Controlling Fake Reviews*

Yuta Yasui†

January 31, 2023

Abstract

This paper theoretically analyzes fake reviews on a platform market using models where a seller creates fake reviews through incentivized transactions, and its sales depend on its rating based on a review history. The platform can control the incentive for fake reviews by changing the parameters of the rating system, such as weights placed on old and new reviews and its filtering policy. At equilibrium, the number of fake reviews increases as quality increases but decreases as reputation improves. Since fake reviews have a positive relationship with a product's underlying quality, rational consumers find a rating more informative when fake reviews exist, while credulous consumers suffer from a bias caused by boosted reputation. A stringent filtering policy can decrease the expected amount of fake reviews and the bias of credulous consumers, but at the same time, it can decrease the informativeness of a rating system for rational consumers. In terms of the weight placed on the review history, rational consumers benefit from higher weights on past reviews than from optimal weights without fake reviews.

*"Fully ref**ded after RE**W*

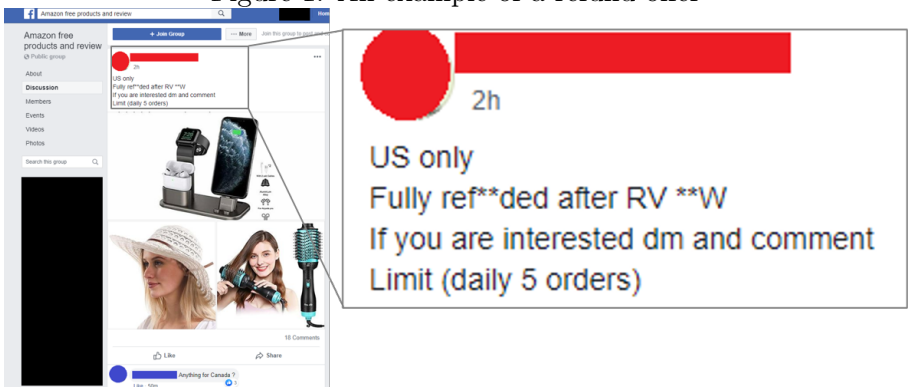
If you are interested dm and comment"

— a post on Facebook

*For continual guidance, I thank my advisor Ichiro Obara. For helpful duscussions, I thank Stepan Aleksenko, Simon Board, Brett Hollenbeck, Akina Ikudo, Akira Ishide, Jacob Kohlhepp, Jay Lu, Toshihiro Matsumura, Moritz Meyer-ter-Vehn, Tomasz Sadzik, Susumu Sato, Nikhil Vellodi and seminar participants at Kochi University of Technology, UCLA, University of Tokyo, Hitotsubashi University, 12th Paris Conference on Digital Economy, International Conference on Oligopoly Theory at Cohnnam National University.

†Kochi University of Technology, yasui.yuta@kochi-tech.ac.jp

Figure 1: An example of a refund offer



Person Red, who is suspected as a seller on Amazon, posts pictures of its products and offers full refunds of the products after reviews of them. About an hour after of the post, Person Blue, who is suspected as a fake reviewer, shows an interest on the products and refunds.

1 Introduction

Online platform markets are growing worldwide, such that both businesses and their customers increasingly rely on reviews on the platforms.[1] At the same time, incentives for sellers to make fake reviews are also growing. Washington Post (Dwoskin and Timberg, 2018) reports that based on fake review detection algorithms, 50.7% of reviews for Bluetooth headphones, 58.2% for Bluetooth speakers, 55.6% for weight loss pills, and 67.0% for testosterone boosters on Amazon are suspicious. How do sellers make fake reviews? The sellers can post information of their products with refund offers, which are typically finalized via PayPal after purchases and positive reviews on Amazon. (See Fig. 1 for an example of such an offer.)[2] These reviews correspond to verified purchases and are reflected to the star rating (until they are detected by Amazon).[3] He et al. (2020) connect such refund offers on Facebook with product listings on Amazon and show a positive correlation between refund offers on Facebook and a product's performance on Amazon such as its ratings, sales ranking, and the number of reviews. Regulators have been concerned about fake reviews, and their attitude toward fake reviews is becoming stringent. For instance, in 2019, the Federal Trade

[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 institutional details and data analysis on platforms and ratings, see also Belleflamme and Peitz (2018)

[2] For more details on evasive practice by incentivized reviewers and agents who contact buyers to incentivize them to write reviews, see Oak (2021).

[3] Offers of such fake reviews from fake reviewers have been found on eBay.

Commission (FTC) filed the first case against paid fake reviews by CureEncapsulations on Amazon. Online platforms have restricted fake reviews in their own ways, but regulators put increasingly high pressure on online platforms to maintain a stricter attitude against fake reviews. [4]

However, the impact of fake reviews on consumers on a platform is not clear. First, consumers might not be fooled by fake reviews if they know that there are fake reviews. In the standard work of Holmström (1999), the market can correctly anticipate the behavior of long-lived players and debias the signal. Furthermore, customers might be able to elicit additional information from fake reviews. If only high-quality sellers make fake reviews to boost their initial reputation, the boosted rating can be an even better signal of good quality. Such a behavior might be possible if low quality is revealed via word of mouth, and only a high-quality seller can reap benefits from future sales, as suggested by Nelson (1970,1974) in the context of advertising. [5]

In this study, we examine a theoretical model in which sales are determined by the seller's reputation level and the seller chooses the amount of positive fake reviews at each instance. Consumers perceive a seller's reputation based on the potentially boosted ratings displayed on the platform. The platform can control how strictly it filters fake reviews and how much the rating reflects the information of past feedback (i.e., how fast the rating evolves). A key assumption in this study is that it becomes harder for a seller to make fake reviews as its reputation improves because of the higher reimbursement necessary to incentivize reviewers due to the higher price.[6] This brings more fake reviews from the seller with low reputation. This also generates the dependence of fake reviews on the seller's quality-type. Because high-quality sellers benefit more from their high reputation, high-quality sellers generate more fake reviews at equilibrium. Because of this positive relationship between the number of fake reviews and quality, consumers sometimes benefit from lenient policies on fake reviews. In the literature on signaling promotion, the complementarity between quality and reputation is understudied because, in most research, promotion is done only once at the beginning

[4]For instance, in 2019, the Competition and Markets Authority (CMA) in U.K. launched work programme "has written to Facebook and eBay this week urging them to conduct an urgent review of their sites to prevent fake and misleading online reviews from being bought and sold". In responses, both Facebook and eBay have immediately deleted posts identified by CMA, and updated their policy to explicitly prohibit offers of fake reviews. In 2020, May, CMA has launched new investigation into online websites on how they currently detect fake or misleading reviews.

[5]Ananthakrishnan et al. (2020) analyze the display of fake reviews from a different perspective and 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.

of a game. In this study, the complementarity comes from the future cost-saving effect rather than an increase in revenue.

The opposite dependence of fake reviews on a reputation about quality and on the underlying true quality also provides some cautions on empirical analysis on signaling promotion. That is, reputation-based indices, such as customer rating, can be a bad proxy for a product's underlying quality. Researchers can estimate opposite results if they use customer rating as a proxy for quality. Furthermore, even if the true quality is measured, it is important to control for the reputation level when estimating the relationship between promotion and the underlying quality. Fig. 2 exemplifies the possibility of an omitted variable issue; that is, the promotion level and the true quality of a product can be negatively correlated without being conditioned upon a firm's reputation level, even though quality and promotion have a positive relationship, *ceteris paribus*.

The negative relationship between fake reviews and a firm's reputation also increases the speed at which the rating changes. That is, in the presence of fake reviews, when the rating goes down (up), it more quickly goes up (down) than when the rating system has no fake reviews. This distorts the informativeness of the rating system. How fast the rating changes relates to the relative weight of new information in the rating system. The greater is the weight of new information (and the lower the weight of old information), the faster is the transition of the rating. Thus, the equilibrium effect that makes the transition faster has the effect of distorting upward the weight of the new information (and downward the weight of the old information). Therefore, given the existence of fake reviews, the platform needs to make some adjustments. The platform should set a lower weight for new information (and higher weight for old information) compared with a rating system that has no fake reviews.

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

[7]For instance, Federal Trade Commission (FTC)'s mission is "[p]rotecting consumers and competition by preventing anticompetitive, deceptive, and unfair business practices through ...". (https://www.ftc.gov/about-ftc)

rameters.

Thus, the regulator might face a trade-off between the precision of the information for rational consumers and the bias that credulous consumers suffer from. This study provides a framework for analyzing such a trade-off.

The remainder of this paper is organized as follows. In Section 2, we review related literature. In Section 3, we analyze a model with rational buyers. In Section 4, we introduce credulous consumers. Section 5 concludes. Most of the proofs are deferred to the Appendix.

2 Literature Review

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

The first strand is often framed in the context of certification, such as the works of 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 wherein the quality is endogenously chosen by the seller rather than the exogenous variable.

This paper relates to another strand of literature, as it analyzes how to aggregate the players' 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's (1999) signal jamming/career concern model, Hörner and Lambert (2021) show that the effort level of a long-lived player is maximized by a rating that is linear to past observations. Vellodi (2020) analyzes the impact of rating on the entry/exit behavior of a firm and derives an optimal rating that prevents high-quality sellers from exiting from the market due to a reputation trap of failing to accumulate good reputation because of initial bad luck. Bonatti and Cisternas (2020) examine

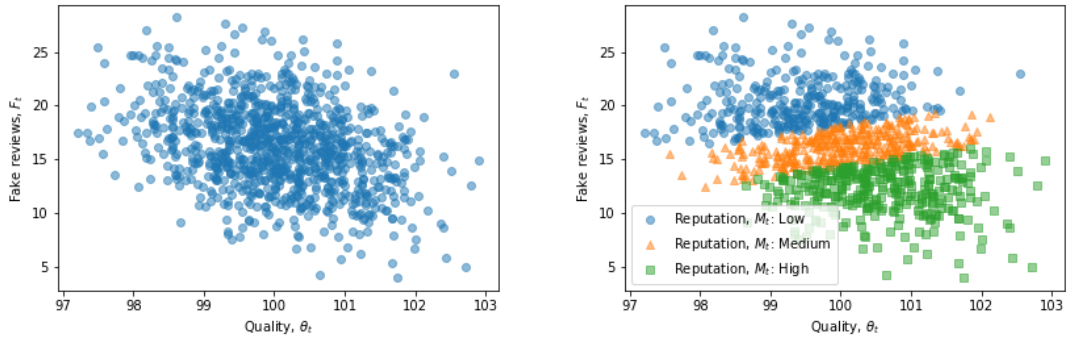
a long-lived consumer's Ratchet effect. The consumers try to hide its willingness to pay to avoid the personalized pricing by short-lived monopolist, so that the consumption does not perfectly reflect their willingness to pay. Similarly to Hörner and Lambert (2021) and Bonatti and Cisternas (2020), this study examines the relationship between a signal-jamming structure and a linear rating system. In contrast to Hörner and Lambert (2021), the equilibrium strategy is dependent on the hidden quality and reputation, such that the seller's strategy changes the informativeness of the rating on the equilibrium path, as in Bonatti and Cisternas (2020).[8] In contrast to Bonatti and Cisternas (2020), where the effect of the manipulation is endogenously determined via the short-lived player's belief, in this study, the platform controls for the effectiveness of the manipulation so that we can analyze the impact of censorship by the platform. In addition, this study departs from the literature by analyzing the impact of manipulation on naive/credulous consumers, which is often the concern of regulators.

This paper also contributes to the literature on promotion and signaling. Nelson (1970, 1974) argues that even if the promotion does not have any intrinsic information, "burning money" itself can be a signal of good quality because such a signal pays off only for high-quality firms through repeated purchases 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. Using a one-shot signal-jamming framework instead of a signaling model, Mayzlin (2006) shows a negative relationship between promotion through fake reviews and quality, and Dellarocas (2006) generalizes conditions for the positive/negative correlation in a one-shot signal-jamming model. Bar-isaac and Deb (2014) examine the effects of vertically/horizontally heterogeneous preferences, and Grunewald and Kräkel (2017) examine the effect of competition between firms. Most studies on the signaling role of promotion are based on models with one-shot promotion, except for Horstmann and MacDonald (1994), where the experience of the product is an imperfect signal of the quality, and the signaling via advertising is done only after establishing a reputation so that it is hard for low-quality sellers to mimic high-quality sellers' behavior.[9]. In this study, I examine a dynamic signal-jamming model,

[8] Another contrast to Hörner and Lambert (2021) is that they start from a general information structure so that they can represent any reputation by changing the information structure. Then, they can focus on the resulted process of reputation level in a similar way that researchers focus on the resulted outcome by the revelation principle in the context of the mechanism design. On the other hand, this paper and Bonatti and Cisternas (2020) use more specific information structure, so that we should examine how the consumers interpret the resulted rating.

[9] Aside from the context of the rating system or the signaling promotion, Drugov and Troya-Martinez (2019) examine the biasing behavior of the seller in a model a. la. Holmström (1999) incorporating a detection rule and

Figure 2: 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.

where reputational concern is the driving force for the positive correlation between quality and promotion. It also generates non-degenerate dynamics consistent with an observation by Luca and Zervas (2016) that strategic manipulation increases after a drop in reputation.

The dependence of fake reviews on reputation also provides some implications for the empirical literature on signaling promotion. The literature has had weak support regarding the correlation between quality and promotion. For instance, Kwoka (1984) observes that optometrists with more advertisements provide less thorough eye examination, and Horstmann and Moorthy (2003) observe that advertising is hump-shaped in terms of quality among restaurants in New York. Recently, Sahni and Nair (2019) implement a quasi-experiment to isolate the intrinsic information and signaling effect of burning money and show that the consumer positively responds to the burning of money. They point out that it is difficult to show the relationship between quality and promotion level because it is difficult to obtain a reliable measure of quality. This paper emphasizes this point. A reputation-based index, such as customer rating, can be a bad proxy for the underlying quality. The reputation level and the underlying quality level have opposite impacts on the promotion level in equilibrium. Furthermore, even if the true quality is measured, it is important to control for the reputation level. As shown in Fig. 2 , the level of promotion and the true quality can be negatively correlated without being conditioned upon the reputation level, even though quality and promotion have a positive relationship, *ceteris paribus*.

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 study, we examine both models with rational consumers and naive consumers. In this section, we first introduce a baseline model with a mass of rational consumers. The consumers rationally expect that a long-lived seller makes fake reviews following a linear strategy. However, they cannot induce the seller's exact action at time t because the quality is still hidden, even though the strategy and the current reputation are known to the consumers.

Then, in the next section, we introduce a market with naive consumers who do not expect any fake reviews while the seller makes fake reviews, such that the reputation is biased upward. In each model, we examine the impact of the platform's filtering/censoring policy on reviews, the weights of new and old reviews, and the precision of genuine reviews.

3.1 Model

The model is in a 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 units 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 the rating of the product at time t.[10] The price being a representation of the reputation of the hidden quality is the standard assumption in the literature on reputation. The quality θ_t governs consumers' willingness to pay for the product, so the price is high when the expected quality of the product is high. A more specific underlying model, that can incorporate naive consumers is suggested in the Appendix.

The quality, θ_t, and 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:

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

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

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 $\left(Z_t^\theta,\, Z_t^\xi\right)$ is a standard Brownian motion; 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, and σ_θ and σ_ξ govern the standard deviations of the disturbance. The exogenous mean-reverting process of θ_t is understood as resulting from the competition over quality among sellers. The relative quality of a firm's product might decrease due to the rise of other sellers with even higher quality. The firm's product's relative quality might increase when a competitor increases its product's price. The transition of the rating, Y_t, is interpreted in a 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. After filtering suspicious reviews, the new reviews consist of two components: "organic" reviews and the remaining fake reviews. The second and third terms of Eq. (3) correspond to organic reviews. Higher quality tends to generate high reviews, and the information becomes precise when there is feedback from many transactions (i.e., high q) or a high response rate (i.e., high b). The disturbance, $\sigma_\xi dZ_t^\xi$, is caused by the heterogeneity of the criteria among customers.[11] 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 reviews enter as $aF_t dt$. Thus, a small a implies stringent censorship. As in **?**, Vellodi (2020), and Bonatti and Cisternas (2020), the rating, Y_t, does not exactly capture 5-star rating on Amazon, Yelp, or some other online platform. The level of Y_t is dependent on the mean of θ_t and other parameters. By this specification of the rating, we can rely on the normality to simplify the analysis.

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 denotes the transaction fees imposed by the platform. The first term is the total revenue from all transactions, including those corresponding to fake reviews, and the second term is the reimbursement cost to the fake reviewers. The last term expresses that generating more fake reviews

[11] In this paper, the mechanism behind the customer feedback is abstracted and 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.

is harder. The seller might find it challenging to search for incentivized reviewers through communities such as Facebook. Some fake review services may charge a higher price for fake reviews. Furthermore, increasing the number of fake reviews come with a higher risk of being detected by the platform. The cost of production is abstracted out from the model. [12] 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 when it is rewritten as follows:

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

Without the second term in eq. (4), the model becomes effectively a special case of Hörner and Lambert (2021), which is based on Holmström's (1999) signal-jamming model and uses a general information structure as a rating. However, due to the existence of this term, the marginal cost of the manipulation depends on the current reputation level. Therefore, the equilibrium manipulation level depends on the current rating in contrast to Hörner and Lambert (2021), where the equilibrium action turns out to be state-independent. Instead of relying on the time- and state-invariant action, we apply the idea of Bonatti and Cisternas (2020) to focus on a linear strategy, and a Gaussian stationary distribution of (θ_t, Y_t). Then, the Hamilton-Jacobi-Bellman equation gives a simple quadratic value function, which is solved by the guess-and-verify method. It is verified that as τ approaches zero, the equilibrium strategy becomes invariant to θ_t, Y_t, (and t).

The interaction between the current reputation and the current action is considered as the driving force of the non-degenerate Markov equilibrium strategy. In this study, this interaction between reputation and manipulation is derived from the reimbursement to fake reviewers; however, such an interaction can be more commonly observed in the context of fake reviews. For instance, if the reputation is high, then a large demand can crowd-out fake reviews, such that the effective fake reviews can be more costly given the high reputation. In the Appendix, an alternative model with such an interpretation is discussed. A model with a changing quantity that is isomorphic to the main model is discussed in Appendix C.

[12] Whether the high quality seller or low quality seller face high costs of production is arguable by itself. If high quality come from the seller's high productivity, the high quality seller can produce with lower costs. If the low quality is by the seller's choice rather than the difference in the production technology among sellers, the low quality product would be associated with low production cost. The different specifications on the production costs can cause different pattern in fake reviews, but those extensions are deferred to the future research.

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 the 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. $F = \arg\max_{(\tilde{F}_t)_{t\geq 0}} E_0\left[\int_0^\infty e^{-tr}\pi_t\right]$ 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 a stationary Gaussian.

We do not 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, given a linear strategy, the condition for $(\theta_t,\, Y_t)_{t\geq 0}$ to be a stationary Gaussian is simply characterized by an inequality—similar to Bonatti and Cisternas (2020)—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 $(\theta_t,\, Y_t)_{t\geq 0}$ to have a stationary distribution (otherwise, the process of Y_t diverges). When $(\theta_t,\, Y_t)$ is a 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 is stationary, all expectations in Eq.(6) are constants. By letting $\lambda \equiv \frac{Cov(\theta_t, Y_t)}{Var(Y_t)}$ and $\nu \equiv E[Y_t]$ (and $\mu = E[\theta_t]$ 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. *(Stationarity 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 a 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. $(\theta_0, Y_0)' \sim \mathcal{N}\left([\mu, \nu]', \Gamma\right)$ *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 a stationary distribution. Now, the 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 function and the linear strategy via the matching coefficient.

3.2 Equilibrium Characterization

The equilibrium strategy is characterized by guessing the quadratic value function and the linear strategy and by matching coefficients α, β, δ, $(v_k)_{k=0}^5$ of the first-order conditions, envelop conditions, and the stationarity condition characterized in Lemma 1. In the proof, the characterizing conditions are summarized into one equation $h(L) = 0$ with an aggregator $L \equiv a\lambda\beta$, and then all the equilibrium coefficients are derived as a function of L. Aggregator L is interpreted as an equilibrium effect on the speed of the rating transition or the equilibrium effect on the relative weight of new information. When L is positive, the rating transition effectively speeds up because the low rating is soon boosted back to the average rating by fake reviews.

By analyzing the existence and uniqueness of the aggregator L and examining the corresponding equilibrium coefficients, we obtain the following theorem:

Theorem 1 (Existence and uniqueness)**.** *There is always 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 coefficients α, β, and δ are differentiable in the 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 the transition speed of the rating, and κ is the transition speed of the quality. The required inequality is reasonable as long as the rating system is meant to help estimate the current quality. For instance, even if the true quality does not drift much (i.e., $\kappa \simeq 0$), the rating should drift toward the underlying true quality (i.e., $\phi > 0$).

Intuition of the Equilibrium Strategy In Theorem 1, it is shown that high-quality types make more fake reviews ($\alpha > 0$), conditional on its reputation level. and high-reputation type makes fewer fake reviews ($\beta < 0$) conditional on the quality type. Given the logic of Nelson (1970; 1974), $\alpha > 0$ (and $\beta < 0$) might look intuitive, but this model adds different reasons than the previous research.

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

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

Then, $\beta = -\frac{\tau}{c} + \frac{2a\lambda}{c} v_4$. Furthermore, the envelope condition gives an expression for v_4 so that it is rewritten as $\beta = -\frac{\tau}{c} - \frac{\tau}{c} \frac{a\beta\lambda}{(-a\beta\lambda + r + 2\phi)}$. 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 the fake review is high. Therefore, the seller will make fewer fake reviews given a higher reputation. The second term corresponds to the fake review's 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 increases with the reputation. This is because the future self will reduce the amount of fake reviews after observing the boosted reputation due to today's fake reviews. Furthermore, this effect increases with M_t because the future reputation M_{t+dt} tends to be high given a high M_t, so the interaction term

$$\tau M_{t+dt} F_{t+dt} = \alpha\tau M_{t+dt}\theta_{t+dt} + \tau\beta M_{t+dt}^2 + \delta\mu\tau M_{t+dt} \tag{8}$$

decreases quadratically given a negative β. It turns out that the first term dominates the second term; thus β remains negative.

The intuition of positive α comes from the complementarity between the quality, θ, and the reputation, M, in the seller's value function. With high quality θ_t today, the reputation in the future tends to be higher than the case with low quality today, given the same level of reputation M_t today. Furthermore, as previously stated, the future benefit from the reputation boost is higher given a higher reputation in the future. Thus, high quality results in a high incentive for fake reviews. Mathematically, the equilibrium coefficient α is characterized as

$$\alpha = a\lambda v_5 = \frac{a\lambda}{\kappa + r + \phi} \{2 (a\alpha + bq) \lambda v_4 - \alpha\tau\} \tag{9}$$

The first equality reveals that the sign of α comes from the complementarity of θ and M in the value function. In the last expression, $(a\alpha + bq)\lambda$ indicate that the high θ_t results in a high M_{t+dt}. It is multiplied with positive v_4, which represents an increasing marginal value with respect to M_{t+dt}. This is the driving force of the positive α. The remaining term of Eq. (9), $-\alpha\tau$, 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; thus, today's fake reviews increase the cost in the future via the first term of Eq. (8).

In summary, the driving force of $\beta < 0$ is the incentive to reduce $\tau M_t F_t$ today given a high M_t.

α is positive because of the complementarity of θ_t and M_t through cost savings. Readers might wonder why an increase in revenue (like Nelson, 1970, 1974) does not appear in the above argument. 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 a high reputation through 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 the normative properties of the equilibrium, we check some positive properties of the equilibrium.

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

Proposition 1. *$E[F_t]$ increases with L and L increases with a. Furthermore, $E[F_t] \geq 0$ holds for sufficiently large a.*

Thus, the model can represent a reasonable situation under some parameters where fake reviews have non-trivial effect (i.e., a is significantly high). There still remains a small probability that F_t becomes negative due to the normal distribution, but the model can approximate a reasonable distribution of the fake reviews, under which the negative revenue is rarely observed, as shown in Fig. 2.

The precision of "organic" feedback from normal customers also monotonically changes the expected amount of fake reviews. When the organic feedback from customers varies a lot, it is hard for the seller to manipulate the reputation because 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 decreases the expected amount of fake reviews, as shown in Proposition 1, it does not imply that the seller's strategy gets closer to the no-fake strategy of

$\{\alpha, \beta, \delta\} = \{0, , 0, 0\}$. Moreover, the stringent policy might have unintentional effects of increasing the absolute value of the equilibrium coefficients.

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

Under a stringent policy (small a), the marginal benefit of fake review decreases because fake reviews are reflected less in the rating; but at the same time, the dependence of the marginal benefit on the current reputation also decreases. Mathematically, the second term of $\beta = -\frac{\tau}{c} + \frac{\tau}{c}\frac{-a\beta\lambda}{(-a\beta\lambda + r + 2\phi)}$ decreases while the marginal cost still depends on the current reputation regardless of the censoring policy. Therefore, $|\beta|$ increases owing to the less countervailing effect.

In the proof of the proposition, the intensity of dynamic consideration is also captured by an aggregator $L = -a\lambda\beta$, which is the equilibrium effect on the reputation transition speed. L becomes smaller when the dynamic incentive becomes smaller; thus, 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 increases in $\frac{a\tau}{c}$ and decreases 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 does not necessarily increase in a because α is a function in a and L, so the change in a affects directly and indirectly via L, and the net impact is not clear. $|\beta|$ does not necessarily decrease in $\frac{\tau}{c}$ for an analogous reason even though a limit of $\tau \to 0$ is known.

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

Some limits of the equilibrium strategy are worth noting before jumping into a normative analysis. Since the negative β comes from the interaction term in the cost of the fake reviews, whose coefficient is τ, β approaches zero as τ approaches zero. At the same time, α also approaches zero because the complementarity of θ and M is caused by future cost savings via negative β. In this limit, the fake reviews become constant as in Holmström (1999). This is summarized in the following proposition.

Proposition 4. $|\alpha|, |\beta| \to 0$ *as* $\tau \to 0$.

3.3 Optimal Rating System for Rational Consumers

In this study, we focus on the informativeness of the rating system as a normative criterion for two reasons. First is from the viewpoint of consumer protection: as the rating system gets more informative about the quality of a product, the price is likely to be close to the underlying quality. Thus, it becomes less likely that consumers would face huge regret from the purchase of the product. Second is from the viewpoint of the platform: the informativeness of the rating is crucial to attracting consumers in the long run. If consumers find it uninformative, they, as well as the sellers, can move to other platforms, given less consumers in the market. Thus, the informativeness of the rating would be the first priority when the platform controls it.

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

Given an equilibrium strategy, the stochastic differential equations—Eqs. (1) and (5)—give us ρ^2 as a function of the parameters and the equilibrium strategy. Therefore, the change of a parameter directly affects ρ^2 and indirectly affects it via a change of the equilibrium strategy. Fortunately, by representing the equilibrium coefficients α and β as functions of the equilibrium aggregator $L = a\beta\lambda$, all the direct and indirect effects of the censorship (a) are expressed as an effect through L. Comparative statics about other parameters, such as ϕ and σ_ξ/σ_θ, can also be examined by the indirect effect through L and the direct effect.

Lemma 3. *At the equilibrium,* ρ^2 *is expressed as a function:*

$$\rho^2(L;\,\phi,\,\kappa,\,\sigma_\xi,\,\sigma_\theta,\,r,\,bq) = \frac{(\phi+L)}{(\kappa+\phi+L)}\frac{(A(L;\,\phi,\,\kappa,\,r,\,bq)+1)^2}{((A(L;\,\phi,\,\kappa,\,r,\,bq)+1)^2+\kappa(\sigma_\xi/\sigma_\theta)^2(\kappa+\phi L))}$$

on which a, c, τ *have effects only through* L.

$A(L;\,\phi,\,\kappa,\,r,\,bq)$ *summarizes all the direct and indirect effects of* a *on the informativeness as a function of* L.

3.3.1 Filtering/Censoring Reviews

First, we analyze the impact of a filtering/censoring policy, a. Do fake reviews damage the informativeness of the rating system compared with the case without fake reviews? Does filtering or censoring the reviews (i.e., decrease in a) increase the rating's informativeness?

As a benchmark, we derive informativeness *without* fake reviews. By construction, we can do this by letting $\alpha = \beta = \delta = 0$.[13] The same informativeness is also replicated by setting $L = 0$ in $\rho^2(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta)$ to make it easier to compare with the informativeness at the equilibrium.

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

Note that $L = 0$ does not necessarily mean $\alpha = \beta = \delta = 0$. For instance, L approaches 0 as a approaches 0; but at the same time, β converges to some negative value. The lemma says that even under such a situation, informativeness is the same as that without fake reviews. Lemma 2, which is about the relationship between L and parameters, and Lemma 4 together lead us to the following proposition:

Proposition 5. *The informativeness of the rating system in 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 the no-fake strategy.

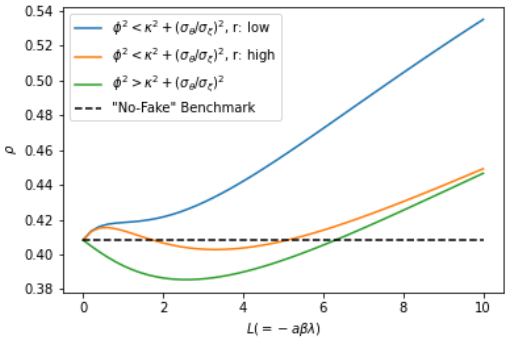
By analyzing the behavior of $\rho^2(L;\, \phi,\, \kappa,\, \sigma_\xi,\, \sigma_\theta)$ with respect to L, we can conclude that the informativeness can be even higher under some parameters where a positive amount of the fake reviews is expected. In other words, stringent censorship can decrease the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

Proposition 6. *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 $\phi^2 < \kappa^2 + \frac{\sigma_\theta^2}{\sigma_\xi^2}$*

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

Figure 3: Change of the informativeness in the aggregator L



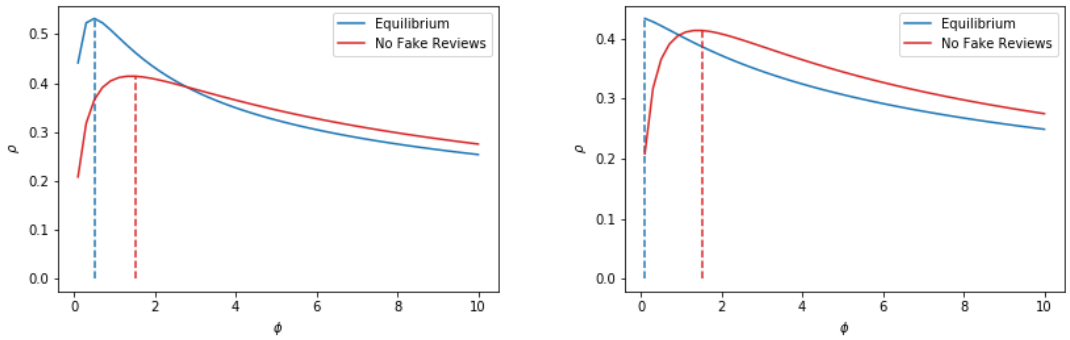
The graph indicates that the informativeness is (i) increasing in L if ϕ and r are relatively low, (ii) increasing in L around zero, then decreasing, and then increasing if ϕ is relatively low but r is relatively high, and (iii) decreasing in L around zero and then increasing in L if ϕ is relatively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

Fig. 3 shows the behavior of ρ^2 with respect to L. The first part of the proposition comes from the fact that ρ^2 converges to 1 as L approaches infinity. Since L is increasing $\frac{a\tau}{c}$ from zero to infinity, the equilibrium informativeness surpasses that of the no-fake benchmark at some point as $\frac{a\tau}{c}$ increases. The second part is derived from the behavior of ρ^2 around $L = 0$. The derivative of ρ^2 with respect to L 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 decreases in L; thus, decreases in $\frac{a\tau}{c}$. [14]

The intuition of this proposition consists of two parts: (i) As mentioned in Subsection 3.2.1, the sensitivity of fake reviews to θ_t decreases as the strategic effect L decreases. Thus, the strict censorship policy, which reduces the equilibrium aggregator L, decreases the signaling effect of the fake reviews. (ii) Meanwhile, $L > 0$ increases the effective transition speed of reputation to $\phi + L$. It can be good or bad in terms of informativeness, depending on the original transition speed, ϕ. More specifically, the threshold of $\sqrt{\kappa^2 + \sigma_\theta^2/\sigma_\xi^2} \equiv \phi^0$ is the informativeness-maximizing ϕ, given no fake reviews. Therefore, if ϕ is smaller than ϕ^0, the faster transition improves informativeness. It turns out that the first effect dominates in the case of a large L and the second effect dominates in the case of L close to zero.

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

Figure 4: 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.

3.3.2 Weights on New/Previous Reviews

Next, we analyze the optimal weights of the new and old 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 to this equation. Meanwhile, at equilibrium, ϕ changes the equilibrium aggregator L. Thus, the optimal weight at equilibrium is characterized by $\frac{d\rho^2}{d\phi}=\frac{\partial}{\partial\phi}\rho^2(L;\phi,\kappa,\sigma_\xi,\sigma_\theta)+\frac{\partial}{\partial L}\rho^2(L;\phi,\kappa,\sigma_\xi,\sigma_\theta)\frac{dL}{d\phi}=0$. Let the solution of this equation be ϕ^*. Now, we have the following proposition. [15]

Proposition 7. $\frac{d\rho^2}{d\phi}<0$ *at* $\phi=\phi^0$. *Furthermore, if* r *is sufficiently small, then* $\rho^2(L(\phi^*);\phi^*,\kappa,\sigma_\xi,\sigma_\theta)>\rho^2(0;\phi^0,\kappa,\sigma_\xi,\sigma_\theta)$.

The first part of the proposition states that the platform should reduce the speed of transition ϕ, given the existence the fake reviews. Intuitively, this is explained as follows. At equilibrium, the transition of the rating score Y_t is $\phi+L$ where L is non-negative. Therefore, to cancel the strategic impact on the transition speed, the platform should decrease ϕ, compared with the no-fake benchmark ϕ^0. Again, the transition speed is interpreted as the relative weight of the new information. At the equilibrium, the number of fake reviews decreases in the current rating; thus, the fake reviews cancel the past performance. In other words, the new information is effectively weighted more than the platform intends. Thus, the platform can increase the informativeness by

[15] ϕ corresponding to disaggregated information, ϕ^d, is an alternative benchmark as in Bonatti and Cisternas (2020). In this model, we obtain a mixed result for the comparison of ϕ^* and ϕ^d. See the appendix for more details.

adjusting it downward.

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 than the no-fake review benchmark by adjusting the speed of updating the rating. The implication is similar to Proposition 5, but is slightly different from it. The right panel of Fig. 3 illustrates that informativeness at equilibrium is greater than that without fake reviews under some parameters (e.g., $\phi = 0.9$), as shown in Proposition 5, but it can still be lower than the maximum informativeness without fake reviews (maximized around $\phi = 1.6$). The second part of Proposition 6 states that even when we compare the maximum informativeness of the rating with and without fake reviews, the one with fake reviews will be higher if the seller cares enough about the future as shown in the left panel of Fig. 3.

3.3.3 The Precision of Genuine Reviews

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

Proposition 8. *The informativeness at the equilibrium decreases in* $\frac{\sigma_\xi}{\sigma_\theta}$.

Thus, the precise organic feedback increases 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 model, but in the context of customer reviews, it is natural to consider the impact on naive consumers who do

not expect any fake reviews. The regulator often tries to protect customers from fake reviews, with the 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. They are modeled by assuming that the reputation (and the price) is characterized as $\widetilde{M}_t = \mu + \widetilde{\lambda}\left[Y_t - \widetilde{\nu}\right]$ where $\widetilde{\lambda}$ and $\widetilde{\nu}$ are characterized by the stochastic differential equations Eqs. (1) and (5), where $\alpha = \beta = \delta = a = 0$. Meanwhile, the long-lived seller faces the same problem as in the previous chapter, except for the definition p_t.[16]

4.1 Model / Equilibrium Characterization

In this section, the price is assumed to be a convex combination of a rational reputation M and a naive reputation $\widetilde{M}$.

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

One interpretation is that each consumer can be partially rational. Their expectation about the quality of the product is somewhere in between the totally sophisticated expectation and the totally naive expectation. The rationality of each consumer is captured by η.

Another interpretation is that η is the ratio of rational consumers among all consumers. Then, the market price is set somewhere in between the rational expectation and the naive expectation. When the ratio of rational consumers increases, it converges to the rational expectation. The linear specification captures such a relationship in a simple manner. Furthermore, it can be rationalized as an equilibrium price given a specific utility function of buyers. Suppose that there are n consumers in the market and $\eta \cdot n$ of them are rational and the other $(1-\eta)\cdot n$ are naive. Consumer $i \in [0,\, n]$ feels $u_{t,i} = \theta_t + \epsilon_{t,i} - p_t$ if the consumer buys the product, and 0 otherwise, where $\epsilon_{t,i}$ is identically and independently distributed. Rational and naive consumers differ only in terms of how they form

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

their expectation based on the same observation of the rating Y_t. Conditional on Y_t, a rational consumer purchases the product if and only if $M_t + \epsilon_i - p \geq 0$, while a naive consumer purchases it 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 the 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 because Y keeps track of both M and $\widetilde{M}$ in a simple manner:

$$\begin{aligned} rV(\theta, Y) = \sup_{F \in \mathbb{R}} & (1 - \tau) p \cdot q - \tau p \cdot F - \frac{c}{2} F^2 \\ & - \kappa(\theta - \mu) V_\theta \\ & + \{-\phi Y_t + aF_t dt + bq\theta_t\} V_Y \\ & + \frac{\sigma_\theta^2}{2} V_{\theta\theta} \\ & + \frac{b^2 q^2 \sigma_\xi^2}{2} V_{YY} \end{aligned} \tag{10}$$

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

Theorem 2. *For any $\eta \in [0, 1]$, a stationary linear Markov equilibrium always exists. 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 coefficients α, β, and δ are differentiable in the 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, surprisingly, the existence of naive consumers reduces the seller's strategic behavior.

Proposition 9. *The equilibrium with naive consumers ($\eta \in [0, 1)$) generates a smaller $|\alpha|$, a larger $|\beta|$,and a smaller $E[F_t]$ compared with the equilibrium without naive consumers ($\eta = 1$).*

This is because rational consumers are more sensitive to the change in ratings compared with naive consumers. Rational consumers know that the rating is boosted, but they also know that the rating is boosted more by a firm with a high quality product. Therefore, rational consumers attribute the boosted rating to high quality, and set a high price for such a boosted rating. Meanwhile, naive consumers are unaware of such a strategic correlation between quality and a rating. Therefore, with naive consumers, the price is less responsive to the boost of the ratings; thus, the seller faces a smaller *marginal* benefit of fake reviews, which leads fewer fake reviews in expectation.

Readers might wonder why the seller does not become more exploitative of naive consumers. This is simply because the fake review strategy against rational consumers generates more fake reviews for different reasons than exploiting consumers. If only a small number of naive consumers exist and observe the ratings, naive consumers would form even more biased estimates because the seller makes more fake reviews to send a good signal to rational consumers.

4.2 Optimal Rating System for Naive Consumers

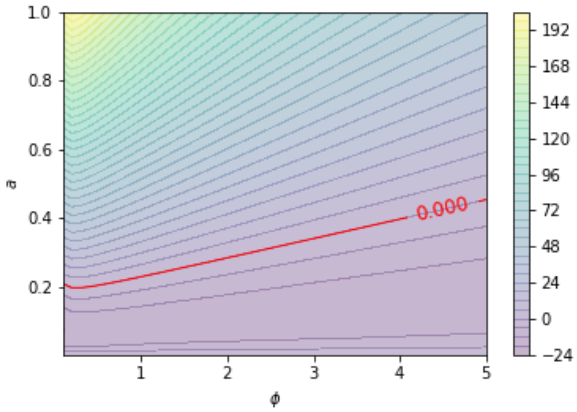
Criteria: Bias in the Reputation. In this section, we evaluate the impact of fake reviews on naive consumers. To do so, we introduce a bias in the naive consumer's expectation caused by the boosted rating:

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

where $\widetilde{\lambda}$ is the sensitivity of the reputation to the rating, and ν and $\widetilde{\nu}$ are the actual mean of the rating and the estimate of the mean of the rating by the naive consumers, respectively. The decomposition of the bias, as shown above, is intuitive: the positive bias is due to the boosted reputation. Because naive consumers do not expect any fake reviews, they interpret a high rating ($Y_t > \widetilde{\nu}$) as a result of high quality, even though it is actually the average level of the rating at equilibrium ($Y_t = \nu > \widetilde{\nu}$).

Therefore, as long as the seller makes a positive amount of fake reviews (in expectation) to boost the rating, naive consumers are positively biased. This intuition is verified in the following lemma.

Figure 5: Impact of censorship intensity and the weights of reviews on naive consumer's bias.



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

4.2.1 Filtering/Censoring Reviews

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

First, we examine the impact of a filtering policy, for which regulators are arguably concerned the most. The following proposition provides a theoretical background of a stringent policy that procect the naive customers. Note that even though the statement seems pretty intuitive, it is not trivial because the model predicts a non-monotonouse relationship between censorship and bias in general. Fortunately, in a realistic range of parameters, where naive consumers suffer from a positive bias in their reputation, stringent censorship will reduce such a bias.

Proposition 10. *Suppose Bias* ≥ 0*; then, Bias increases in a.*

Combined with Lemma 5, the condition for a stringent policy to work for naive consumers is translated as the condition of a measure observable by the platform.

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

Thus, as long as a positive number of fake reviews are observed, the stringent policy is beneficial for naive consumers, even though it can reduce informativeness of rating for the rational consumers.

4.2.2 Weights on New/Previous Reviews

As shown in Fig. 5, the bias tends to be hump-shaped in ϕ. This is intuitive because fake reviews 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 emphasizes that the trade-off between bias and informativeness can be an inherent feature of fake reviews.

Some readers might want an integrated criteria for bias and the informativeness. The mean squared error (MSE) is a natural candidate. It does not provide a clear-cut prediction, but a simulation of MSE is provided in the Appendix.

5 Conclusions

In this study, the effects of fake reviews on rational and credulous consumers are analyzed. The key assumption is that a high reputation results in a high cost of fake reviews. This is rationalized by the high reimbursement to reviewers or high demand for the product and the substantial, authentic feedback crowding-out the fake reviews.

At equilibrium, the amount of fake reviews increases (decreases) as product quality (firm reputation) increases (improves), which implies difficulties in the empirical analysis of signaling promotion. Stringent censorship reduces the expected amount of fake reviews, while decreasing the signaling effect and increasing the effective transition speed of the rating.

This leads to a normative result wherein the rating under a less strict filtering policy can be more informative than the rating under a strict policy or the rating with no fake reviews. In terms of the weights of new and old information in a rating system where fake reviews exist, the platform should reduce the weight of new information to maximize the informativeness of the rating, compared with a rating system that does not have fake reviews. Since fake reviews effectively attenuate the impact of old information and increase the relative weight of the new information, the platform should make the necessary adjustments.

The existence of credulous consumers decreases the expected amount of fake reviews since they are less responsive to the rating without being aware of the positive relationship between fake reviews and the quality. Moreover, they are vulnerable to fake reviews and pay more than the true

quality in expectation. The model predicts that as long as a positive amount of the fake reviews is observed, the regulator or the platform can reduce such biased behaviors by enhancing censorship.

The results emphasize that regulators or platforms would face a trade-off between the degree of informativeness and the bias caused by fake reviews. As long as the rating is considered informative, the incentive to make fake reviews arises.

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*, 2020, *87* (2), 750–791.

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.

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

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

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.

He, Sherry, Brett Hollenbeck, and Davide Proserpio, "The Market for Fake Reviews," *SSRN Electronic Journal*, 2020, pp. 1–38.

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*, 2021, *88* (4), 1892–1935.

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

$$V = v_0 + v_1\theta + v_2 M + v_3\theta^2 + v_4 M^2 + v_5\theta M \tag{11}$$

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,

$$\frac{c}{a\lambda}\alpha = v_5 \tag{12}$$

$$\frac{c}{2a\lambda}\left(\beta + \frac{\tau}{c}\right) = v_4 \tag{13}$$

$$\frac{\delta\mu c}{a\lambda} = v_2 \tag{14}$$

By the Envelop condition w.r.t. M,[17]

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

By inserting the derivatives of eq.(11) and equating the coefficients of θ, M, and constants on LHS and RHS,

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

[17]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.

Then, inserting eq.(12) to eq (14),

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

$$2\,(r+\phi)\frac{c}{2a\lambda}\left(\beta+\frac{\tau}{c}\right) = -\tau\beta + \{a\lambda\beta - \phi\}\, 2\frac{c}{2a\lambda}\left(\beta+\frac{\tau}{c}\right) \tag{16}$$

$$(r+\phi)\frac{\delta\mu c}{a\lambda} = (1-\tau)\,q - \tau\delta\mu + \kappa\mu\frac{c}{a\lambda}\alpha + \{a\lambda\delta\mu + \phi\mu - \phi\lambda\nu\}\, 2\frac{c}{2a\lambda}\left(\beta+\frac{\tau}{c}\right) \tag{17}$$

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 aggregator $L = -a\beta\lambda$ so that the stationarity condition is easier to verify. First, by replacing λ to $-\frac{L}{a\beta}$ in the above four equations,

$$0 = -\frac{bq(\beta c+\tau)}{a} + \alpha\tau - \alpha(\beta c+\tau) - \frac{\alpha\beta c\kappa}{L} - \frac{\alpha\beta c\phi}{L} - \frac{\alpha\beta cr}{L} \tag{18}$$

$$0 = \beta\tau - \beta(\beta c+\tau) - \frac{2\beta\phi(\beta c+\tau)}{L} - \frac{\beta r(\beta c+\tau)}{L} \tag{19}$$

$$0 = \frac{\nu\phi(\beta c+\tau)}{a} - \delta\mu(\beta c+\tau) + \frac{\alpha\beta c\kappa\mu}{L} - \frac{\beta c\delta\mu\phi}{L} + \frac{\beta\mu\phi(\beta c+\tau)}{L} - \frac{\beta c\delta\mu r}{L} + \delta\mu\tau + q\tau - q \tag{20}$$

$$-\frac{L}{a\beta} = \frac{\sigma_\theta^2(L+\phi)(a\alpha+bq)}{\sigma_\theta^2(a\alpha+bq)^2 + \kappa bq\sigma_\xi^2(\kappa+L+\phi)} \tag{21}$$

By solving (19) 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 (18) 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 (21), 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 $-a\beta\lambda \equiv L > 0$. □

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 $h'(L) < 0$ □

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

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 aggregator 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 $E[F_t]$ only through the aggregator 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 $E[F_t]$ increasing in a and decreasing in $\frac{\sigma_\xi}{\sigma_\theta}$.

Since $E[F_t] > 0$ for sufficiently large L and $L \to \infty$ as $a \to \infty$, $E[F_t] > 0$ 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

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

Thus, the informativeness without fake reviews is

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

.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:

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

Note that $\rho^2\left(0;\,\phi,\,\kappa,\,\sigma_\xi,\,\sigma_\theta\right)=\frac{\phi(bq)^2}{(\kappa+\phi)\left((bq)^2+\kappa bq\left(\sigma_\xi/\sigma_\theta\right)^2(\kappa+\phi)\right)}$ coincides 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 characterized 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: $\rho^2\left(0;\,\phi^0,\,\kappa,\,\sigma_\xi,\,\sigma_\theta\right)<\rho^2\left(L\left(\phi^0\right);\,\phi^0,\,\kappa,\,\sigma_\xi,\,\sigma_\theta\right)\leq$

$\rho^2\left(L\left(\phi^*\right);\,\phi^*,\,\kappa,\,\sigma_\xi,\,\sigma_\theta\right)$. 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. Now, we look for a quadratic value function

$$V=v_0+v_1\theta+v_2Y+v_3\theta^2+v_4Y^2+v_5\theta Y \tag{22}$$

satisfying the HJB equation:

$$\begin{aligned}rV\left(\theta,\,Y\right)=&\sup_{F\in\mathbb{R}}\left(1-\tau\right)p\cdot q-\tau p\cdot F-\frac{c}{2}F^2\\&-\kappa\left(\theta-\mu\right)V_\theta\\&+\left(aF+bq\theta-\phi Y\right)V_Y\\&+\frac{\sigma_\theta^2}{2}V_{\theta\theta}\\&+\frac{bq\sigma_\xi^2}{2}V_{YY}\\&\text{s.t. }p=\mu-\left(\eta\lambda+\left(1-\eta\right)\tilde{\lambda}\right)Y+\left(\eta\lambda\nu+\left(1-\eta\right)\tilde{\lambda}\tilde{\nu}\right)\end{aligned}$$

The first order condition and gives

$$v_5 = \frac{\alpha c}{a} \tag{23}$$

$$v_4 = \frac{\beta c + \hat{\lambda}\tau}{2a} \tag{24}$$

$$v_2 = \frac{c\delta\mu + \mu\tau - \widehat{\lambda\nu}\tau}{a} \tag{25}$$

where $\hat{\lambda} = \left(\eta\lambda + (1-\eta)\,\tilde{\lambda}\right)$ and $\widehat{\lambda\nu} = \left(\eta\lambda\nu + (1-\eta)\,\tilde{\lambda}\tilde{\nu}\right)$, and the envelop condition gives

$$0 = \hat{\lambda}\alpha\tau - 2a\alpha v_4 - 2bqv_4 + rv_5 + \kappa v_5 + v_5\phi \tag{26}$$

$$0 = -2a\beta v_4 + \beta\hat{\lambda}\tau + 2rv_4 + 4v_4\phi \tag{27}$$

$$0 = -2a\delta\mu v_4 + \delta\mu\hat{\lambda}\tau + \hat{\lambda}q\tau - \hat{\lambda}q + rv_2 - \kappa\mu v_5 + v_2\phi \tag{28}$$

By inserting eq.(24) into (27) and solving it for $\hat{\lambda}$ and by letting $L = a\beta$, we obtain

$$\hat{\lambda} = \frac{cL(L + r + 2\phi)}{a\tau(r + 2\phi)} \equiv \hat{\lambda}\,(L)$$

On the other hand, the stochastic differential equation for $(\theta,\, Y)$ gives

$$\lambda = \frac{bq\sigma_\theta^2(L+\phi)\,(A\,(L)+1)}{\sigma_\theta^2\,(bqA\,(L)+bq)^2 + \kappa bq\sigma_\xi^2(\kappa + L + \phi)} \equiv \lambda\,(L)$$

$$\tilde{\lambda} = \frac{bq\sigma_\theta^2\phi}{\sigma_\theta^2\,(bq)^2 + \kappa bq\sigma_\xi^2(\kappa+\phi)} = \lambda\,(0)$$

Then, by rearranging

$$\hat{\lambda} = \left(\eta\lambda + (1-\eta)\,\tilde{\lambda}\right)$$

$$\Rightarrow 1 = \frac{\eta\lambda\,(0) + (1-\eta)\,\lambda\,(L)}{\hat{\lambda}\,(L)} \equiv h\,(L;\eta)$$

Note that $\lim_{L\to 0} h\,(L;\eta) = \infty$ and $\lim_{L\to\infty} h\,(L;\eta) = 0$. Then, $h_L\,(L;\eta) < 0$ holds for any $\eta \in [0,\, 1]$ as long as $h_L\,(L;1) < 0$. □

Proof of Proposition 8. Since $\lambda(0) \leq \lambda(L)$ for any $L \geq 0$, we have $h(L;\eta) \leq h(L;1)$ for any $\eta \in [0, 1]$. Thus, the equilibrium L will be smaller given $\eta < 1$ than the equilibrium L given $\eta = 1$.

The expected amount of the fake reviews is

$$E[F_t] = \alpha\mu + \beta\nu + \delta\mu$$

By plugging the equilibrium conditions and taking derivative with respect to L, we can show $\frac{\partial}{\partial L} E[F_t] \geq 0$. □

Proof of Proposition 9. At the equilibrium, $\frac{\partial bias}{\partial L} \geq 0$ always holds and $\frac{\partial bias}{\partial a} \geq 0$ holds if $bias \geq 0$. □

B An interpretation of the pricing rule

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.

C An Alternative Model with Changing q

The same results with the base line model can be generated with a slightly different specification of the model with the quantity level dependent on the reputation level.

Now, suppose that the seller sells q_t units of the product at a fixed price of p, and makes F_t units of fake reviews. The quality of the product is denoted as θ_t. A sufficiently large mass of consumers

forms a belief on the quality $E[\theta_t|Y_t] \equiv M_t$ and the demand function based on that. Since the price is fixed, high reputation results in large quantity: $q_t = M_t$.

The quality θ_t evolves in the same way as the main model. The new information as

$$aF_t dt + bq_t \left(\theta_t dt + \sigma_\xi dZ_t^\xi\right) \tag{29}$$

The difference from the main model is that the quantity varies over time and the coefficient of dZ_t^ξ is now defined as $bq_t\sigma_\xi$ instead of $\sqrt{bq_t}\sigma_\xi$. In this specification, we can analyze the effect of the organic reviews crowding out the fake reviews, but not the effect of the large transaction generating intrinsically more precise information by the large sample.

The seller's instantaneous payoff is defined as:

$$\pi_t = (1-\tau)\, p\,(q_t + F_t) - p \cdot F_t - \frac{c}{2}\left(\frac{F_t}{q_t}\right)^2$$

where τ is transaction fees imposed by the platform. The specification of the quadratic cost is now different from the base line model: the seller needs to pay a large cost if the seller tries to increase the share of the fake reviews among the all the reviews. The revenue and the reimbursement cost is still the same as the baseline model.

$$\begin{aligned}\pi_t &= (1-\tau)\, pq_t - \tau p \cdot F_t - \frac{c}{2}\left(\frac{F_t}{q_t}\right)^2 \\ &= (1-\tau)\, pM_t - \tau p \cdot M_t \frac{F_t}{M_t} - \frac{c}{2}\left(\frac{F_t}{M_t}\right)^2\end{aligned}$$

By changing the choice variable of the seller from F_t to $\frac{F_t}{M_t}$, which is the combination of the original variable and a constant at time t, we can write the instantaneous profit isomorphic to one in the baseline model. To simplify the analysis, we assume that the platform use an average information at time t to update the ratings:

$$d\xi = \frac{a}{b}\frac{F_t}{M_t}dt + \theta_t dt + \sigma_\xi dZ_t^\xi \tag{30}$$

The model is then isomorphic to the baseline model, so generates the same results as those from

the baseline model.

D Simulation Results

D.1 Mixture of the Rational and Naive Consumers

In the main part, the correlation of the rating with the underlying true quality for rational consumers, and the bias for the naive consumers are examined. There is a trade-off of the correlation and the bias. Then, natural questions are (i) how to integrate such indices into one objective function, and (ii) how it changes as the market's rationality changes from totally naive to totally rational. In this section, we suggest a mean squared error of the price since the price is considered as the whole market's prediction about the underlying quality. The minimization of the mean squared errors minimizes the customers' *ex post* regret on average, so increases the value-added of the platform, and attracts the customers in long-run.

D.1.1 Mean Squared Error

The mean squared errors of the price is defined and written with the equilibrium variables as follows:

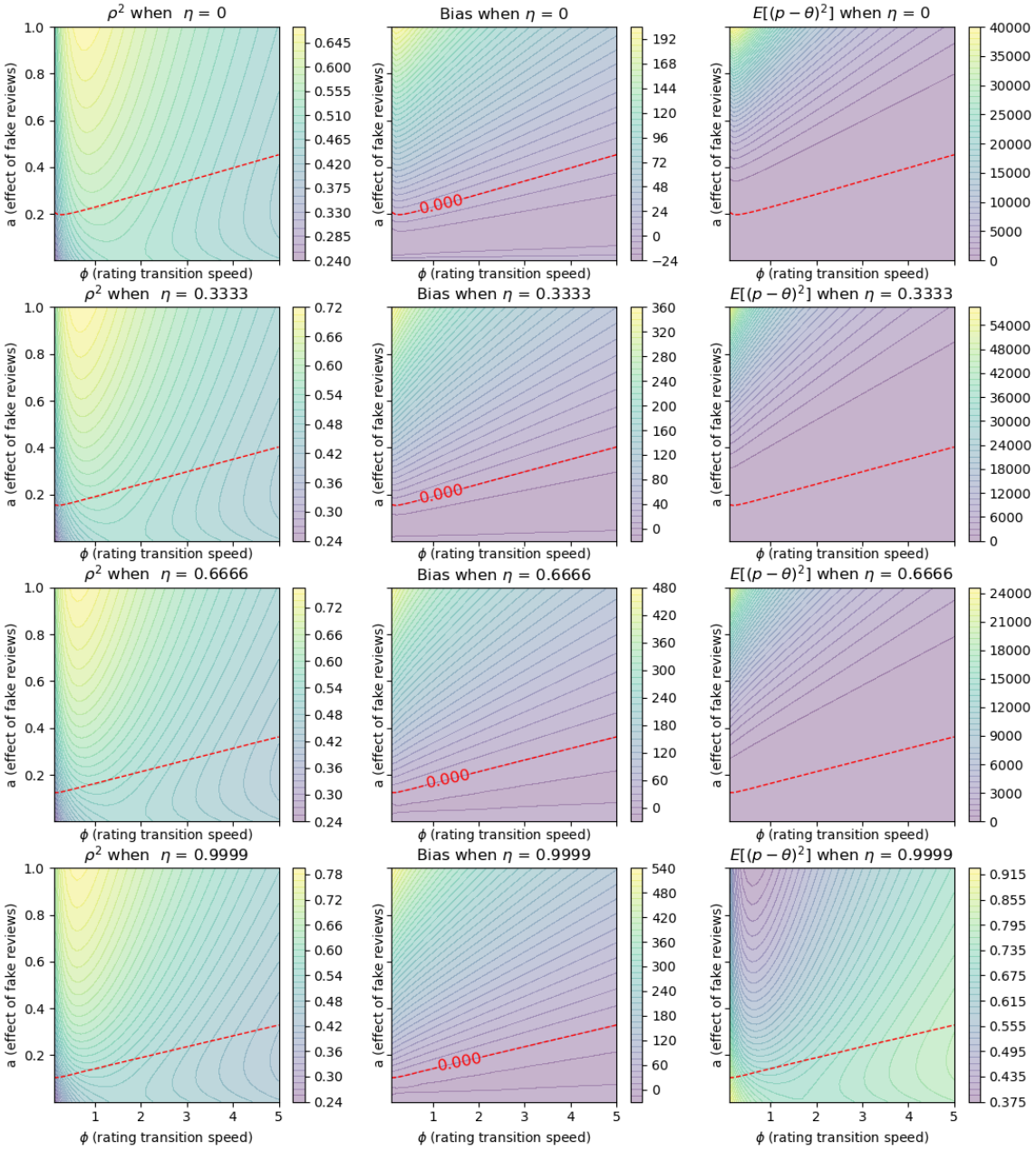
$$\begin{aligned} MSE_p &= E\left[(p_t - \theta_t)^2\right] \\ &= E\left[\left(\eta\left\{\mu + \lambda\left[Y_t - \nu\right]\right\} + (1-\eta)\left\{\mu + \widetilde{\lambda}\left[Y_t - \widetilde{\nu}\right]\right\} - \theta_t\right)^2\right] \\ &= Var\left(Y_t\right)\left\{\left(\eta\lambda + (1-\eta)\widetilde{\lambda}\right)^2 - 2\left(\eta\lambda + (1-\eta)\widetilde{\lambda}\right)\lambda\right\} + (1-\eta)^2 Bias^2 + Var\left(\theta_t\right) \end{aligned}$$

Note that, when $\eta = 1$, minimization of MSE is reduced to maximization of the correlation of the rating Y_t and θ_t:

$$\begin{aligned} MSE_p &= -\lambda^2 Var\left(Y_t\right) + Var\left(\theta_t\right) \\ &= Var\left(\theta_t\right)\left\{1 - \frac{Cov\left(Y_t, \theta_t\right)^2}{Var\left(Y_t\right)^2}\frac{Var\left(Y_t\right)}{Var\left(\theta_t\right)}\right\} \\ &= Var\left(\theta_t\right)\left\{1 - \rho^2\right\} \end{aligned}$$

For different levels of η, we calculate the correlation of Y and θ as a criteria for the rational consumers, the bias as a criteria for naive consumers, and the mean squared error as a criteria for the whole market. See fig.5 for the simulation results. The correlation of the rating with the underlying quality show the similar pattern regardless of the level of η, while it is scaled up as the rationality increases. So does the bias the naive consumers faces. This is consistent with Proposition 9. As the market becomes more rational, the consumers takes the signaling effect of the seller's fake reviews ($\alpha > 0$), so the market becomes more sensitive to the rating. Then, the seller will have more incentive to make fake reviews, resulting in more bias for naive consumers. At the same time, the signaling effect ($\alpha > 0$) is also enhanced by this increased manipulation by the seller. Therefore, the rating becomes more informative for rational consumers. Roughly speaking, the mean squared error integrates the correlation and the bias into one. As the ratio of the rational consumers increases, the correlation becomes more important. As the ratio of the naive consumers increases, the bias comes more important. Fig. 5 exhibits this. For $\eta = 0$, 0.3333, 0.6666, the MSE shows the similar pattern as the bias, while the MSE shows the similar pattern as the correlation for $\eta = 0.9999$. Given other parameters used in the simulation, the bias is the dominant force in MSE for most of η. This results depends on the parameter setting, so is ultimately an empirical question, but suggests that decreasing the bias is more important than increasing the informativeness for rational consumers.

Figure 6: Correlation, bias, and mean squared errors



From top to the bottom, the rationality of the market is increased from 0, 0.3333, 0.6666, to 0.9999. The left panels are contours of the correlation of the rating Y_t with θ_t based on rational expectations taking the seller's strategy into account. The middle panels show biases the naive consumers faces. The right panels show the mean squared errors of the market price as a whole market's prediction of the underlying quality. Red dashed lines border sets of parameters which predict realistic positive bias (positive number of positive fake reviews) at the equilibrium. Areas above red lines corresponds to the positive number of positive fake reviews.