

Let

$$\bar{\tau}(m) \equiv \tau_1(m)p + \tau_0(m)(1-p)$$

Bayes' rule in (7) immediately implies:

- $\hat{p}_{f^*}(m) > p$ *if and only if* $\tau_1(m) > \bar{\tau}(m)$;
- $\hat{p}_{f^*}(m) = p$ *if and only if* $\tau_1(m) = \bar{\tau}(m)$;
- $\hat{p}_{f^*}(m) < p$ *if and only if* $\tau_1(m) < \bar{\tau}(m)$;

[5] This proposition can be extended to the case of many states. We would like to thank Sergiu Hart for suggesting a very elegant proof .

Therefore:

$$\int_m (\hat{p}_{f^*}(m) - p)(\tau_1(m) - \bar{\tau}(m))dm > 0 \tag{8}$$

where the strict inequality follows from the fact that given the properties of t_θ *and* f^**, there exists an* m *for which* $\hat{p}_{f^*}(m) = p$*.Inequality 8 can be written as*

$$\int_m (\hat{p}_{f^*}(m)\tau_1(m)dm - \int_m (\hat{p}_{f^*}(m)\bar{\tau}(m)dm - p\int_m \tau_1(m)dm + p\int_m \bar{\tau}(m)dm > 0 \tag{9}$$

since

$$\int_m \tau_1(m)dm = 1 \text{ and } \int_m \bar{\tau}(m)dm = 1.$$

We can write inequality 9 as

$$\int_m (\hat{p}_{f^*}(m)\tau_1(m)dm - \int_m (\hat{p}_{f^*}(m)\bar{\tau}(m)dm > 0$$

which can be written as

$$E[\hat{p}_{f^*}(m) \mid \theta = 1] - E\left[\hat{p}_{f^*}(m)\right] > 0$$

and since $E\left[\hat{p}_{f^*}(m)\right] = p$ *we get*

$$E[\hat{p}(m) \mid \theta = 1] - p > 0.$$

■

Thus, even when the probability of the sender being honest is very small the receiver's posterior moves towards the true state in expectation.

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

5.1 Two-Sided Faking

Up to this point, we have assumed that Fake's objective is to increase the receiver's belief that the state is one. We now allow Fake to 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 $q > 0$ and that he is 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 can show that S_f^1's equilibrium strategy is the same as that described above and S_f^0's equilibrium strategy is its mirror image. More specifically, 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 $1 - q_0$.

Example 2 *Consider the linear case discussed in Example 1. Figure 4 and 5 depict the Fake-0 (green) and Fake-1 (red) strategies, as well as the receiver's posterior as a function of the message he receives, for the parameters* ($q = \frac{9}{19}, q_1 = \frac{9}{19}, q_0 = \frac{1}{19}$)*:*

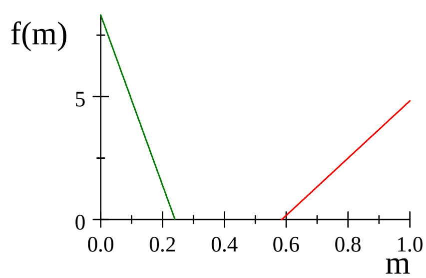

Figure 4

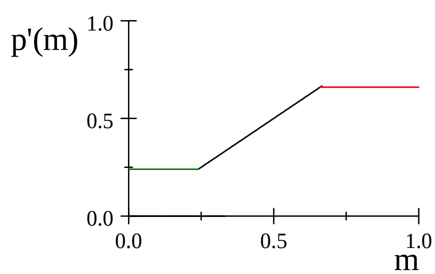

Figure 5

The following remark establishes that assuming that the honest sender simply reports his signal is without loss of generality.

Remark 1 ***Strategic "Honest" Sender.*** *If the honest sender is also strategic, then the equilibrium outcome described above will still be an equilibrium outcome. Namely, consider the case where the honest sender is replaced by a strategic player whose payoff decreases with the distance of the receiver's posterior from the sender's belief, given the signal he received and the prior. Assume, for simplicity, that there is only one sender and that he can either be 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) as a function of the signal received by the sender. There exists an equilibrium in which the "honest" sender truthfully reports his signal while the fake one uses the strategy* f^* *described in Theorem 4. To see this, assume that these are indeed the two strategies and hence the receiver's posterior likelihood ratio is given by* $\hat{P}(m \mid f^*)$ *as described in 2. 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.*

6 N-Period Model

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

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.

6.1 N periods with one fake sender, one honest sender and one receiver

There are N periods. The state of the world (which is the same throughout the entire 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 (for now) that there is only one receiver who forms her posterior about the state after observing the N messages. The fake sender's objective is to maximize the receiver's posterior belief, after the 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, σ_N specifies for each period $n \in N$ the mixed action of the fake sender *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 3 *σ_N^* is unique and $f_n^* \equiv f^*$ for all $n = 1, ..., N$.*

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

$$\hat{P}(m_1, ...m_n \mid \sigma_N^*) = \frac{Pr(\theta = 1 | (m_1, ...m_n), \sigma_N^*)}{Pr(\theta = 0 | (m_1, ...m_n), \sigma_N^*)} = P \frac{Pr(m_1, ...m_n | \theta = 1, \sigma_N^*)}{Pr(m_1, ...m_n | \theta = 0, \sigma_N^*)}. \quad (10)$$

Consider the last period $n = N$. Using our assumption that in every period the type *of the sender is independently drawn, we can rewrite 10 as:*

$$\begin{aligned}\hat{P}\left(m_1,...m_N \mid \sigma_N^*\right) &= P\frac{Pr\left(m_N|\theta=1,\sigma_N^*,(m_1,...,m_{N-1})\right)}{Pr(m_N|\theta=0,\sigma_N^*,(m_1,...,m_{N-1}))}\frac{Pr\left(m_1,...,m_{N-1}|\theta=1,\sigma_N^*\right)}{Pr\left(m_1,...,m_{N-1}|\theta=0,\sigma_N^*\right)} \\ &= P\frac{qt_1(m_N)+(1-q)f_N^*\left(m_1,...,m_N\right)}{qt_0(m_N)+(1-q)f_N^*\left(m_1,...,m_N\right)}\frac{Pr\left(m_1,...,m_{N-1}|\theta=1,\sigma_N^*\right)}{Pr\left(m_1,...,m_{N-1}|\theta=0,\sigma_N^*\right)}.\end{aligned} \tag{11}$$

With (11) 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 $(m_1,...,m_{N-1})$, $f_N^*\left(m_1,...,m_{N-1},1\right)>0$. *Also for any message* $\tilde{m}_N \neq 1$ *such that* $f_N^*\left(m_1,...,\tilde{m}_N\right)>0$ *it must be that*

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

which implies that

$$\frac{qt_1(\tilde{m}_N)+(1-q)f_N^*\left(m_1,...,m_{N-1},\tilde{m}_N\right)}{qt_0(\tilde{m}_N)+(1-q)f_N^*\left(m_1,...,m_{N-1},\tilde{m}_N\right)}=\frac{qt_1(1)+(1-q)f_N^*\left(m_1,...,m_{N-1},1\right)}{qt_0(1)+(1-q)f_N^*\left(m_1,...,m_{N-1,}1\right)}. \tag{12}$$

Note the similarity between equation (12) in Section 2 and equation (3) above. By following step by step the arguments in the proof of Theorem 4 one can easily establish that

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

for all $(m_1,...,m_{N-1})$ *. 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.* ■

6.2 The general N-period case

We are now ready to present the main result which states that the one-period strategy of a fake sender coincides with his equilibrium move in every period in which he is drawn to send a message, under the following conditions:

a. There are many senders, some of whom are fake and share his objective, some of whom are fake and share the opposite objective, and some of whom are honest.

b. There is a finite number of periods and in every period one of the senders (who may have already moved before) is randomly chosen to send a message and he might not observe all the previous messages sent by other senders.

c. In every period, a new receiver forms her belief about the state of the world on the basis of her own prior and the (possibly partial) history of messages up to and including that period.

Formally, 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_θ.

In every period, sender l is selected with probability q_l, $l = 1, ..., L$ 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 $f_{\bar{q}^0}^*$ and $f_{\bar{q}^1}^*$ be, respectively, Fake-0 and Fake-1's equilibrium strategies in the two-sided one-period model, where, the sender is of type Fake-θ with probability $\bar{q}^\theta$ and he is honest with probability $1-\bar{q}^0-\bar{q}^1$. Let σ_N^{l*} 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 3 (hence the proof is omitted).

Proposition 4 *Let l be a fake sender. Then σ_N^{l*} is unique and for all $n = 1, ..., N$, $f_n^{l*} \equiv f_{\bar{q}^\theta}^*$ if l is type Fake-θ.*

In order to put Proposition 4 in 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 and which 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 4 fully characterizes his strategy.

It follows from Proposition 2 and 4 that in the general N-period model, the receivers' beliefs shift over time (in expectation) towards the true state of the world. The assumed stationarity of q_l, $l \in \{0, 1, h\}$ together with Proposition 4 imply that the posterior follows a Martingale sequence and hence, in the limit (as $N \to \infty$) the posterior is fully revealing. Namely, the posterior converges to the truth almost surely, that is:[6]

Proposition 5 *For any $q > 0$, if $\theta = \theta_T$ is the true state, then*

$$\Pr[\lim_{N\to\infty} \hat{p}(m_1, ..., m_N \mid \theta = \theta_T, \sigma_N^{0*}, \sigma_N^{1*}) = \theta_T] = 1$$

Fake senders manage to steer the belief their way to some extent every time they are chosen to send a message and therefore certainly slow down correct learning on average. While the proportion of fake reviews affects the speed of learning, asymptotic arrival at the truth is preserved as long as the proportion of true reviews is positive.

7 Conclusion

Nowadays online reviews are ubiquitous and have become an essential part of a consumer's everyday decisions. But the credibility of these reviews has been seriously

[6] We thank Phil Reny for suggesting a very simple proof for this proposition.

undermined by the incentives of businesses 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 that offer online reviews for nearly every existing product and service nowadays. They try to combat fake reviews by filtering them with word algorithms, though plenty of false positives and negatives remain as the fakers improve 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 looked more suspicious (see Luca and Zervas (2016) and the Economist, Oct. 22 2015).

The aim of the model we proposed is to faithfully represent the product review market above described. Namely, there is an ocean of reviews for any given product: any customer may read only some of them and is aware that the reviews 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 and/or develop different beliefs after seeing different sets of reviews. Despite the generality of this setup, the characterization of the fake reviewer strategy is unique and crucially does not depend on the specific details described above, but only on the true/fake review shares. This allows us to deliver general and robust predictions on information transmission in the presence of true and fake reviews: fake reviewers manage to persuade and shift beliefs individually, but only enough to slow down not to derail correct long-run information aggregation. Reading reviews is evidently costly and time consuming and therefore is it worthwhile? This depends among other things on 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 our model can also serve as a platform to study other scenarios or interactions where fake information transmission may take place. Consider, for example, a news outlet (such as a newspaper or radio station) or a blog, where the 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 her beliefs about the sender's type, which in turn influence 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 presented 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 remain for future research.

8 References

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

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

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

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

5. 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.

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

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

8. Levy G., I. Moreno de Barreda and R. Razin: "Persuasion with Correlation Neglect". Working paper.

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

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

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