Controlling Fake Reviews

Yuta Yasui

July 19, 2020

*"Fully ref**ded after RE**W*
If you are interested dm and comment"

— a post on Facebook

1 Introduction

Online markets are growing all over the world, and the role of the ratings on the platforms is getting more and more important.[1] At the same time, incentives for sellers to make fake reviews is also growing. Washington Post (Dwoskin and Timberg, 2018) reports that 50.7% of reviews for Bluetooth headphones, 58.2% of Bluetooth speakers, 55.6% of weight-loss pills, 67.0% of testosterone booster on Amazon are suspicious. How do the sellers make fake reviews. The sellers can find reviewers on SNS such as Facebook and offer them "free" samples in a form of refund after purchase and the review, so that the review is counted as a review with a verified purchase and reflected to the star rating.[2] Both platform and regulators prohibit such behavior. Especially the regulators attitude is getting more stringent. For instance, Federal Trade Commission (FTC) filed the first case against paid fake reviews, by CureEncapsulations on Amazon in 2019.[3] Platforms' reaction to fake reviews varies, but the regulators puts more pressure to the platforms to have more strict attitude against the fake reviews.[4]

However, the impact of the fake reviews on consumers on platform is not very clear. Consumers might not be fooled by fake reviews. For instance, in a standard work of Holmström (1999), the market can correctly anticipate the long-lived player's behavior and debias the signal. Or, customers might be able to elicit some information even from fake reviews. If only high quality sellers make

[1] Hollenbeck (2018); Hollenbeck et al. (2019) show that ratings work as a substitute of other form of advertisement or brand names, and this pattern is getting stronger over time in the hotel industry. Reimers and Waldfogel (2020) exhibit that the existence of star ratings has 15 times as the impact on consumer surplus as the professional reviews on New York Times. For the insitutional details and data analysis on platforms and ratings, see also Belleflamme and Peitz (2018)

[2] Amazon also shows ligitimate promotional reviews along with organic reviews, but weights on such reviews in the calculation is reportedly low. [reference to Amazon's terms of service] Therefore, the sellers still have incentives to make fake reviews through fake purchase.

[3] [Regulation in EU, UK]

[4] [reference]

fake reviews to boost their initial reputation, even the fake reviews can be a signal of good quality. Such a behavior might be possible if low quality will be revealed through the word-of-mouth and only high quality seller can reap benefits from the future sales as suggested by Nelson (1970;1974) in the context of advertising. [5]

In this paper, we examine a theoretical model where seller's sales is determined by the reputation level and the seller chooses the amount of fake reviews at each time. Consumers form a reputation based on the potentially boosted rating displayed on the platform. The platform can control how strictly it filters fake reviews and how much the rating reflects the information of the past feedback (ie, how fast the rating evolves). A key assumption in this paper is that it gets harder for the seller to make fake reviews when its reputation gets higher. In the main part, this is explained by the higher reimbursement to incentivized reviewers due to higher price.[6] This causes the dependence of the amount of fake reviews on the current rating/reputation level. Furthermore, it is shown that such dependence will be greater with more stringent censoring policy (Proposition 3). This causes an unwanted effect for the platform: the stringent filtering policy might decrease the informativeness of the rating on the platform. This is because the fake reviews depend more on the current reputation, but less on the underlying true quality.

The above discussion is based on the assumption of the rational consumers, who knows the seller's strategy. However, the regulator's concern is not necessarily such a sophisticated consumers, but more naive consumers, who is vulnerable to fake reviews.[7] In this paper, we also incorporate such consumers, and show how much they are biased due to the fake reviews by the sellers. Even though the relationship between the bias and the censorship policy is not monotone in general, the stringent censorship reduces the naive consumer's bias in a reasonable range of parameters.

Thus, the regulator might face a trade-off between the precision of the information for the rational consumers and the bias which the credulous consumers suffer from. This paper provides a framework to analyze such a trade-off.

2 Literature Review

This paper mainly contributes two lines of the literature: rating design and signaling through promotion. The literature of the rating design can be divided to two strands: (i) how to reveal the known quality levels or an estimated quality index (i.e. whether to reveal the full information or add noise/coarsen the information), and (ii) how to make the index of unknown quality based on the multiple sources of information on the performance of the player.

[5]With a different perspective of the issue of fake reviews, Ananthakrishnan et al. (2020) also show that the consumers form more trust on the platform if it shows the fake reviews with flags indicating them as fake reviews, rather than deleting them from the platform.

[6]We can see the interaction between fake reviews and reputation more commonly. For instance, fake reviews might be crowded out if the seller receives many organic feedback due to large demand caused by high reputation. Then, the effective fake review would be costly for such a seller.

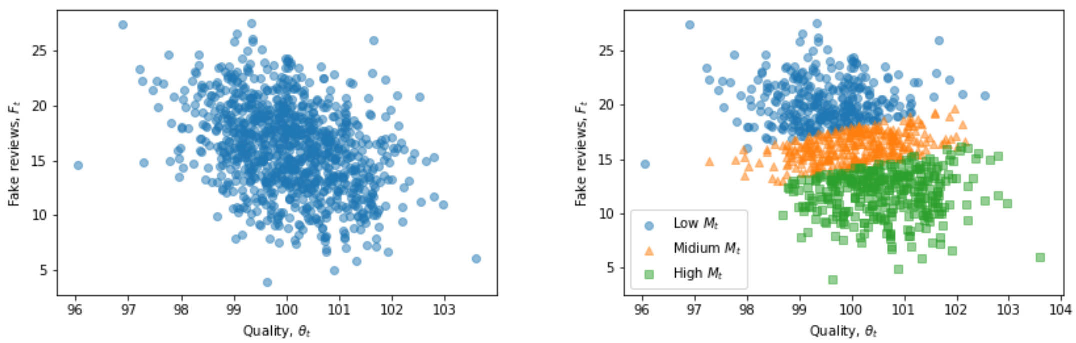
[7]For instance, Competition and Market Authority (CMA) in the U.K. says the role of review sites as "ensuring that consumers can trust the reviews they see" [reference]

The first strand is often framed in the context of certification such as Lizzeri (1999), Ostrovsky and Schwarz (2009), Boleslavsky and Cotton (2015), Harbaugh and Rasmusen (2018), Hopenhayn and Saeedi (2019), Hui et al. (2018). Some models are made tractable by the representation with posterior distribution in the line of Bayesian persuasion proliferated by Rayo and Segal (2010) and Kamenica and Gentzkow (2011). Saeedi and Shourideh (2020) extend the framework where the quality is endogenously chosen by the seller rather than the exogenous variable.

This paper fits in another strand analyzing how to aggregate the player's actions into a single index. In a one-shot model, Ball (2019) analyzes the optimal way to aggregate the various sources of potentially manipulated signals. In a dynamic setting based on Holmström (1999)'s signal jamming/career concern model, Hörner and Lambert (2018) show that the effort level of the long-lived player is maximized by a rating linear in the past observations. Vellodi (2020) analyzes the impact of the rating on the entry/exit behavior of the firm, and derive an optimal rating preventing high-quality seller exiting from the market due to reputation trap of failing to accumulate good reputation because fo initial bad luck. Bonatti and Cisternas (2019) examine a long-lived consumer's Ratchet effect. The consumers try to hide its willingness to pay to avoid the personalized pricing by short-lived monopolist, so that the consumption does not perfectly reflect their willingness to pay. Like as Hörner and Lambert (2018) and Bonatti and Cisternas (2019), this paper examines a relationship between a signal-jamming structure and a linear rating system. In contrast to Hörner and Lambert (2018), the equilibrium strategy is dependent on the hidden quality and the reputation so that the seller's strategy changes the informativeness of the rating on equilibrium path as in Bonatti and Cisternas (2019). In contrast to Bonatti and Cisternas (2019), where the effect of the manipulation is totally endogenously determined via the second order belief of the short-lived player, in this paper, the platform controls the effectiveness of the manipulation so that we can analyze the impact of the censorship by the platform. Besides, this paper depart from the literature by analyzing the impact of the manipulation on naive/credulous consumers, which is often the concern of the regulators.

This paper also contributes the literature on the promotion and signaling. Nelson (1970, 1974) argues that even if the promotion does not have any intrinsic information, the fact of burning money itself can be a signal of a good quality because such a signal pays off only for high quality type through the repeated purchase in the future. This idea is formalized later by Kihlstrom and Riordan (1984), Milgrom and Roberts (1986a) and many others as separating equilibria in signaling models. Via a one-shot signal jamming framework rather than a signaling model, Mayzlin (2006) showed a negative relationship between a promotion through fake reviews and quality, and Dellarocas (2006) generalizes conditions for the positive/negative correlation in one-shot signal jamming model. Bar-isaac and Deb (2014) examine effects of vertically/horizontally heterogeneous preference, and Grunewald and Kräkel (2017) examine the effect of competition between firms Most of the research on the signaling role of the promotion are based on models with one-shot promotion. An exception is Horstmann and MacDonald (1994), where the experience of the product is an imperfect signal of the quality and the signaling via the advertising is done only after establishing some reputation so

Figure 1: A simulated distribution of quality levels and the amount of fake reviews



The left panel show that the the amount of the fake reviews is negatively correlated with the quality level, unconditional on the level of reputation. On the other hand, the right panel shows that the amount of the fake reviews is increasing in the quality level, conditional on the reputation level.

that it is hard for the low-quality seller to mimic the high-quality seller's behavior.[8]. In this paper, I examine a dynamic signal-jamming model, where a reputational concern is a driving force of the positive correlation between the quality and the promotion. It also generate a non-degenerate dynamics consistent with an observation by Luca and Zervas (2016) that strategic manipulation increases after a drop of reputation.

The dependence of fake reviews on the reputation also gives some implication to the empirical literature on the signaling promotion. The literature have had a weak support on the correlation of the quality and the promotion. For instance, Kwoka (1984) observes that optimetrists with more advertisement provide less thorough eye examination, and Horstmann and Moorthy (2003) observe that the advertising is hump-shaped in quality among restaurants in Now York. Recently, Sahni and Nair (2019) implement a quasi-experiment to isolate the intrinsic information and the signaling effect of burning money and show that the consumer positively respond to the money burning. They pointed out that it is difficult to show the relationship between the quality and the promotion level since it is hard to obtain a reliable measure of quality. This paper emphasizes this point. The reputation-based index such as the customer rating can be a bad proxy for the underlying quality. The reputation level and the underlying quality level have the opposite impact on the promotion level in the equilibrium. Furthermore, even if the true quality is measured somehow, it is important to control reputation level. As shown in fig. 1 , the level of promotion and the true quality can be negatively correlated without conditioning on the reputation level, even though the quality and the promotion has positive relationship, *ceteris paribus*.

.

[8]Aside from the context of the rating system or the signaling promotion, Grugov and Troya-Martinez (2019) examine the biasing behavior of the seller in a model a. la. Holmström (1999) incorporating a detection rule and credulous consumers, and show that the biasing behavior increases as the authority requires stricter rule and the share of credulous consumer increases.

3 Rating Design for Rational Consumers

In this paper, we examine both models with rational consumers and with naive consumers. In this section, we first introduce a baseline model with a mass of rational consumers. The consumers rationally expect that the long-lived seller makes fake reviews following a linear strategy. However, they cannot expect the seller's action at time t precisely, due to the hidden quality.

Then, in the next section, we introduce a market with naive consumers, who does not expect any fake reviews, while the seller still makes fake reviews. In each model, we examine the impact of the platform's filtering/censoring policy on reviews, the weights on new/previous reviews, and the precision of the genuine reviews.

3.1 Model

The model is continuous time and infinite horizon, $t \in [0, \infty)$. At each instance t, a long-lived seller sells q units of its product, whose quality is denoted as θ_t, and makes F_t amount of fake reviews. A sufficiently large mass, n, of consumers forms a demand function such that the price $p_t = E[\theta_t|Y_t] \equiv M_t$ clears the market, where Y_t is a rating of the product at time t.[9] In this paper, we interpret this pricing rule as a result of competition among heterogeneous consumers, to which we can easily introduce a mixture of rational and naive consumers in the next section.

Suppose that consumer $i \in [0, n]$ feels $u_{t,i} = \theta_t + \epsilon_{t,i} - p_t$ if the consumer buy the product, and 0 otherwise, where $\epsilon_{t,i}$ is identically and independently distributed. Then, given the rating shown on the platform, Y_t, the consumer will choose to purchase the product if and only if $E[\theta_t|Y_t] + \epsilon_{t,i} - p_t \geq 0$. Therefore, the demand function is expressed as $n \cdot (1 - F(p_t - M_t))$ where $F(\cdot)$ is a c.d.f. of the random variable $\epsilon_{t,i}$. By letting $n = 2q$ and assuming that $\epsilon_{t,i}$ is distributed symmetrically around zero. We obtain $p_t = M_t$ as the market clearing price.

The quality θ_t and the rating Y_t change over time. The quality θ_t follows an exogenous mean reverting process:

$$d\theta_t = (-\kappa\theta_t + \mu)\, dt + \sigma_\theta dZ_t^\theta \tag{1}$$

while the rating Y_t is characterized by the following differential equation:

$$dY_t = -\phi Y_t + d\xi_t \tag{2}$$

where $d\xi_t$ is defined as

$$d\xi_t = aF_t dt + bq\theta_t dt + \sqrt{bq}\sigma_\xi dZ_t^\xi \tag{3}$$

where a is the effectiveness of the fake review, b is the feedback rate from customers, μ is the mean of θ_t in the stationary distribution, σ_θ and σ_ξ govern the standard deviations of the disturbance. The transition of the rating Y_t is interpreted in discrete time analogue that the future rating Y_{t+dt} is a weighted sum of the new reviews $d\xi_t$ and the previous reviews Y_t with weights of 1 and $1 - \phi dt$, respectively. The new reviews consist of two components: "organic" reviews and remaining fake

[9] Saeedi (2019) showed that the reputation is the measure determinant of the price on eBay market.

reviews after filtering. The second and third term of eq. (3) corresponds to organic reviews. Higher quality tends to generate high reviews, and the information gets more precise when there is more feedback due to more transaction q or higher response rate b. The disturbance $\sigma_\xi dZ_t^\xi$ is caused by the heterogeneity of criteria among the customers.[10] The first term is the effect of the fake reviews. The seller tries to boost the average review through fake reviews, but some of them are detected by the platform and the remaining enters as $aF_t dt$. Thus, small a implies stringent censorship. The exogenous mean-reverting process of θ_t is understood as a resut of the competition on quality among sellers. Relative quality of the own product might go down due to the rise other sellers with even higher quality.

The seller's instantaneous payoff is defined as:

$$\pi_t = (1-\tau)\, p_t \,(q + F_t) - p_t \cdot F_t - \frac{c}{2} F_t^2$$

where τ is transaction fees imposed by the platform. The first term is the total revenue from all the transactions including the own fake reviews, and the second term is the reimbursement cost to the fake reviewers. The last term is expressing that making more fake reviews is harder. The seller might find it difficult to search incentivized reviewers through communities such as Facebook, or some fake review services ask the seller a higher price for fake reviews because making more fake reviews comes with higher risk of being detected by the platform. The long-lived seller maximizes its discounted present value by choosing $(F_t)_{t\geq 0}$.

The instantaneous profit becomes easier to compare with the previous research by rewriting it as

$$\pi_t = (1-\tau)\, M_t \cdot q - \tau M_t \cdot F_t - \frac{c}{2} F_t^2. \tag{4}$$

Without the second term in eq. (4), the model becomes effectively a special case of Hörner and Lambert (2018), which is based on Holmström (1999)'s signal jamming model. However, due to the existence of this term, the marginal cost of the manipulation depends on the current reputation level, so the equilibrium manipulation level depends on the current rating in contrast to Hörner and Lambert (2018) where the equilibrium action turns out to be state dependent. In stead of relying on the constant action, we apply the idea of Bonatti and Cisternas (2019) to focus on a linear strategy and Gaussian stationary distribution of (θ_t, Y_t). Then, the HJB equation gives a simple the quadratic value function, solved by guess-and-verify method. It is verified that as τ goes to zero, the equilibrium strategy becomes invariant to θ_t, Y_t, (and t). Thus, the the interaction between the current reputation and the current action is considered as a driving force of non-degenerate Markov equilibrium strategy. In this paper, this interaction between the reputation and the manipulation is derived from the reimbursement to the incentivized reviewers, however, such an interaction can be more commonly observed in the context of fake reviews. For instance, if the reputation is high, then large demand can crowd out fake reviews, so that the effective fake reviews can be more costly

[10] In this paper, the mechanism behind the customer feedback is abstracted, assumed that the fixed portion of consumers keep reviewing. For detailed analysis on the customer feedback, see Chevalier et al. (2018) and the literature cited in it. They analyze the relationship with managerial responses to reviews.

given the high reputation. In the appendix, an alternative model with such an interpretation is discussed.

Definition of the Equilibrium As mentioned above, we focus on a linear Markov strategy equilibria, where a linear Markov strategy maximizes the seller's discounted present value among any admissible strategies.

A linear strategy (in θ_t and Y_t) is defined as:

$$F_t = \hat{\alpha}\theta_t + \hat{\beta}Y_t + \hat{\gamma}$$

Note that θ_t does not directly appear in the instantaneous payoff function, but it appears in the transition of payoff relevant state variable, Y_t. Thus, the seller is potentially sensitive to the level of θ_t. Now the equilibrium is defined as follows:

Definition 1. A linear Markov strategy $F = (F_t)_{t \geq 0}$ s.t. $F_t = \hat{\alpha}\theta_t + \hat{\beta}Y_t + \hat{\gamma}$ is a stationary Gaussian linear Markov equilibrium if

1. $F = \arg\max_{\left(\tilde{F}_t\right)_{t\geq 0}} E_0\left[\int_0^\infty e^{-tr}\pi_t\right]$ where $\left(\tilde{F}_t\right)_{t\geq 0}$ is admissible,
2. $M_t = E\left[\theta_t | Y_t\right]$, and
3. $(\theta_t,\, Y_t)_{t\geq 0}$ induced by F is stationary Gaussian

We don't know that $(\theta_t,\, Y_t)_{t\geq 0}$ is stationary or Gaussian *ex ante* because Y_t is endogenously determined by F_t, however, the the condition for $(\theta_t,\, Y_t)_{t\geq 0}$ to be stationary Gaussian is simply characterized by an inequality, similarly to Bonatti and Cisternas (2019). By eqs. (2) and (3), and the definition of the linear strategy,

$$\begin{aligned} dY_t &= -\phi Y_t dt + aF_t dt + bq\theta_t dt + \sqrt{bq}\sigma_\xi dZ_t^\xi \\ &= -\left(\phi - a\hat{\beta}\right)Y_t dt + \left(a\hat{\alpha} + bq\right)\theta_t dt + a\delta\mu dt + \sqrt{bq}\sigma_\xi dZ_t^\xi \end{aligned} \tag{5}$$

Thus, an inequality $\phi - a\hat{\beta} > 0$ must hold for the $(\theta_t,\, Y_t)_{t\geq 0}$ to have the stationary distribution. (Otherwise, the process of Y_t diverges.) When $(\theta_t,\, Y_t)$ is stationary Gaussian, by the projection theorem on the Gaussian distribution,

$$M_t \equiv E\left[\theta_t | Y_t\right] = E\left[\theta_t\right] + \frac{Cov\left(\theta_t,\, Y_t\right)}{Var\left(Y_t\right)}\left[Y_t - E\left[Y_t\right]\right] \tag{6}$$

. Furthermore, if it's stationary, all the expectations in eq.(6) are constants. By letting $\lambda \equiv \frac{Cov(\theta_t,\, Y_t)}{Var(Y_t)}$ and $\nu \equiv E\left[Y_t\right]$ (and $\mu = E\left[\theta_t\right]$ by construction), eq.(6) is written as $M_t = \mu + \lambda[Y_t - \nu]$. In the following part of this section, we use M_t instead of Y_t as a state variable for the sake of expositional simplicity. Then, the linear strategy is redefined as

$$F_t = \alpha\theta_t + \beta M_t + \delta\mu$$

The stationary condition is summarized as follows:

Lemma 1. *(Stationality and the characterization of the long-run moments) Suppose* $F_t = \alpha\theta_t + \beta M_t + \delta\mu$ *where* $M_t \equiv E[\theta_t | Y_t]$ *for all* $t \geq 0$*. Then, a process* $(\theta_t, Y_t)_{t\geq 0}$ *is stationary Gaussian if and only if*

i. $M_t = \mu + \lambda [Y_t - \nu]$ *for all* t

ii. $a\lambda\beta - \phi < 0$*, and*

iii. $(\theta_0, Y_0)' \sim \mathcal{N}([\mu, \nu]', \Gamma)$ *is independent of* $\left(Z_t^\theta, Z_t^\xi\right)_{t\geq 0}$ *where* Γ *is the variance-covariance matrix in the stationary distribution.*

The third condition is required so that the game starts from stationary distribution. Now, HJB equation is simply written by using Ito's lemma:

$$\begin{aligned} rV(\theta, M) = & \sup_{F\in\mathbb{R}} (1-\tau) M \cdot q - \tau M \cdot F - \frac{c}{2}F^2 \\ & - \kappa(\theta - \mu) V_\theta \\ & + \left\{a\lambda F + bq\lambda\theta - \phi\left[M - \bar{\theta} + \lambda\bar{Y}\right]\right\} V_M \\ & + \frac{\sigma_\theta^2}{2} V_{\theta\theta} \\ & + \frac{bq\lambda^2\sigma_\xi^2}{2} V_{MM} \end{aligned} \tag{7}$$

By guessing the quadratic form of the value function, $V = v_0 + v_1\theta + v_2 M + v_3\theta^2 + v_4 M^2 + v_5\theta M$, and the linear strategy, we can verify the existence and uniqueness of the value fiction and the linear strategy via the matching coefficient.

3.2 Equilibrium Characterization

The following theorem shows that the equilibrium strategy exists and it is unique under a reasonable range of parameters.

Theorem 1 (Existence and uniqueness)**.** *There always exits a stationary linear Markov equilibrium. For any equilibrium,* $\alpha > 0$*,* $\beta \in \left(-\frac{\tau}{c}, 0\right)$*,* $\lambda > 0$ *and* $L > 0$ *hold. Furthermore, if* $h'(L) < 0$ *holds, then such an equilibrium is unique and the equilibrium strategy* α*,* β*,* δ *is differentiable in parameters.*

$h'(L) < 0$ holds for any $L > 0$ if $6\kappa\phi + 4r^2 + 2\kappa r + 17r\phi + 19\phi^2 > \kappa^2$.

Note that $6\kappa\phi + 4r^2 + 2\kappa r + 17r\phi + 19\phi^2 > \kappa^2$ is a loose and reasonable condition. ϕ is a transition speed of the rating and κ is the transition speed of the quality. This is reasonable as long as the rating system is meant to help estimating the current quality rather than the previous quality. For instance, even if the true quality does not drift much (ie, $\kappa \simeq 0$), the rating should drift to reflect new information about the unknown quality (ie, $\phi > 0$).

Intuition of the Equilibrium Strategy In Theorem 1, it is shown that high-quality type makes more fake reviews ($\alpha > 0$), conditional on its reputation level. and high reputation type makes less fake reviews ($\beta < 0$) conditional on the quality type. Given the logic of Nelson (1970; 1074), $\alpha > 0$ (and $\beta < 0$) might look intuitive but the reasons of such signs are different from those in the previous research.

I start from the negative β. From the first order condition, the optimal strategy is expressed as

$$F_t = -\tau M + a\lambda \underbrace{\{v_2 + 2M_t v_4 + \theta v_5\}}_{=V_M}$$

Then, $\beta = -\tau + 2a\lambda v_4$, furthermore, the envelop condition gives an expression for v_4 so that it is rewritten as $\beta = -\tau - \lambda \frac{a\beta\tau}{(-a\beta\lambda+r+2\phi)}$. The first term comes from the interaction of the reputation level and the fake reviews in the cost term, $\tau M_t F_t$. If the reputation is high, then the marginal cost of fake review is high. Therefore, the seller will make less fake reviews given the higher reputation. The second term is corresponding to the fake review's the marginal benefit in the future. Given the equilibrium strategy, $v_4 = -\frac{\beta\tau}{2(-a\beta\lambda+r+2\phi)}$ is positive, meaning that the marginal benefit in the future is increasing in the reputation. This is because the future self will reduce the amount of the fake reviews after observing the boosted reputation due to today's fake reviews. Furthermore, this effect is increasing in M_t, because the future reputation M_{t+dt} tends to be high given high M_t, so the reduction of F_{t+dt} reduces $\tau M_{t+dt} F_{t+dt}$ a lot. It turns out that the first term dominates the second term so that β is negative.

The intuition of the positive α comes from similar logic with the counteracting effect in the last paragraph. From the first order condition, $\alpha = a\lambda v_5$, and the envelop condition gives us $v_5 = \frac{1}{\kappa+r+\phi}\{2(a\alpha + bq)\lambda v_4 - \alpha\tau\}$. Note that $2(a\alpha + bq)\lambda v_4$ captures the effect that the boosted reputation stays high for a long time given high quality. The cost reduction explained in the last paragraph will be in effect for a long time if θ_t is high. This gives the positive incentive for the seller to make more reviews given high θ_t. On the other hand,$-\alpha\tau$ states that such an incentive is attenuated because the quality in the near future θ_{t+dt} tends to be high given high θ_t, and the future self produce high F_{t+dt} resulting in high cost.

In summary, the driving force of $\beta < 0$ is the incentive to reduce $\tau M_t F_t$ today given high M_t. α is positive because the boosted reputation will stay long given high θ_t and it will reduce the cost in the future. Readers might wonder why the logic of Nelson (1970; 1974) does not apply here. If θ_t is high, the boosted revenue would stay high for a long time, but in this model, such a product would eventually achieve high reputation through the organic feedback even without fake reviews. Therefore, the *marginal future revenue* $\frac{dp_s}{dF_t}$ $(s \geq t)$ is independent of θ_t. It is worth noting that the same intuition applies even in a variant of the model with a fixed price p and time-varying quantity

q_t discussed in the appendix.

3.2.1 Properties of the equilibrium

Before examining normative comparative statics of the equilibrium, we check some positive properties of the equilibrium.

First, the expected amount of the fake reviews is increasing in a. This is simply because the marginal benefit of the fake review in the future would increase if the platform loosen the censorship policy. The model does not guarantee the positive amount of the fake review, but it is also shown that the expected amount of the fake review is positive under some parameters.

Proposition 1. *$E[F_t]$ is increasing in L, furthermore, increasing in a. Furthermore, $E[F_t] \geq 0$ holds for sufficiently large a.*

Thus, the model can represent reasonable situation given some parameters where the fake review has non-trivial effect. There remains a small probability that F_t becomes negative, but the model can approximate a reasonable distribution of the fake reviews as shown in fig. 1.

The precision of "organic" feedback from normal customers also monotonically changes the expected amount of the fake reviews. When the organic feedback from customers disperses a lot, it is hard for the seller to manipulate the reputation since a boosted rating is attributed to a large variation in the feedback.

Proposition 2. *$E[F_t]$ is decreasing in $\left(\frac{\sigma_\xi}{\sigma_\theta}\right)$.*

Even though a stringent policy decrease the expected amount of the fake reviews as shown in Proposition 1, it does not imply that the seller's strategy gets closer to no-fake strategy of $\{\alpha, \beta, \delta\} = \{0, , 0, 0\}$. Moreover, the stringent policy might have unintentional effects.

Proposition 3. *$|\alpha|$ is increasing in $\frac{\tau}{c}$ and decreasing in $\frac{\sigma_\xi}{\sigma_\theta}$. $|\beta|$ is decreasing in a and increasing in $\left(\frac{\sigma_\xi}{\sigma_\theta}\right)$.*

For small a by a stringent policy, the marginal benefit of the fake review decreases, but at the same time, the dependence of the marginal benefit on the current reputation also decreases. Thus, the second term of $\beta = -\tau - \lambda \frac{a\beta\tau}{(-a\beta\lambda + r + 2\phi)}$ decreases while the marginal cost still depends on the current reputation regardless of the censoring policy. Therefore, $|\beta|$ gets larger due to the smaller attenuation effect.

In the proof, the intensity of dynamic consideration is captured by an aggregater $L = -a\lambda\beta$, which is the equilibrium effect on the reputation transition speed. L becomes smaller when the dynamic incentive becomes smaller, so the α which only comes from the future marginal benefit becomes smaller, and $|\beta|$, to which the future marginal benefit only works as a counteracting effect, becomes greater because the present cost reduction incentive prevails. L is shown to be increasing in $\frac{a\tau}{c}$ and decreasing in $\frac{\sigma_\xi}{\sigma_\theta}$.

Lemma 2. *L at the equilibrium is increasing in $\frac{a\tau}{c}$ and decreasing in $\frac{\sigma_\xi}{\sigma_\theta}$. Furthermore, $L \to 0$ as $\frac{a\tau}{c} \to 0$ and $L \to \infty$ as $\frac{a\tau}{c} \to \infty$.*

This concludes Proposition 3. α is not necessarily increasing in a because α is a function in a and L, so the change in a impact directly, and indirectly via L, and the net impact is not clear. $|\beta|$ is not necessarily decreasing in $\frac{\tau}{c}$ by the analogous reason.

Proposition 3 implies less signaling (smaller α) and more noise (greater $|\beta|$) when the aggregater on the strategic effect L is smaller. This suggests small information from the rating system when the strategic effect L is small. In the following section, we formally examine this effect.

3.3 Optimal Rating System for Rational Consumers

Since the rational customers can form an unbiased estimate from any current rating, $M_t = E\left[\theta_t | Y_t\right]$, the informativeness of the rating is defined by the precision of the customer's estimate on the quality. Due to the normality assumption, This is rewritten as $Var\left(\theta_t | Y_t\right) = Var\left(\theta_t\right)\left(1 - \rho^2\right)$ where ρ^2 is the correlation between θ_t and Y_t. Therefore, we use ρ^2 as the criteria on the informativeness of the rating.

Given an equilibrium strategy, the stochastic differential equation eqs. (1) and (5) gives us ρ^2 as a function of the parameters and the equilibrium strategy. Therefore, change of a parameter directly affects ρ^2 and indirectly affects through change of the equilibrium strategy. Fortunately, by representing equilibrium coefficients α, β as a function of the equilibrium aggregater $L = a\beta\lambda$, both the direct and indirect effects of the censorship (a) are expressed as an effect only through L. Comparative statics about other parameters such as ϕ and σ_ξ/σ_θ can be also expressed by the indirect effect through L and the direct effect.

Lemma 3. *When $(\alpha,\, \beta,\, \delta)$ is the equilibrium strategy, ρ^2 is expressed a function: $\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)$ on which a, c, τ affect only through L.*

3.3.1 Filtering/Censoring Reviews

First, we analyze the impact of filtering/censoring policy. Does the fake reviews damage the informativeness of the rating system compared to the case without fake reviews? Does the filtering/censoring the reviews increase the informativeness?

As a benchmark, we derive the informativeness *without* fake reviews. By construction, we can do so by letting $\alpha = \beta = \delta = 0$. [11] The same informativeness is also replicated by setting $L = 0$ in $\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)$, so that it gets easier to compare the informativeness at the equilibrium and at the no-fake benchmark.

Lemma 4. *$\rho^2\left(0;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)$ coincides with ρ^2 under the no-fake strategy.*

[11] Actually, δ does not enter in the formula for the informativeness, so $\delta = 0$ does not matter in terms of the informativeness.

Note that $L = 0$ does not necessarily mean $\alpha = \beta = \delta = 0$. For instance, L goes to 0 as a goes to 0, but at the same time β converges to some negative value. The lemma says that even such a situation, the informativeness is the same as the one without fake reviews. By combining with Lemma 2, this gives us the following proposition:

Proposition 4. *The informativeness of the rating system in the equilibrium converges to that of the "no-fake" strategy as* $\frac{a\tau}{c} \to 0$.

Thus, even though the equilibrium strategy at the limit of $\frac{a\tau}{c}$ is not necessarily the no-fake strategy, the informativeness converges to that of no-fake strategy.

Now, by analyzing the behavior of $\rho^2 (L; \phi, \kappa, \sigma_\xi, \sigma_\theta)$ with respect to L, we can conclude that the informativeness can be even higher under the equilibrium strategy where the positive amount of the fake reviews is expected. In other words, the stringent censorship can decrease the informativeness of the rating system.

Proposition 5. *The equilibrium strategy is more informative than no-fake strategy under a set of parameters such that*

1. $\frac{a\tau}{c}$ *is sufficiently large, or*
2. $\frac{a\tau}{c}$ *is sufficiently small and* $\phi^2 < \kappa^2 + \frac{\sigma_\theta^2}{\sigma_\xi^2}$

Fig. 2 exhibits the behavior of ρ^2. The behavior of ρ^2 around $L = 0$ is determined by the relative size of ϕ^2 and $(\kappa^2 + \sigma_\theta^2/\sigma_\xi^2)$. If $\phi^2 < \kappa^2 + \frac{\sigma_\theta^2}{\sigma_\xi^2}$, then ρ^2 is decreasing in L, which is increasing in $\frac{a\tau}{c}$, vice versa. ρ^2 converges to 1 as L goes to infinite. Thus, it surpasses ρ^2 of the no-fake benchmark at some point.[12]

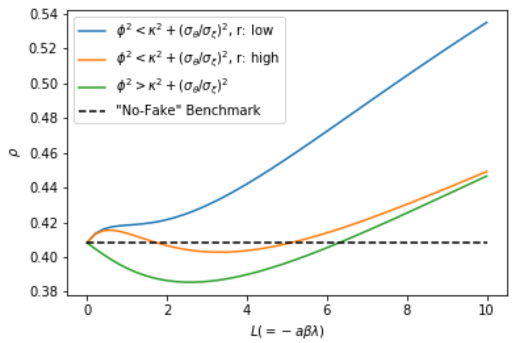
Intuition consists of two parts: (i) As mentioned in Subsection 3.2.1, the sensitivity to θ_t decreases and that of M_t increases as the strategic effect L decreases. Thus, the strict censorship policy, which reduces the equilibrium aggregater L, decreases signaling effect and increase the noise from the sensitivity to M_t. (ii) Note that $L > 0$ increase the effective transition speed of reputation to $\phi + L$. It can be good or bad for the informativeness, depending on the original transition speed, ϕ. To be more specific, the threshold of $\phi^2 = \kappa^2 + \frac{\sigma_\theta^2}{\sigma_\xi^2}$ is the informativeness-maximizing ϕ, given no fake reviews. Therefore, if ϕ is smaller than ϕ^0, the equilibrium effect L increases the informativeness, and vice versa. The first effect dominates for large L and the second effect dominates L close to zero.

3.3.2 Weights on New/Previous Reviews

Next, we analyze the optimal weights on the previous reviews. Again, the informativeness without fake reviews is expressed by $\rho^2 (0; \phi, \kappa, \sigma_\xi, \sigma_\theta)$. Therefore, the optimal weight at this benchmark is simply characterized by $\frac{\partial}{\partial \phi}\rho^2 (0; \phi, \kappa, \sigma_\xi, \sigma_\theta) = 0$. Let ϕ^0 be the solution of this equa-

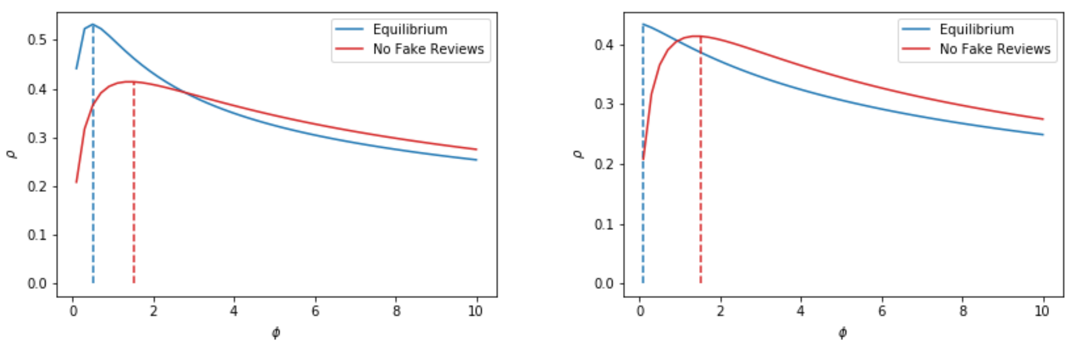
[12] Note that $E[F_t]$ is increasing in L (Proposition 1). Thus, the high informativeness is not due to negative fake reviews, but associated with the positive amount of fake reviews.

Figure 2: Change of the informativeness in the aggregater L



The graph indicates that the informativeness is (i) increasing in L if ϕ and r are relatively low, (ii) increasing in L around zero, then decreasing, and then increasing if ϕ is relatively low but r is relatively high, and (iii) decreasing in L around zero and then increasing in L if ϕ is relatively high. It also indicates the rating becomes more informative than the no-fake benchmark as L gets large.

Figure 3: Change of the informativeness in ϕ



The left panel shows change of the informativeness in ϕ when r is relatively low, while the right panel shows that of a relatively high r. The informativeness is maximized at a lower ϕ under the equilibrium than the maximizer under the no-fake benchmark.

tion.[13] On the other hand, at the equilibrium, ϕ changes the equilibrium aggregater L as well. Thus, the optimal weight at the equilibrium is characterized by $\frac{d\rho^2}{d\phi} = \frac{\partial}{\partial\phi}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) + \frac{\partial}{\partial L}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)\frac{dL}{d\phi} = 0$. Let the solution of this equation be ϕ^*. Now, we have the following proposition.

Proposition 6. $\frac{d\rho^2}{d\phi} < 0$ *at* $\phi = \phi^0$. *Furthermore, if* r *is sufficiently small, then* $\rho^2\left(L\left(\phi^*\right);\, \phi^*,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) > \rho^2\left(0;\, \phi^0,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)$.

The first part of the proposition states that the platform should reduce the speed of transition ϕ, the given the existence of the fake reviews. Intuitively, this is explained as follows: At the equilibrium, the transition of the rating score Y_t is $\phi + L$ where L is non-negative. Therefore, to

[13] ϕ corresponding to disaggregated information, ϕ^*, is an alternative benchmark as in Bonatti and Cisternas (2019). In this model, we obtain a mixed result for the comparison of ϕ^L and ϕ^*. See the appendix for more details.

cancel the strategic impact on the transition speed, the platform should decrease ϕ compared to the no-fake benchmark ϕ^0. The second part of the proposition is even more striking. If the seller is sufficiently concerned about the future, the platform can achieve higher informativeness level than no-fake review by adjusting the update speed of the rating.

3.3.3 Precision of the genuine reviews

Lastly, we examine the impact of the precision of the organic feedback, $\frac{\sigma_\xi}{\sigma_\theta}$. As discussed in Subsection 3.2.1, increasing $\frac{\sigma_\xi}{\sigma_\theta}$ and decreasing a have the similar effects on the equilibrium strategy. However, they depart in terms of the impacts on the informativeness. This is because a affects the informativeness only through the equilibrium aggregater L, but $\frac{\sigma_\xi}{\sigma_\theta}$ affects the informativeness directly as well. Intuitively, if the reviews consists of the less precise feed back (ie, higher $\frac{\sigma_\xi}{\sigma_\theta}$), the rating score is less informative about the quality. The indirect effect consists of two parts like as the comparative statics over a: (i) Higher $\frac{\sigma_\xi}{\sigma_\theta}$ implies smaller strategic effect L, which then implies less signaling and more noise. (ii) $L > 0$ increase the reputation transition to $\phi + L$. The following proposition shows that the direct effect and the first indirect effect dominate the second indirect effect for any parameters.

Proposition 7. *The informativeness at the equilibrium is decreasing in* $\frac{\sigma_\xi}{\sigma_\theta}$.

Thus, the precise organic feedback increases the informativeness even though it comes with more fake reviews.

4 Rating Design for Naive Consumers

The model with rational consumers is a standard starting point for any economic models, but in the context of customer reviews, it is natural to consider the impact on naive consumers who don't expect any fake reviews. The regulator often tries to protect customers from the fake reviews with an assumption that the fake reviews can fool or at least confuse consumers. In this section, we assume that some consumers do not expect any fake reviews on the platform. Such customers are modeled by by assuming that the reputation (and the price) is characterized as $\widetilde{M}_t = \mu + \widetilde{\lambda}\,[Y_t - \widetilde{\nu}]$ where $\widetilde{\lambda}$ and $\widetilde{\nu}$ are characterized by the stochastic differential equation eqs. (1) and (5) where $\alpha = \beta = \delta = a = 0$. On the other hand, the long-lived seller face the same problem as in the previous chapter except for the definition p_t.[14]

4.1 Model / Equilibrium Characterization

Let's assume that there is a mass of $\eta \cdot n$ rational consumers and $(1 - \eta) \cdot n$ naive consumers where $\eta \in [0, 1]$. Now, the price at time t is defined as p_t:

[14] **Note to be added**: Similarity to Milgrom and Roberts (1986b) RAND "Relying on the Information of Intersted Parties"]

$$
\begin{aligned}
p &= \eta M + (1-\eta)\widetilde{M} \\
&= \eta\left\{\mu + \lambda\left[Y_t - \nu\right]\right\} + (1-\eta)\left\{\mu + \lambda^{naive}\left[Y_t - \nu^{naive}\right]\right\} \\
&= \mu - \left(\eta\lambda\nu + (1-\eta)\lambda^{naive}\nu^{naive}\right) + \left(\eta\lambda + (1-\eta)\lambda^{naive}\right) Y_t
\end{aligned}
$$

This simple linear combination of the reputations is rationalized by the following assumptions. As in the previous section, suppose that consumer $i \in [0, n]$ feels $u_{t,i} = \theta_t + \epsilon_{t,i} - p_t$ if the consumer buy the product, and 0 otherwise, where $\epsilon_{t,i}$ is identically and independently distributed.

The rational consumers and the naive consumers differ only in terms of how they form the expectation based on the same observation of the rating Y_t. As in the previous section, a rational consumer purchase if and only if $M_t + \epsilon_i - p \geq 0$, while a naive consumer purchase the product if and only if $\widetilde{M_t} + \epsilon_i - p \geq 0$. Therefore, the total demand function is expressed as

$$
\eta \cdot n \cdot (1 - F(p - M)) + (1-\eta) \cdot n \cdot \left(1 - F\left(p - \widetilde{M}\right)\right)
$$

where $F(p)$ is a c.d.f. of the random variable ϵ_i. By letting $n = 2q$ and assuming that ϵ_i is distributed uniformly and symmetrically around zero. We obtain $p = \eta M + (1-\eta)\widetilde{M}$ to clear the market. In this section, we consider a linear strategy $F_t = \hat{\alpha}\theta_t + \hat{\beta}Y_t + \hat{\gamma}$ and the HJB equation with state variables θ and Y, since Y keeps track of both M and $\widetilde{M}$ in a simple manner:

$$
\begin{aligned}
rV(\theta, Y) = &\sup_{F \in \mathbb{R}} (1-\tau) p \cdot q - \tau p \cdot F - \frac{c}{2} F^2 \\
&- \kappa(\theta - \mu) V_\theta \\
&+ \left\{-\phi Y_t + aF_t dt + bq\theta_t\right\} V_Y \\
&+ \frac{\sigma_\theta^2}{2} V_{\theta\theta} \\
&+ \frac{b^2 q^2 \sigma_\xi^2}{2} V_{YY}
\end{aligned}
\tag{8}
$$

The following theorem states that, even with credulous consumers, we have the existence and the uniqueness given the same condition as the baseline model.

Theorem 2. *For any $\eta \in [0,1]$, there always exits a stationary linear Markov equilibrium. For any equilibrium, $\alpha > 0$, $\beta \in \left(-\frac{\tau}{c}, 0\right)$, $\lambda > 0$ and $L > 0$ hold. Furthermore, if $h'(L) < 0$ holds, then such an equilibrium is unique and the equilibrium strategy α, β, δ is differentiable in parameters.*

$h'(L) < 0$ holds for any $L > 0$ if $6\kappa\phi + 4r^2 + 2\kappa r + 17r\phi + 19\phi^2 > \kappa^2$..

In addition, somehow surprisingly, the existence of the naive consumers reduce the sellers strategic behavior.

Proposition 8. *The equilibrium with naive consumer ($\eta \in [0,1)$) generates smaller $|\alpha|$, larger*

$|\beta|$,*and smaller* $E[F_t]$ *compared to the equilibrium without naive consumers.*

This is because the naive consumers are less reactive to the change of ratings compared to the rational consumers. Therefore, the price is less responsive to the boost of the ratings, so the *marginal* benefit of the fake review is smaller with naive consumers. Thus, the seller makes less fake reviews in expectation.

Readers might wonder why the seller does not become more exploitative with naive consumers. It is simply because the fake review strategy against rational consumers generates more fake reviews by the different reasons than exploiting consumers. If only a small number of naive consumer exist and observe the ratings, which is essentially for rational consumers, the naive consumers would form even more biased estimate.

4.2 Optimal Rating System for Naive Consumers

Criteria: Bias in the Reputation In this part we evaluate the impact of fake reviews on naive consumers. To do so, we focus on the bias caused by the boosted rating:

$$\begin{aligned} bias &= E\left[\widetilde{M}_t - \theta_t\right] \\ &= E\left[\mu - \theta_t + \widetilde{\lambda}\left[Y_t - \widetilde{\nu}\right]\right] \\ &= \widetilde{\lambda}\left[\nu - \widetilde{\nu}\right] \end{aligned}$$

The above decomposition of the bias is intuitive: the positive bias is due to the boosted reputation. Since the naive consumers do not expect any fake reviews, they interpret a high rating as a result of the high quality even though it is actually the average level of the rating at the equilibrium (ie, $\nu > \nu^{naive}$). Thus, as long as the seller makes the positive amount of fake reviews (in expectation), the naive consumers are positively biased.

Lemma 5. *Bias* ≥ 0 *if and only if* $E[F_t] \geq 0$.

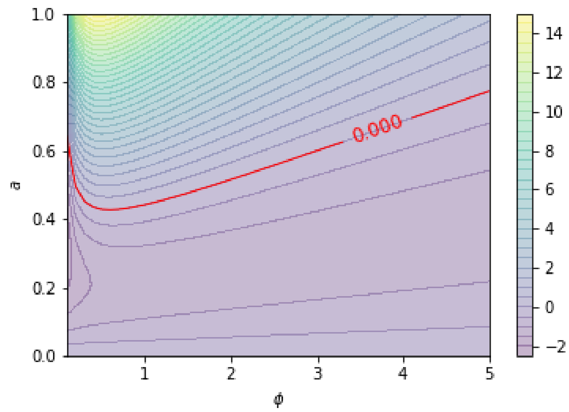
4.2.1 Filtering/Censoring Reviews

In the following part, for the sake of the tractability, I focus on the case of $\eta = 0$, where only the naive consumers exists in the market. Numerical exercises in cases of $\eta \in (0, 1)$ can be found in the Appendix.

First, we examine the impact of filtering which the regulator and the platform are concerned about. The following proposition provides a theoretical background of such a policy protecting the naive customers. Note that even though the statement seems pretty intuitive, it is not trivial since the model predicts non-monotone relationship between censorship and the bias in general. Fortunately, in a realistic range of parameters where we predict the positive amount of fake reviews, the stringent censorship will reduce the bias.

Proposition 9. *Suppose Bias* ≥ 0. *Then, Bias is increasing in* a.

Figure 4: Impact of censorship intensity and the weights of reviews on naive consumer's bias.



Combining with Lemma 5, it can be rephrased as following corollary:

Corollary 1. *The stringent censorship reduces the naive consumers' bias whenever the expected amount of the fake reviews is positive.*

4.2.2 Weights on New/Previous Reviews

As shown in fig. 4, the bias tends to be hump shaped in ϕ. This is intuitive because the fake review would be effective only when the rating is believed to be informative by the consumers so that the consumers react to the rating. Since the informativeness is hump shaped in ϕ, so is the bias caused by the fake reviews.

This suggest a trade off between the bias and the informativeness. Some readers might want a criteria integrating the bias and the informativeness. The mean squared error (MSE) is a natural candidate. It does not give us a clear-cut prediction, but a simulation about MSE is provided in the Appendix.

5 Summary

In this paper, effects of fake reviews on rational and credulous consumers are analyzed. The key assumption is that the high reputation causes high cost of the fake reviews. This is rationalized by the high reimbursement to reviewers or high demand crowding out the fake reviews.

In the positive analysis, we showed that the amount of the fake reviews is increasing in quality and decreasing in reputation level, which exhibit the difficulty of empirical analysis on the signaling promotion. The stringent censorship reduces the expected amount of fake reviews while it make the fake reviews more noisy.

This leads to the normative result that the rating under a less strict filtering policy can be more informative compared to the rating under a strict policy or the rating with no fake reviews. The dependence of the fake reviews on the current reputation also changes the transition speed of the rating/reputation. Since the equilibrium effect makes the transition faster, the transition speed should be adjusted downward to maximize the correlation between the quality and the rating.

The existence of credulous consumers decreases the expected amount of fake reviews since they are less responsive to the rating. On the other hand, they are vulnerable to fake reviews and pay more than the true quality in expectation. The model predict that, as long as the positive amount of the fake reviews is expected, the regulator or the platform can reduce such a biased behavior by enhancing the censorship.

Thus, the regulator or the platform face a trade-off between the precision of the informativeness and the bias caused by the fake reviews. This paper provides a framework to analyze such a trade-off.

References

Ananthakrishnan, Uttara M., Beibei Li, and Michael D. Smith, "A Tangled Web: Should Online Review Portals Display Fraudulent Reviews?," *Information Systems Research*, jun 2020, *Article in*, 1–22.

Ball, Ian, "Scoring Strategic Agents," 2019, (November), 1–57.

Bar-isaac, Heski and Joyee Deb, "What is a Good Reputation? Career Concerns with Heterogeneous Audiences," *International Journal of Industrial Organization*, 2014, *34*, 44–50.

Belleflamme, Paul and Martin Peitz, "Inside the Engine Room of Digital Platform: Reviews, Ratings, and Recommendations," in "Economic Analysis of the Digital Revolution" 2018, pp. 75–114.

Boleslavsky, Raphael and Christopher Cotton, "Grading standards and education quality," *American Economic Journal: Microeconomics*, 2015, *7* (2), 248–279.

Bonatti, Alessandro and Gonzalo Cisternas, "Consumer Scores and Price Discrimination," *Review of Economic Studies*, 2019.

Chevalier, Judith A., Yaniv Dover, and Dina Mayzlin, "Channels of Impact: User Reviews When Quality Is Dynamic and Managers Respond," *Marketing Science*, 2018, *37* (5), 688–709.

Dellarocas, Chrysanthos, "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science*, 2006, *52* (10), 1577–1593.

Dwoskin, Elizabeth and Craig Timberg, "How merchants use Facebook to flood Amazon with fake reviews," 2018.

Grugov, Mikhail and Marta Troya-Martinez, "Vague Lies and Lax Standards of Proof: On the Law and Economics," *Journal of Economics & Management Strategy*, 2019, *28*, 298–315.

Grunewald, Andreas and Matthias Kräkel, "Advertising as signal jamming," *International Journal of Industrial Organization*, 2017, *55*, 91–113.

Harbaugh, Rick and Eric Rasmusen, "Coarse Grades: Informing the Public by Withholding Information," *American Economic Journal: Microeconomics*, 2018, *10* (1), 210–235.

Hollenbeck, Brett, "Online Reputation Mechanisms and the Decreasing Value of Chain Affiliation," *Journal of Marketing Research*, oct 2018, *55* (5), 636–654.

_ , Sridhar Moorthy, and Davide Proserpio, "Advertising Strategy in the Presence of Reviews: An Empirical Analysis," *Marketing Science*, 2019, *38* (5), 793–811.

Holmström, Bengt, "Managerial Incentive Problems: A Dynamic Perspective," *Review of Economic Studies*, 1999, *66* (1), 169–182.

Hopenhayn, Hugo and Maryam Saeedi, "Optimal Ratings and Market Outcomes," *NBER Working Paper Series*, 2019, pp. 1–39.

Hörner, Johannes and Nicolas Lambert, "Motivational Ratings," *Review of Economic Studies*, 2018, *forthcomin.*

Horstmann, Ignatius J. and Glen M. MacDonald, "When is Advertisnig a Signal of Product Quality?," *Journal of Economics and Management Strategy*, 1994, *3* (3), 561–584.

_ and Sridhar Moorthy, "Advertising Spending and Quality for Services: The Role of Capacity," *Quantitative Marketing and Economics*, 2003, *1* (3), 337–365.

Hui, Xiang, Maryam Saeedi, Giancarlo Spagnolo, and Steve Tadelis, "Certification, Reputation and Entry: An Empirical Analysis," 2018, pp. 1–59.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Kihlstrom, Richard E. and Michael H. Riordan, "Advertising as a Signal," *Journal of Political Economy*, 1984, *92* (3), 427–450.

Lizzeri, Alessandro, "Information Revelation and Certification Intermediaries," *RAND Journal of Economics*, 1999, *30* (2), 214–231.

Luca, Michael and Georgios Zervas, "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Management Science*, 2016, *62* (12), 3412–3427.

Mayzlin, Dina, "Promotional Chat on the Internet," *Marketing Science*, 2006, *25* (2), 155–163.

Milgrom, Paul and John Roberts, "Price and Advertising Signals of Product Quality," *Journal of Political Economy*, 1986, *94* (4), 796–821.

_ and _ , "Relying on the Information of Interested Parties," *RAND Journal of Economics*, 1986, *17* (1), 18–32.

Ostrovsky, Michael and Michael Schwarz, "Information disclosure and unraveling in matching markets," *American Economic Journal: Microeconomics*, 2009, *2* (2), 34–63.

Rayo, Luis and Ilya Segal, "Optimal Information Disclosure," *Journal of Political Economy*, 2010, *118* (5), 949–987.

Reimers, Imke C. and Joel Waldfogel, "Digitization and Pre-Purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," *National Bureau of Economic Research*, 2020.

Saeedi, Maryam, "Reputation and Adverse Selection: Theory and Evidence from eBay," *RAND Journal of Economics*, 2019, *50* (4), 822–853.

_ **and Ali Shourideh**, "Optimal Rating Design," 2020, pp. 1–47.

Sahni, Navdeep S and Harikesh S Nair, "Does Advertising Serve as a Signal? Evidence from a Field Experiment in Mobile Search," *The Review of Economic Studies*, 2019, (October 2019), 1529–1564.

Vellodi, Nikhil, "Ratings Design and Barriers to Entry," *SSRN Electronic Journal*, 2020, pp. 1–63.

A Proofs

Proof of Theorem 1. By $M_t = \mu + \lambda [Y_t - \nu] \Leftrightarrow \lambda Y_t = M_t - \mu + \lambda\nu$, and the linear strategy $F_t = \alpha\theta_t + \beta M_t + \delta\mu$, the increment of M_t is written as

$$\begin{aligned} dM_t &= d(\lambda Y_t) \\ &= (-\phi + a\lambda\beta) M_t dt \\ &+ (a\lambda\alpha + bq\lambda)\theta_t dt \\ &+ (\phi\mu - \phi\lambda\nu + a\lambda\delta\mu) dt \\ &+ bq\lambda\sigma_\xi dZ_t^\xi \end{aligned}$$

Now, we look for a quadratic value function

$$V = v_0 + v_1\theta + v_2 M + v_3\theta^2 + v_4 M^2 + v_5\theta M \tag{9}$$

satisfying the HJB equation:

$$\begin{aligned}
rV(\theta, M) = \sup_{F \in \mathbb{R}} & (1-\tau) M \cdot q - \tau M \cdot F - \frac{c}{2} F^2 \\
& - \kappa (\theta - \mu) V_\theta \\
& + \left\{ a\lambda F + bq\lambda\theta - \phi \left[M - \bar{\theta} + \lambda \bar{Y} \right] \right\} V_M \\
& + \frac{\sigma_\theta^2}{2} V_{\theta\theta} \\
& + \frac{bq\lambda^2 \sigma_\xi^2}{2} V_{MM}
\end{aligned}$$

By the first-order condition,

$$\begin{aligned}
0 &= -\tau M - cF + a\lambda V_M \\
\Leftrightarrow F &= -\frac{\tau}{c} M + \frac{a\lambda}{c} V_M \\
&= \frac{a\lambda}{c} v_5 \theta + \left(2\frac{a\lambda}{c} v_4 - \frac{\tau}{c} \right) M + \frac{a\lambda}{c} v_2
\end{aligned}$$

By matching coefficients with $F = \alpha\theta + \beta M + \delta\mu$,

$$\begin{aligned}
\alpha &= \frac{a\lambda}{c} v_5 \\
\beta &= 2\frac{a\lambda}{c} v_4 - \frac{\tau}{c} \\
\delta\mu &= \frac{a\lambda}{c} v_2
\end{aligned}$$

By solving them for v_k's,

$$\frac{c}{a\lambda}\alpha = v_5 \tag{10}$$

$$\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) = v_4 \tag{11}$$

$$\frac{\delta\mu c}{a\lambda} = v_2 \tag{12}$$

By the Envelop condition w.r.t. M,[15]

$$\begin{aligned}
rV_M = & (1-\tau) q - \tau F \\
& - \kappa (\theta - \mu) V_{\theta M} \\
& - \phi V_M \\
& + \{ a\lambda F + bq\lambda\theta - \phi [M - \mu + \lambda\nu] \} V_{MM}
\end{aligned}$$

[15] The envelop condition w.r.t. θ gives conditions characterizing v_1 and v_3, and one characterizing v_5, which coincides with the condition from the envelop condition w.r.t. M.

By inserting the derivatives of eq.(9) and equating the coefficients of θ, M, and constants on LHS and RHS,

$$
\begin{aligned}
(r+\phi)\, v_5 &= -\tau\alpha - \kappa v_5 + \{a\lambda\alpha + bq\lambda\}\, 2v_4 \\
2\,(r+\phi)\, v_4 &= -\tau\beta + \{a\lambda\beta - \phi\}\, 2v_4 \\
(r+\phi)\, v_2 &= (1-\tau)\, q - \tau\delta\bar{\theta} + \kappa\mu v_5 + \{a\lambda\delta\mu + \phi\mu - \phi\lambda\nu\}\, 2v_4
\end{aligned}
$$

Then, inserting eq.(10) to eq (12),

$$
(r+\phi+\kappa)\,\frac{c}{a\lambda}\alpha = -\tau\alpha + \{a\lambda\alpha + bq\lambda\}\, 2\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) \tag{13}
$$

$$
2\,(r+\phi)\,\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) = -\tau\beta + \{a\lambda\beta - \phi\}\, 2\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) \tag{14}
$$

$$
(r+\phi)\,\frac{\delta\mu c}{a\lambda} = (1-\tau)\, q - \tau\delta\mu + \kappa\mu\frac{c}{a\lambda}\alpha + \{a\lambda\delta\mu + \phi\mu - \phi\lambda\nu\}\, 2\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) \tag{15}
$$

By combining with the consistency of λ: $\lambda = \frac{(a\alpha+bq)\sigma_\theta^2(\phi-a\beta\lambda)}{(\phi-a\beta\lambda+\kappa)\kappa bq\sigma_\xi^2+\sigma_\theta^2(a\alpha+bq)^2}$, we can characterize α, β, δ, λ. In the following, I do so by using an aggregater $L = -a\beta\lambda$ so that the stationarity condition is easier to verify. First, by replacing λ to $-\frac{L}{a\beta}$ in the above four equations,

$$
0 = -\frac{bq(\beta c+\tau)}{a} + \alpha\tau - \alpha(\beta c+\tau) - \frac{\alpha\beta c\kappa}{L} - \frac{\alpha\beta c\phi}{L} - \frac{\alpha\beta c r}{L} \tag{16}
$$

$$
0 = \beta\tau - \beta(\beta c+\tau) - \frac{2\beta\phi(\beta c+\tau)}{L} - \frac{\beta r(\beta c+\tau)}{L} \tag{17}
$$

$$
0 = \frac{\nu\phi(\beta c+\tau)}{a} - \delta\mu(\beta c+\tau) + \frac{\alpha\beta c\kappa\mu}{L} - \frac{\beta c\delta\mu\phi}{L} + \frac{\beta\mu\phi(\beta c+\tau)}{L} - \frac{\beta c\delta\mu r}{L} + \delta\mu\tau + q\tau - q \tag{18}
$$

$$
-\frac{L}{a\beta} = \frac{\sigma_\theta^2(L+\phi)(a\alpha+bq)}{\sigma_\theta^2(a\alpha+bq)^2 + \kappa bq\sigma_\xi^2(\kappa+L+\phi)} \tag{19}
$$

By solving (17) for β, we get $\beta = -\frac{\tau}{c}\left(\frac{r+2\phi}{r+2\phi+L}\right) \equiv B\,(L)$. By inserting this into (16) and solving it for α, we get $\alpha = \frac{bq}{a}\frac{L^2}{(r+2\phi)(r+\phi+\kappa+L)} \equiv A\,(L)$. By plugging $\beta = B\,(L)$ and $\alpha = A\,(L)$ into (19), we obtain an equation characterizing L:

$$
-\frac{L}{aB\,(L)} = \frac{\sigma_\theta^2(L+\phi)(aA\,(L)+bq)}{\sigma_\theta^2(aA\,(L)+bq)^2 + \kappa bq\sigma_\xi^2(\kappa+L+\phi)}
$$

Rearranging it , we get

$$
\begin{aligned}
1 &= \frac{\sigma_\theta^2(L+\phi)(aA\,(L)+bq)}{\sigma_\theta^2(aA\,(L)+bq)^2 + \kappa bq\sigma_\xi^2(\kappa+L+\phi)}\frac{-aB\,(L)}{L} \\
&\equiv h\,(L)
\end{aligned}
$$

To evaluate $h(L)$, the sign of L is useful to characterize.

Lemma 6. *$\beta < 0$ and $L > 0$ under the linear stationary Gaussian equilibrium.*

Proof. By the stationarity, we must have $\phi + L > 0$. Then,

$$\begin{aligned}\beta &= -\frac{\tau}{c}\left(\frac{r+2\phi}{r+2\phi+L}\right)\\ &= -\frac{\tau}{c}\left(\frac{r+2\phi}{r+\phi+\phi+L}\right)\\ &< 0\end{aligned}$$

Then, $\alpha = \frac{bq}{a}\frac{L^2}{(r+2\phi)(r+\phi+\kappa+L)} > 0$ and $\lambda = \frac{(a\alpha+bq)\sigma_\theta^2(\phi+L)}{(\phi+L+\kappa)b^2q^2\kappa\sigma_\xi^2+\sigma_\theta^2(a\alpha+bq)^2} > 0$. Now, we can conclude $-a\beta\lambda \equiv L > 0$. $\square$

Now, it is shown that $\lim_{L\to 0} h(L) = \infty$ and $\lim_{L\to\infty} h(L) = 0$. Then, combined with the continuity of $h(L)$, there exist some L such that $h(L) = 1$. The uniqueness is proved by checking whether $h'(L) < 0$ holds. It is shown that

$$h'(L) = -\,h_1(L)\left\{h_2(L) + L^4\left(-\kappa^2 + 6\kappa\phi + 4r^2 + 2\kappa r + 17r\phi + 19\phi^2\right)\right\}$$

where $h_1(L), h_2(L) > 0$ for all $L > 0$. Thus, $6\kappa\phi + 4r^2 + 2\kappa r + 17r\phi + 19\phi^2 > \kappa^2$ is sufficient for $h'(L) < 0$ $\square$

Proof of Lemma 2. By plugging $\alpha(L)$ and $\beta(L)$ in to h, it can be written as $h(L) = \frac{a\tau}{c}\frac{h_3}{L(L+r+2\phi)\left(h_4+\left(\sigma_\xi/\sigma_\theta\right)^2 h_5\right)}$ $\square$

where $h_3 = (L+\phi)(r+2\phi)^2(\kappa+L+r+\phi)\left(L^2+L(r+2\phi)+(r+2\phi)(\kappa+r+\phi)\right)$, $h_4 = bq\left(L^2+L(r+2\phi)+(r+2\phi)(\kappa+r+\phi)\right)^2$, $h_5 = \kappa(r+2\phi)^2(\kappa+L+\phi)(\kappa+L+r+\phi)^2$. Note that h_3, h_4, h_5 are positive and independent of a and σ_ξ/σ_θ. Thus, h is increasing in $\frac{a\tau}{c}$ and decreasing in σ_ξ/σ_θ. Since $h'(L) < 0$ is shown in the proof of Theorem 1, the implicit function theorem tells that L is increasing in a and decreasing in σ_ξ/σ_θ. Furthermore, $h(L) \to \infty$ if L is bounded above and $\frac{a\tau}{c} \to \infty$. Thus, to satisfy the equilibrium condition: $1 = h(L)$, L goes infinite as $\frac{a\tau}{c}$ goes infinite. Similarly, $h(L) \to 0$ if L is bounded away from zero and $\frac{a\tau}{c} \to 0$. Thus, L goes infinite as $\frac{a\tau}{c}$ goes infinite to satisfy the equilibrium condition.

Proof of Proposition 1 and 2. Since $E[M_t] = E[E[\theta_t|Y_t]] = \mu$, we have $E[F_t] = E[\alpha\theta_t + \beta M_t + \delta\mu] = (\alpha+\beta+\delta)\mu$. By expressing α, β, δ as a function of the equilibrium aggregater L, it is written as $E[F_t] = \frac{cLq(1-\tau)(L+r+2\phi)-\mu\tau^2\left(r^2+3r\phi+2\phi^2\right)}{c\tau(L^2+L(r+2\phi)+r^2+3r\phi+2\phi^2)}$ and the partial derivative with respect to L is $\frac{\partial E[F_t]}{\partial L} = \frac{\left(r^2+3r\phi+2\phi^2\right)(2L+r+2\phi)\left(cq(1-\tau)+\mu\tau^2\right)}{c\tau(L^2+L(r+2\phi)+r^2+3r\phi+2\phi^2)^2} > 0$.

Since a, σ_ξ, and σ_θ affects $E\left[F_t\right]$ only through the aggregater L, we can show the effects of a and $\frac{\sigma_\xi}{\sigma_\theta}$ by analyzing the sign of $\frac{dL}{da}$ and $\frac{dL}{d(\sigma_\xi/\sigma_\theta)}$. By Lemma 2, we can conclude $E\left[F_t\right]$ increasing in a and decreasing in $\frac{\sigma_\xi}{\sigma_\theta}$.

Since $E\left[F_t\right] > 0$ for sufficiently large L and $L \to \infty$ as $a \to \infty$, $E\left[F_t\right] > 0$ holds for sufficiently large a. □

Proof of Proposition 3. The equilibrium condition gives $\alpha = \frac{bq}{a}\frac{L^2}{(r+2\phi)(r+\phi+\kappa+L)}$ and $\beta = -\frac{\tau}{c}\left(\frac{r+2\phi}{r+2\phi+L}\right)$. Furthermore, it is shown that $\frac{\partial\alpha}{\partial L} > 0$ and $\frac{\partial\beta}{\partial L} > 0$. Then, Lemma 2 concludes the proposition. □

Proof of Lemma 3 and 4. An arbitrary strategy α, β, δ satisfying $\phi - a\beta\lambda$ (not necessarily the equilibrium strategy) generates a stationary distribution. Using the vairance-covariance matrix of the stationary distribution, the informativeness is written as

$$\rho^2 = \frac{(\phi - a\beta\lambda)(a\alpha + bq)^2}{(\kappa + \phi - a\beta\lambda)\Big((a\alpha + bq)^2 + \kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa + \phi - a\beta\lambda)\Big)}$$

Thus, the informativeness without fake reviews is

$$\rho^2 = \frac{\phi(bq)^2}{(\kappa + \phi)\Big((bq)^2 + \kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa + \phi)\Big)}$$

.On the other hand, at the equilibrium, $-a\beta\lambda$ can be replaced to L, and $a\alpha$ is written as a function in L: $a\alpha = bq\frac{L^2}{(r+2\phi)(r+\phi+\kappa+L)}$ such that $a\alpha = 0$ when $L = 0$. Note that a does not appear in the RHS, so the direct and indirect effects of a on $a \cdot \alpha$ are all captured by L. Now the equilibrium informativeness is written as:

$$\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) = \frac{(\phi + L)(a\alpha + bq)^2}{(\kappa + \phi + L)\Big((a\alpha + bq)^2 + \kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa + \phi + L)\Big)}.$$

Note that $\rho^2\left(0;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) = \frac{\phi(bq)^2}{(\kappa+\phi)\left((bq)^2+\kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa+\phi)\right)}$ coinsides with the informativeness without fake reviews. This concludes Lemma 4. □

Proof of Proposition 5. The first part is proved by the limit as $L \to \infty$:

$$\begin{aligned}
&\lim_{L\to\infty} \rho^2 \left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) \\
=&\lim_{L\to\infty} \frac{(\phi+L)}{(\kappa+\phi+L)} \frac{(a\alpha+bq)^2}{\left((a\alpha+bq)^2+\kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa+\phi+L)\right)} \\
=&1
\end{aligned}$$

The second part comes from the derivative of ρ^2 with respect to L around zero. □

Proof of Proposition 6. The optimal ϕ without fake reviews is chatacterized by $\frac{\partial}{\partial\phi}\rho^2\left(0;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) = 0$, which yields $\phi^0 = \sqrt{bq\left(\sigma_\theta/\sigma_\xi\right)^2+\kappa^2}$ as the optimal level. On the other hand, the effect of ϕ at the equilibrium is

$$\begin{aligned}
\frac{d\rho^2}{d\phi} &= \frac{\partial}{\partial\phi}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) + \frac{\partial}{\partial L}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)\frac{dL}{d\phi} \\
&= \frac{\partial}{\partial\phi}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) - \frac{\partial}{\partial L}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)\frac{\partial h}{\partial\phi}/\frac{\partial h}{\partial L}
\end{aligned}$$

By evaluating this at $\phi = \phi^0$, we obtain $\frac{d\rho^2}{d\phi}|_{\phi=\phi^0} < 0$.

The second part is proved by two inequalities: $\rho^2\left(0;\, \phi^0,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) < \rho^2\left(L\left(\phi^0\right);\, \phi^0,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) \le \rho^2\left(L\left(\phi^*\right);\, \phi^*,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)$. The first inequality is proved as follows. For any $L > 0$,

$$\begin{aligned}
&\rho^2\left(L;\, \phi^0,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) - \rho^2\left(0;\, \phi^0,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) \\
=&r \cdot g_1 + g_2
\end{aligned}$$

where g_1 is polynomial in r and L and $g_2 > 0$ is polynomial in L and does not depend on r. Since $L \to C$ for some $C > 0$ as $r \to 0$, $r \cdot g_1 + g_2$ converges to a positive number. Thus, for sufficiently small r, the first inequality holds. The second inequality holds by definition. □

Proof of Proposition 7. Similarly to Proposition 6, the total effect of σ_ξ/σ_θ is written as $\frac{d\rho^2}{d(\sigma_\xi/\sigma_\theta)} = \frac{\partial}{\partial(\sigma_\xi/\sigma_\theta)}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right) - \frac{\partial}{\partial L}\rho^2\left(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta\right)\frac{\partial h}{\partial(\sigma_\xi/\sigma_\theta)}/\frac{\partial h}{\partial L}$. It is shown that $\frac{d\rho^2}{d(\sigma_\xi/\sigma_\theta)} < 0$. □

Proof of Theorem 2. [sketch] The characterizing equation $h(L) = 1$ is modified as $1 = \frac{\eta f(L)+(1-\eta)C}{g(L)} \equiv h\left(L;\eta\right)$ where $\lim_{L\to 0} h\left(L;\eta\right) = \infty$ and $\lim_{L\to\infty} h\left(L;\eta\right) = 0$. Then, $h_L\left(L;\eta\right) < 0$ holds for any $\eta \in [0,\, 1]$ as long as $h_L\left(L;1\right) < 0$,

[Proof of Proposition 8] [sketch] it is shown that $h\left(L;\eta\right) \le h\left(L;1\right)$ for any $\eta \in [0,\, 1]$. Thus, the equilibrium L will be smaller with $\eta < 1$ compared to $\eta = 1$.

[Proof of Proposition 9][sketch] $\frac{\partial bias}{\partial L} > 0$ if $bias > 0$. $\square$

B An Alternative Model with Changing q

C Simulation Results

C.1 Comparison to Disaggregate Information

C.2 Mixture of the Rational and Naive Consumers

C.3 Mean Squared Error as a Criteria