University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 307

Bad News Turned Good: Reversal Under Censorship

Aleksei Smirnov and Egor Starkov

November 2018

Bad News Turned Good: Reversal Under Censorship*

Aleksei Smirnov†, Egor Starkov‡

November 13, 2018

Abstract

Sellers often have the power to censor the reviews of their products. We explore the effect of these censorship policies in markets where some consumers are unaware of possible censorship. We find that if the share of such "naive" consumers is not too large, then rational consumers treat any bad review that *is* revealed in equilibrium as good news about product quality. This makes bad reviews worth revealing and allows the high-type seller to use them as a costly signal of his product's quality to rational consumers.

Keywords: Censorship, dynamic games, disclosure, moderated learning

JEL Codes: D82, D83, D90

*Starkov is deeply indebted to Jeffrey Ely, Yingni Guo, and Wojciech Olszewski for continuous support and invaluable advice. We are also grateful to Simon Board, Benjamin Casner, Georgy Egorov, Alexey Makarin, David Miller, Nick Netzer, Georg Nöldeke, Alessandro Pavan, Marek Pycia, Rene Sarant, Armin Schmutzler, Bruno Strulovici, seminar participants at the University of Zürich and Northwestern University, and participants of 2017 Econometric Society European Winter Meeting, 2018 Midwest Economic Theory Conference, Econometric Society Summer School, 29th International Conference on Game Theory, and 11th Transatlantic Theory Workshop for valuable feedback and helpful comments.

†Department of Economics, University of Zürich, Blümlisalpstrasse 10, CH-8006, Zürich, Switzerland; e-mail: aleksei.smirnov@econ.uzh.ch.

‡JOB MARKET PAPER. Department of Economics, Northwestern University, 2211 Campus Drive, Evanston, IL 60208, USA; e-mail: starkov@u.northwestern.edu.

1 Introduction

Word of mouth has long been a significant source of information about product features and quality. One of its manifestations in the digital age is online product reviews. Opinions of fellow consumers often seem more trustworthy than sellers' product descriptions, and the sheer numbers of reviews offer a great diversity of viewpoints. However, sellers can undermine this learning channel, and one instrument they often have for doing so is censorship, i.e., removing unfavorable reviews of their own product.[1]

It is straightforward that whenever censorship is possible and its cost is low, it will be employed to at least some extent. A naive conjecture would be that if the seller can censor at will, then no meaningful bad reviews ever remain, and those that do convey absolutely no information. This is because the seller would delete any review that harms sales. However, in practice we observe plenty of informative bad reviews even when censorship opportunities exist (e.g., on sellers' own websites). This paper asks the following questions: first, why might the seller be willing to *not* censor unfavorable reviews? Furthermore, how is the informational content of such reviews affected by censorship? Finally, how can bad reviews increase sales?

Our paper answers these questions in the setting where some consumers do not account for the possibility of censorship itself.[2] We construct a model in which a long-lived seller offers for sale a good of privately known quality to a sequence of short-lived consumers. Consumption utility is suggestive about the product quality and is relayed to future consumers through product reviews, which may be deleted by the seller. Consumers differ in what inferences they make from posted reviews: some consumers are *strategic* – i.e., fully aware of the seller's censorship capabilities – while others are *naive* and ignore censorship.

The main result of the paper (Theorem 2) states that if naive consumers are present in the market but do not dominate it, then there exist equilibria in which bad reviews are revealed on the equilibrium path in a payoff-relevant way. The main driving force behind this phenomenon is explained by our second result (Theorem 1), which states that having no bad reviews should actually be perceived as *bad* by a strategic consumer. In other words, in any equilibrium, any strategic consumer *improves* her belief about the product quality upon observing any review that says the product is *bad*. Bad reviews are only revealed in equilibrium to signal product quality to strategic consumers, and this signal is only credible if it has the potential to hurt sales to naive consumers.

The reason why any bad review has to improve the seller's reputation is two-fold. The main idea is that it only makes sense for the seller to reveal a bad review if it does not have any detrimental effect on future sales. The *direct effect* can then be formulated as follows: if a bad review was published and it harms sales to naive consumers, then it should increase sales to strategic consumers, as otherwise it would not have been published. In particular, this increase in sales is attained by improving the product's reputation in the eyes of strategic consumers.

The *expectancy effect* is more involved. It boils down to the fact that revealing bad reviews creates *expectancy* among strategic consumers in the sense of them losing faith in the product faster,

[1] This is plausible when we are talking about the seller's own online store, where he has absolute power over the content posted on the website – including product reviews. However, censorship is possible in other settings as well, see Section 2 for the discussion.

[2] See Section 2 for empirical evidence of consumers' naivetè when making inferences from product reviews.

so the seller has to be compensated – reputation-wise – for exposing himself to this expectancy. To elaborate, the direct effect above states that driving naive consumers out of the market with a bad review is a strong positive signal for strategic consumers. Then *absence* of bad reviews in such a situation is a strong negative signal – strategic consumers become rapidly disenchanted about the product in the absence of bad reviews. This is what we call a state of high expectancy. Any bad review that brings naive consumers closer to quitting (without actually pushing them out) drives up the expectancy. This hurts the seller because it forces him to lose strategic consumers faster, hence he should be compensated for exposing himself to high expectancy – meaning that the seller's reputation among strategic consumers has to increase after a bad review even if he does not lose naive consumers' demand by revealing it.

Finally, the argument above explains what happens *if* a bad review is revealed in equilibrium, but does not explain why bad reviews actually work as a signaling device. Mechanically, Theorem 1 implies that the low type should be censoring more bad reviews than a high type, so the latter has to be [at least weakly] more willing to let bad reviews out. This is indeed the case: the high-type seller is less reluctant to reveal bad reviews and lose naive consumers because he is more confident about receiving good reviews in the future, which will bring naive consumers back to the market.

Of course, even in the context of product reviews, censorship is not the only way the seller can manipulate the information available to the consumer. Posting fake reviews, be it fake positive reviews of own product or fake negative reviews of competing products, is another activity the seller can engage in.[3] While we mostly focus on censorship in this paper, Section 6 shows that our result continues to hold in the presence of both censorship and fake reviews.

The main focus of this paper is on product reviews – and to keep things clear we will stick to this interpretation throughout – but the model translates naturally to other settings that feature censorship or dynamic disclosure of verifiable information. For example, instead of a seller censoring bad reviews, one may think about the government censoring news stories in an attempt to retain citizens' support. In the context of venture financing, the startup may choose whether to disclose temporary setbacks to the investors or not. A bank may disclose or withhold information about its temporary liquidity deficit in an attempt to prevent a bank run. Our paper implies that in all these settings if some receivers are naive, it may be beneficial for the sender to disclose bad news or failures, since rational receivers would take the mere fact of disclosure as a good signal.

The paper is organized as follows. Section 2 discusses the plausibility of censorship in product reviews and reviews the relevant literature. Section 3 presents a short example to convey the main idea of the paper. In Section 4 we formulate the full model. The main results are presented in Section 5. Section 6 contains some further discussion of the model and its extensions, while Section 7 concludes. All proofs are relegated to the Appendix.

2 Background and Literature Review

2.1 Censorship: Background

The main setting considered in the paper is that of a platform that the seller owns or has moderation rights in. Examples include seller's own website, forum, or Facebook page. In all of

[3] About 16% of Yelp reviews are marked as potentially fake (Luca and Zervas [2016]).

these cases the seller is able to remove bad reviews directly. Such deeds are by definition difficult to document, but some claims may be found.[4]

However, it is important to note that the seller does not need to have direct power to remove bad reviews – he merely needs to convince whoever has this power. For example, some platforms (such as Etsy) allow sellers to try to address buyers' dissatisfaction and ask buyers to remove their negative review if all issues were resolved. While most review aggregators (such as Amazon, Yelp or TripAdvisor) do not allow the sellers to directly remove reviews, convincing, bribing, or harassing consumers into deleting their own reviews are all viable options in those cases. Promising free items or politely asking to contact the company before writing a bad review both have a chance of succeeding at making the consumer remove or alter their bad review, or even not write one in the first place. One extreme method of consumer harassment is SLAPP – Strategic Lawsuit Against Public Participation, – when a seller sues a reviewer primarily to deter other critics from writing negative reviews. While these suits are rarely won in court, they are likely to succeed at forcing the person to delete their review before the suit even reaches court, and/or at intimidating other potential reviewers.[5]

Finally, in some settings the seller may get to choose more favorable reviewers – e.g., a movie distributor picks the critics that get to write the pre-release reviews. In this setting the seller may also ensure that no bad review gets through – either by screening the reviews directly, or by choosing ex ante more favorable reviewers, or through repeated interaction mechanisms.[6]

2.2 Censorship and Review Manipulations

Academic literature on manipulations in product reviews has focused on the issue of fake reviews (in part because those are easier to observe in the data than deleted reviews which are, by definition, missing from the sample). We are not aware of any papers that deal with censorship in product reviews explicitly, apart from Hauser [2018] who models censorship as depressing the rate of reviews arrival (i.e., censorship is indiscriminate in that model). In a spirit similar to censorship, Kovbasyuk and Spagnolo [2017] explore the effects that limited memory (in terms of old records being erased after some time) has on market outcomes. Literature on fake reviews and review manipulation is more prominent. Fake reviews are explored in Dellarocas [2006], Mayzlin [2006], Mayzlin, Dover, and Chevalier [2014], and Luca and Zervas [2016]. Effects of review manipulation are investigated in Aköz, Arbatli, and Çelik [2017] and Harbaugh, Maxwell, and Shue [2017].

Political censorship, however, has received a lot more attention in the literature. Sun [2018] explores a model of dynamic censorship similar to ours, but without naive receivers. Besley and Prat [2006] present a model, in which an incumbent may bribe the media to conceal a bad signal

[4] "Amazon looking into claims that employees delete bad reviews for cash" (ArsTechnica), "EA Rep Reportedly "Bribed" Reddit Mods to Remove Certain Star Wars Battlefront Posts" (PlayStation Lifestyle), and "Airbnb guests accuse it of deleting negative reviews and boosting bad hosts" (Quartz).

[5] Some examples of [unsuccessful] application of this technique are described in the following news articles from ArsTechnica: "Jeweler tries to sue anonymous woman who left 1-star Yelp review", "Router company that threatened a reviewer loses Amazon selling license", and "Valve bans developer from Steam after it sues customers over bad reviews". One case of such suit being won is described in: "Disgruntled bride ordered to pay $115K after defamatory posts ruin Chinese wedding-photo business" (CBC News). Again, it is by definition more difficult to find documented instances when such harassment was successful – not in the sense of suit being won but in the sense of it forcing consumers to remove their bad reviews.

[6] A recent case here is Disney banning LA Times from pre-release screenings of its movie in retaliation for other recent articles. More details in: "LA Times: You can't read our Thor review because Disney is mad" (ArsTechnica).

about himself, but focus on the effects of media diversity and independence on political outcomes. Other papers about political censorship include Shadmehr and Bernhardt [2015], Edmond [2013], Egorov, Guriev, and Sonin [2009], Eraslan and Ozerturk [2017], and Chen and Yang [2017], but all of them explore issues that are very different from those that we focus on.

2.3 Naivete

Evidence of consumers' naivete when making inferences from product reviews has been provided by Brown et al. [2012] and Li and Hitt [2008]. These papers show that at least some consumers ignore the correlation between other players' actions and their private information. This notion of naivete (which is also implied in our model) has been formalized by Eyster and Rabin [2005], who use the term "cursedness" for this type of irrationality. Market interactions under cursedness have been studied by, e.g., Baumann and Rasch [2017] and Ispano and Schwardmann [2018].

More broadly, a wide array of empirical and experimental literature has demonstrated that people play naively even in very basic disclosure games: see Jin [2005], Deversi, Ispano, and Schwardmann [2018], Jin, Luca, and Martin [2018], and Sheth [2018]. In particular, Deversi et al. [2018] and Jin et al. [2018] show that people on the receiving side of their disclosure game either play in a way that is very close to a rational player's strategy, or play in way that is very naive. I.e., their players can be separated relatively well into naive and rational receivers. At the same time, when on the senders' side all the same people play rationally. This means that naivete is very robust in disclosure setting: even after adopting the role of the sender and going through full strategic reasoning of the sender, people switch roles and still play a naive strategy as a receiver.

Finally, if we are talking about political censorship rather than product reviews, then it is worth mentioning that bipartite structure of the electorate (existence of naive voters alongside rational voters) is quite prominent in the political economy literature. Baron [1994] and Grossman and Helpman [1996] are the seminal models with this feature.

2.4 Disclosure

Our paper belongs to the literature on disclosure of verifiable information, since in our model the seller can only decide whether to disclose any bad reviews written by consumers, but cannot actually write fake reviews on his own. Static games of disclosure were originally studied by Grossman [1981], Dye [1985], and Jung and Kwon [1988]; for a recent survey of literature on static disclosure see Dranove and Jin [2010].

Papers in dynamic disclosure include Acharya, DeMarzo, and Kremer [2011], Guttman [2010], Guttman, Kremer, and Skrzypacz [2014], Orlov, Skrzypacz, and Zryumov [2016], and Gratton, Holden, and Kolotilin [2018]. All of them focus on the sender's (or senders') choice of timing of disclosure, while our main interest remains in the informational content of the message.

2.5 Good News, Bad News, and Countersignaling

Some of the empirical evidence points to the fact that bad reviews are not necessarily harmful: see Resnick, Zeckhauser, Swanson, and Lockwood [2006], Berger, Sorensen, and Rasmussen [2010]

and Maslowska, Malthouse, and Bernritter [2017]. Our paper provides one possible channel through which this phenomenon may occur in some settings.

In the classic disclosure models it is weakly (Grossman [1981]) or strictly (Jung and Kwon [1988]) optimal to not reveal the lowest signal. The literature has since provided multiple explanations for why this result may not hold in practice. A variety of papers have been written on "the importance of being honest" – situations where an agent has an incentive to disclose unfavorable information (verifiable or not) in order to establish reputation for being trustworthy. See Sobel [1985], Kartik and McAfee [2007], Dziuda [2011], and Beyer and Dye [2012] for some examples. In all of these papers, however, the incentives for honesty are driven purely by the desire to mimic some behavioral type of sender who never lies. Teoh and Hwang [1991], Marinovic and Varas [2016], and Corona and Randhawa [2018] show that bad news may be worth revealing in settings where they may with some probability be discovered by the receiver regardless. Thordal-Le Quement [2014] and Ispano [2018] show that revealing bad news can be used as a signal of the amount of information the sender possesses. Our paper provides a novel motivation for revealing bad reviews in the presence of a mixed audience. On the other hand, Crawford [2003] shows that sender's uncertainty about the receiver's sophistication level can lead to meaningful communication in the cheap talk extension of even a zero-sum game. We show how these forces manifest themselves in the context of disclosure.

Harbaugh and To [2017] generate non-monotone strategies in the context of disclosure. In the equilibria of these models only medium types choose to disclose their information, while neither high nor low types do. The high types prefer to rely on the exogenous signal instead. This phenomenon is generally known as "countersignaling" and can be observed in other contexts as well, e.g., costly signaling (see Feltovich, Harbaugh, and To [2002], Angeletos, Hellwig, and Pavan [2006], and more recently Kurlat and Scheuer [2017]), Bayesian Persuasion (see Chung and Eső [2013], Example 1 in Inostroza and Pavan [2017], and Guo and Shmaya [2018]), costly disclosure (see Quigley and Walther [2017]), and even cheap talk (see Chen [2009]). Countersignaling has also been empirically observed by Luca and Smith [2015]. Dziuda and Salas [2017] obtain something resemblant of countersignaling in the context of cheap talk with detectable deceit. Countersignaling is similar to our reversal result in that the highest sender types are more likely to send lower signal than lower (medium) types in equilibrium. However, unlike all papers mentioned above, we do not rely on public signals to generate reversal. In addition, we focus on outcomes rather than strategies. Heinsalu [2017] demonstrates signal reversal in the context of dynamic costly signaling (as opposed to disclosure) with no reliance on public signals. In the equilibrium of his model low effort is initially treated as evidence in favor of high type of the sender. The source of this reversal is, however, different from our model.

Some existing papers study asymmetries between good and bad news. Harbaugh et al. [2017] explore a setting in which the sender may distort signals of unknown precision before the receiver (who may be naive about this manipulation) observes them. They discover that when the signal realizations are good on average, the sender prefers all individual signals to be good, while if the average is bad, then the sender wants to introduce more variance and make at least some signals look good. Che and Mierendorff [2016] and Zhong [2017] study the question of whether learning from good or bad news is optimal in learning problems.

2.6 Social learning

One can also look at our model from the consumers' side, asking the question of how social learning proceeds in the presence of censorship. This is related to several other strands of literature. One such strand explores moderated social learning. Kremer, Mansour, and Perry [2014] and Che and Hörner [2018] look at a benevolent planner issuing recommendations to sequentially arriving consumers about which option of unknown quality to choose. Some papers on dynamic persuasion (such as Ely [2017] and Renault, Solan, and Vieille [2017]) study similar problems with more pronounced conflict of interest between the sender and the receivers. All of these papers, however, assume that the sender has commitment power and so may design arbitrary intertemporal disclosure policy which needs not be sequentially rational.

Finally, our paper is resemblant of models of observational learning a la Banerjee [1992], Bikhchandani, Hirshleifer, and Welch [1992], and Smith and Sørensen [2000]. See Smith and Sørensen [2011] for a review of this literature, Eyster and Rabin [2010] and Eyster and Rabin [2014] for versions with naive consumers, and Liu and Schiraldi [2012] and Nikiforov [2015] for examples of how herds may be manipulated by an outside party. The resemblance is that in all these models sequentially arriving consumers learn from those before them. The difference is that in this literature the consumers receive a private signal that is informative about the alternatives and observe the decisions of all past consumers (but not their private information). In our paper the consumers do not receive any private information, but instead observe the payoffs of some of the previous consumers. This allows for a simpler inference problem and mitigates the issue of cascades (herding on a suboptimal alternative). More generally, there exists wide literature on social learning (e.g., Frick and Ishii [2016], Acemoglu, Makhdoumi, Malekian, and Ozdaglar [2017]) which focuses on the assumptions about players' information structures which enable social learning of an unknown state.

3 Illustrative Example

This section presents an example that demonstrates the simplest version of our result – that revealing bad reviews can be profitable for the seller. Assume there are two periods $t \in \{1, 2\}$. A long-lived seller offers for sale a product of privately known persistent quality $\theta \in \{H, L\}$ that he has in infinite supply. Price is fixed at $c > 0$. Low-quality product always yields utility zero to consumers. High-quality product yields utility 1 with probability q and utility 0 with probability $1 - q$.

In each period one short-lived consumer arrives at the market. The first consumer believes that $\theta = H$ with probability p_1. She purchases the product if and only if $p_1 \cdot q \geqslant c$. Conditional on the purchase, she consumes the product immediately and with exogenous probability $\lambda > q$ leaves a review, meaning that she honestly reveals the utility that she received from consuming the product. The seller then decides whether to remove the review. Let r_1^θ denote the [endogenous] probability with which the seller of type θ discloses a bad review (and it never makes sense to delete a good review).

The second consumer initially has prior p_1 that the product is of high quality. Upon arriving at the market she observes the first consumer's review unless it was removed by the seller. With

probability $\frac{1}{2}$ she is *strategic* and uses Bayesian updating to calculate her belief p_2, so that conditional on no review it is such that

$$\frac{p_2^s(\emptyset)}{1-p_2^s(\emptyset)} = \frac{p_1}{1-p_1} \cdot \frac{(1-\lambda)+\lambda(1-q)\left(1-r_1^H\right)}{(1-\lambda)+\lambda\left(1-r_1^L\right)},$$

and conditional on a bad review it is given by

$$\frac{p_2^s(B)}{1-p_2^s(B)} = \frac{p_1}{1-p_1} \cdot \frac{\lambda(1-q)\, r_1^H}{\lambda r_1^L}.$$

Conditional on a good review, we trivially have that $p_2^s(G) = 1$.

On the other hand, with probability $\frac{1}{2}$ the second consumer is *naive* and ignores the possibility that the seller may have deleted the unfavorable review. Then her belief in the case of a bad review is given by

$$\frac{p_2^n(B)}{1-p_2^n(B)} = \frac{p_1}{1-p_1} \cdot \frac{\lambda(1-q)}{\lambda}.$$

In the case of no review we have $p_2^n(\emptyset) = p_1$.

Irrespective of type, the second consumer purchases the product if and only if $p_2 \cdot q \geqslant c$. Suppose then that the parameters are such that

$$p_1 > \frac{c}{q} > p_2^n(B),$$

so that naive consumer in period 2 does not purchase the product after a bad review (but does otherwise). Then it only makes sense for the seller to reveal a bad review in period 1 if losing a sale to a potential naive consumer is offset by generating a sale to a potential strategic consumer who would not have bought the product otherwise, i.e., we should have that $p_2^s(B) \geqslant \frac{c}{q} > p_2^s(\emptyset)$. In other words, the reaction of a strategic consumer to no review must be worse than that to a bad review. If this is the case, then both types of the seller are indifferent between revealing a bad review and deleting it, so any strategy profile (r_1^H, r_1^L) that generates such a belief response would constitute an equilibrium. In particular, $(r_1^H, r_1^L) = (1, 0)$ is a valid equilibrium strategy profile in this case – we have assumed $\lambda > q$, so $p_2^s(\emptyset) < \frac{c}{q}$, while $p_2^s(B) = 1$.

This example demonstrates the "direct effect" of revealing bad reviews: doing so makes the seller lose sales to naive consumers, hence should increase sales to strategic consumers. The "expectancy effect" mentioned in the introduction does not manifest here because it requires richer dynamics. Finally, in this example both types of the seller are indifferent between revealing a bad review at $t = 1$ and not. This indifference is driven by the fact that naive and strategic consumers arrive with equal probabilities in period 2. In the full model we show that this assumption is by no means necessary, and in fact strict preference to reveal bad reviews can be obtained if (and only if) this assumption is violated.

4 The Model

Time is continuous and infinite, $t \in [0, +\infty)$. A long-lived seller offers for sale a product of privately known persistent quality $\theta \in \{H, L\}$, high or low, that he has in infinite supply. Quality θ

is hereinafter referred to as the seller's type. The price of the product is fixed at $c > 0$.[7] Short-lived consumers with a unit demand arrive at the market according to a Poisson process with intensity λ. In other words, the probability that a consumer arrives in any given time interval $[t, t + dt)$ is equal to $\lambda \cdot dt$. All players are assumed to be risk-neutral and evaluate outcomes by their expected values.

Once a consumer arrives, she and the seller instantaneously play the following stage game, specific elements of which are described in more detail in the following subsections. The consumer who arrives at the market observes all information available to her and decides whether to buy the product. If she does, she receives random utility depending on product quality. After the utility is realized, the consumer leaves a review describing her experience and then leaves the market forever. The seller then decides whether to reveal the review or delete it. If the review is revealed, it is then observed by all future consumers before they make their purchase decisions.

We let h_t denote a complete history of the game up to (but not including) time t. It includes current time, the purchase decisions of all consumers who arrived before t, all reviews they wrote, and all respective censorship decisions of the seller. The following subsections elaborate on various parts of the game and introduce notation that will be used throughout.

4.1 Consumers

Consumers arrive at the market according to a Poisson process with intensity λ. Each arriving consumer observes the current time and all reviews written by previous consumers that were not deleted by the seller. The consumer does not observe the purchase decisions of the previous consumers and does not observe whether any reviews were deleted.

The consumers' payoffs are as follows. If a consumer leaves the market without buying the product, she receives utility 0. Consuming a high-quality product yields utility 1 with probability q and utility 0 with probability $1 - q$, while a low-quality product always yields utility 0.[8] Without loss we let consumers' utility be linear in money, so buying the product at price c creates a disutility of c utils.

Each arriving consumer has a "cognitive type" $\gamma \in \{s, n\}$, hereinafter referred to as consumer's type. Strategic consumers ($\gamma = s$, share $\mu \in [0, 1)$ of the population) go through full Bayesian reasoning to infer product quality based on published reviews, taking the seller's censorship strategy into account. Naive consumers ($\gamma = n$, share $1 - \mu$) use Bayesian updating for any reviews they observe but are unaware of possible moderation, i.e., they assume that the seller never removes any reviews. For technical reasons, we also assume that naive consumers do not observe or ignore the times at which reviews were written (unlike strategic consumers).[9] Consumers' types are i.i.d. within the sequence of arriving consumers.

Let $p^\gamma(h_t)$ denote the probability that consumer of type γ assigns to the product being of high

[7] Allowing the seller to set the price would allow for price signaling, while in this paper we focus solely on censorship. That said, giving the seller the control over price could be an interesting (and non-trivial) extension.

[8] This renders any good review into conclusive evidence that the product is of high quality. This assumption of "conclusive good news" is relatively standard in the experimentation literature (see Keller, Rady, and Cripps [2005] and the subsequent literature) since it makes the models a lot more tractable. We relax this assumption in Section 6.3 and show that the main result survives under arbitrary information structures.

[9] This assumption can be easily disposed of at the cost of more complicated assumptions about off-equilibrium-path beliefs (described in Section 4.3).

quality given history h_t. She then buys the product at h_t if and only if her expected consumption utility exceeds price c, i.e.,

$$p^\gamma(h_t) \cdot (q \cdot 1 + (1-q) \cdot 0) + (1 - p^\gamma(h_t)) \cdot 0 \geqslant c,$$

or, equivalently, $p^\gamma(h_t) \geqslant \bar{p}$ where $\bar{p} := c/q$. This behavior will be taken for granted for the remainder of the paper. To avoid triviality, the parameters are assumed such that $\bar{p} \in (0,1)$ and consumers' prior is $p^\gamma(h_0) \geqslant \bar{p}$. We further assume that consumers buy the product when indifferent.

In addition, it will prove convenient to have a separate piece of notation for updated beliefs. Let $f^\gamma(h_t)$ denote the belief p^γ of a consumer arriving in the moment $t + dt$ following history h_t and observing that a bad review was posted at t.

After the utility is realized, the consumer leaves a review. The question of why people decide to leave product reviews is interesting in itself, since in general writing a review costs time and effort and yields little benefit, but this question lies beyond the scope of this paper.[10] In our model, writing a review does not in any way affect the consumer's payoff. Therefore, we model this decision as non-strategic, saying that all consumers leave a review and do so truthfully.[11] Since the utility that a consumer may receive from consumption is binary, so are the reviews. A consumer who decided to leave a review and received utility 1 leaves a good review, while one who received utility 0 leaves a bad review.

4.2 The Seller

The seller is long-lived and discounts future at rate $r > 0$. He always observes the complete history h_t of the game so far, as well as his own type θ. We assume that the seller has zero cost of producing the product and thus receives profit c from every purchase. Therefore, instantaneous expected flow profit for type θ seller at history h_t, given that consumers' purchasing decisions are as described above, is equal to $\lambda c \pi(h_t)$ where $\pi(h_t)$ measures expected sales per arriving consumer:

$$\pi(h_t) = (1-\mu) \cdot \mathbb{I}(p^n(h_t) \geqslant \bar{p}) + \mu \cdot \mathbb{I}(p^s(h_t) \geqslant \bar{p}),$$

with $\mathbb{I}(\cdot)$ being an indicator function. Then the seller's discounted future profit (normalized by $\frac{1}{\lambda c}$) is given by

$$V^\theta(h_t) = \frac{1}{\lambda c} \cdot \mathbb{E}\left[\int_t^{+\infty} e^{-r(u-t)} \pi(h_u) du \,|\, h_t, \theta \right], \tag{1}$$

where the expectation is taken over future histories h_u. Note that seller's type θ enters (1) only through this expectation. Also, conditioning on h_t implies that this value function is evaluated *before* the seller knows whether a consumer (and, consequently, a review) arrives at time t.

The seller only has a nontrivial choice of action at those histories at which a new review arrives.

[10] Li et al. [2016] explore a setting in which the seller may choose to reward the consumer for writing a review. While this may serve as one possible explanation, we feel that other, more psychological, channels are at play as well.

[11] Any model that generates truthful reporting would work in supporting our results. For example, one may assume that consumers experience warm glow from providing truthful information to future consumers. Alternatively, the whole situation can be seen as a reciprocity game with the seller in which the consumer wants to reward good products with good reviews and punish bad products using bad reviews.

The seller then has to decide whether to disclose it or not.[12,13] Any good review perfectly reveals the high quality of the product, guaranteeing that all future consumers of any type will buy the product, and is thus never concealed by a seller. Therefore, the seller only faces a nontrivial choice when a bad review arrives. We denote by $r^\theta(h_t)$ the probability with which seller of type θ reveals (or discloses) a bad review that arrives at history h_t.

4.3 Equilibrium Definition

All consumers in our model are short-lived, so their behavior is myopic. The only strategic player is the seller. He maximizes his value (1) given the consumers' behavior, and the latter only depends on their current and future beliefs $p := (p^n, p^s)$. Given that all available information about future beliefs is contained in current beliefs and the seller's strategy, and that the seller observes all the information that the consumers see, current beliefs $p(h_t)$ are a sufficient statistic of history h_t. Therefore, we can essentially without loss of generality focus on Markov setting with state p and redefine all objects accordingly.[14] For example, the seller's strategy in such a setting is described by $r^\theta(p_t) = r^\theta(p(h_t)) := r^\theta(h_t)$.

We are then looking for Markov Perfect Equilibria of the game, which consist of a strategy profile $\left(r^H, r^L\right)$ and updating rules for beliefs p such that

1. at any state p strategy r^θ for the seller of type θ maximizes $V^\theta(p)$ given the updating rules for p_t and the initial condition $p_0 = p$;

2. beliefs p are updated given strategies $\left(r^H, r^L\right)$ in such a way that

   - p^s is updated using Bayes' rule whenever possible;
   - p^n is updated using Bayes' rule whenever possible under the assumption that $r^H(p) = r^L(p) = 1$ for all p;
   - $p^\gamma = 0$ at histories that a consumer of type γ perceives as being off the equilibrium path.

The latter condition about off-path histories is made purely for convenience and is without loss of generality: if there exists an equilibrium with some off-path beliefs, it can as well be sustained by the most pessimistic off-path belief.

[12] We assume that deleting a review is costless. This may contradict some motivating examples in which deleting reviews is costly, as the company/government has to sustain a customer service/censorship apparatus. However, we argue that *if* the company engages in censorship, the marginal cost of deleting another review is essentially zero.

[13] We assume that reviews cannot be held in a "moderation queue" and revealed later, as well as that published reviews cannot be deleted in the future. The former assumption is made mostly for convenience and has no effect on our results, but the latter assumption is crucial. It can be justified by the folk wisdom that "nothing can be deleted from the Internet" (see the "Streisand effect"). We further do not allow the seller to modify review contents (see Section 2.2 for references to papers that do).

[14] Formally, some equilibria are lost as a result of the restriction to Markov strategies. In particular, many states p admit multiple possible continuation equilibria. In this case we lose [Perfect Bayesian] equilibria which prescribe different continuations at different histories h_t which generate the same state p_t. However, this is not a meaningful loss in the sense that any payoff profile attainable in some PBE of our model can also be generated by some MPE (possibly augmented by a public randomization device).

5 Equilibrium Analysis

This section contains the characterization of equilibria of the game, which culminates in the two main results. Formal proofs of all statements can be found in the Appendix. We start, however, with discussing some preliminaries.

5.1 Preliminaries: Multiplicity

Like most communication games, our model suffers from the multiplicity of self-fulfilling equilibria.[15] The loop for any given state p proceeds as follows: if no bad reviews are ever revealed at p, then strategic consumers are allowed to have arbitrary beliefs in case a bad review *is* revealed at p. In particular, consumers can ascribe $f^s(p) = 0$ after such an off-path event, which makes revealing a bad review at p a weakly dominated action for any seller because the naives' demand cannot increase after a bad review, thus closing the loop. Using this reasoning, we can "ban" disclosure of bad reviews on any subset of the state space.[16] We do not refine such situations away, but our main interest lies in the regions where bad reviews are disclosed.

One particular equilibrium deserves special attention:

Definition 1. *An equilibrium is* fully censored *if* $\left(r^H(p), r^L(p)\right) = (0,0)$ *for all* $p \in [0,1]^2$.

In the fully censored equilibrium, all bad reviews are always deleted. This equilibrium is special in the sense that it always exists, as the reasoning above implies. One of the main contributions of our paper is showing that equilibria exist that are not fully censored (or payoff-equivalent to fully censored equilibrium), i.e., bad reviews are revealed in a payoff-relevant way in such equilibria. In other words, censorship is a trivial phenomenon in equilibrium; it is the *lack* of censorship that is not trivial.

To make the classification of equilibria easier, consider the following piece of notation. Given some equilibrium, let $R \subseteq [0,1]^2$ denote the set of states in which bad reviews are revealed on equilibrium path with positive probability: $R := \{p | \left(r^H(p), r^L(p)\right) \neq (0,0)\}$. Then after observing a bad review at some $p \in R$ the strategic consumer updates her belief p^s using Bayes' rule, while after a bad review at $p \notin R$ the refinement we have adopted implies that $f^s(p) = 0$. The fully censored equilibrium is characterized by $R = \emptyset$.

5.2 Preliminaries: Beliefs

This section explores how beliefs $p = (p^n, p^s)$ are updated throughout the game. At any given history, one of three mutually exclusive events can happen: a good review is posted, i.e., written by a consumer and not deleted by the seller, a bad review is posted, or no new reviews are posted. After any single good review the product is revealed to be good, and beliefs of both types of consumers jump to $p^n = p^s = 1$. Conditional on the other two events, the two types of consumers update their beliefs differently.

Recall that $f^\gamma(p)$ denotes the posterior belief of a consumer of type γ who has observed a bad review posted in state p. For a strategic consumer, Bayes' rule prescribes that the belief is updated

[15] "Self-fulfilling equilibrium" is used as a heuristic notion and not in the sense of any formal definition.

[16] More generally, bad reviews can be banned at arbitrary sets of histories h_t of the game, leading to some non-Markov equilibria.

as

$$\frac{f^s(p)}{1-f^s(p)} = \frac{p^s}{1-p^s} \cdot \frac{(1-q)\cdot r^H(p)}{r^L(p)}. \tag{2}$$

A naive consumer uses the same Bayes' rule to update her beliefs but under the assumption that $r^\theta(p) = 1$ for both θ. Therefore, her belief is updated as

$$\frac{f^n(p)}{1-f^n(p)} = \frac{p^n}{1-p^n} \cdot (1-q)\,. \tag{3}$$

Note that the right-hand side does not depend on p^s or equilibrium strategies $r^\theta(p)$, so $f^n(p)$ in any equilibrium is fully described by the p^n coordinate of the current state. The $(1-q)$ term in (2) and (3) is the "inherent meaning" of a bad review – the fact that absent any other information, the belief should decrease. The $\frac{r^H(p)}{r^L(p)}$ ratio in (2) represents the information about quality θ contained in the seller's strategies.

Finally, if no reviews are published during $[t, t+dt)$ then strategic consumers update their beliefs as

$$\frac{p^s_{t+dt}}{1-p^s_{t+dt}} = \frac{p^s_t}{1-p^s_t} \cdot \frac{(1-\lambda\pi(p)dt) + \lambda\pi(p)dt\cdot(1-q)\cdot\left(1-r^H(p_t)\right)}{(1-\lambda\pi(p)dt) + \lambda\pi(p)dt\cdot(1-r^L(p_t))}.$$

By the usual argument, which involves taking logarithms of both sides and using the approximation $\ln(1+x) \approx x$ for small x, we obtain

$$\dot{p^s} = \lambda p^s\left(1-p^s\right)\pi(p)\cdot\left[(1-q)\cdot\left(1-r^H(p)\right) - \left(1-r^L(p)\right)\right].$$

Hereinafter we use a shorthand notation for the drift term $D(p) := (1-q)\cdot\left(1-r^H(p)\right) - \left(1-r^L(p)\right)$, so the expression above can be written as

$$\dot{p^s} = \lambda p^s\left(1-p^s\right)\cdot\pi(p)D(p). \tag{4}$$

For naive consumers, the similar procedure under the assumption $r^H(p) = r^L(p) = 1$ yields $\dot{p^n} = 0$. Since the intensity λ of reviews' arrival in the absence of censorship is the same for high- and low-quality products, the lack of reviews is uninformative for naive consumers, and their belief stays frozen until a new review is published.

5.3 Preliminaries: Bands and Patience

In the analysis it will prove useful to have measures of demand for the two groups of consumers. Demand here is understood not in the sense of "how much a given consumer buys" but rather "how long it takes until type-γ consumers stop buying the product." We introduce these measures of "how long" in different ways for naive and strategic consumers.

Belief p^n of naive consumers in the absence of good reviews is fully determined by the number of posted bad reviews. In particular, it is independent of time, so naive consumers do not change their purchasing behavior as long as no new reviews are posted. We can, however, keep track of how many more bad reviews they are ready to observe in the absence of good reviews before they stop buying the product. We do this by partitioning the state space (p^n, p^s) into "bands" $\{B_k\}_{k\geqslant 0}$, where k corresponds to the number of bad reviews needed to drop the naive consumers' belief p^n below the

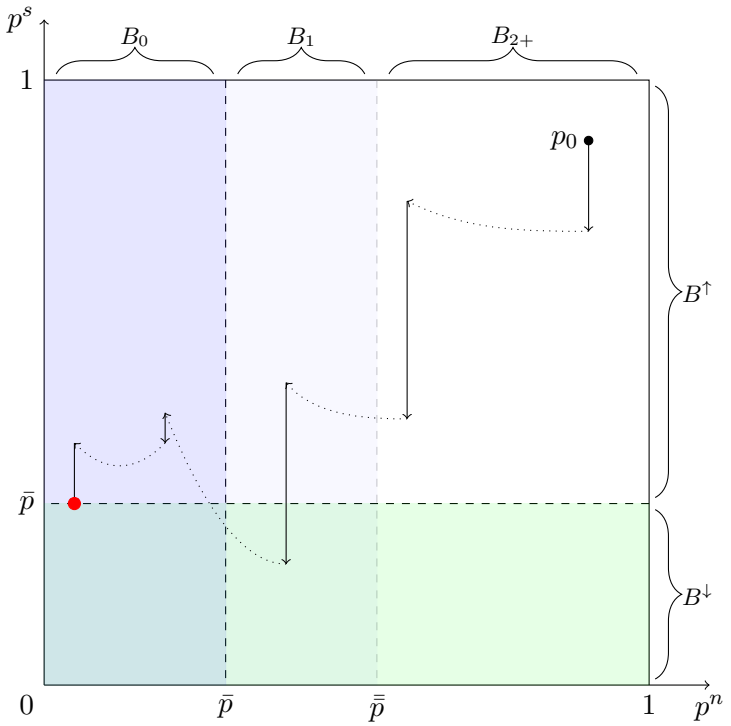

Figure 1: state space, "bands" and an example path of beliefs

threshold $\bar{p}$. In particular, let $B_0 := \{p|p^n \in [0, \bar{p})\}$ and for $k \in \mathbb{N}$ let $B_k := \{p|f(p) \in B_{k-1}\}\backslash B_{k-1}$. Additionally, let $B_{k+} := \cup_{k' \geqslant k} B_{k'}$. We further split each band into two parts according to whether p^s is above or below $\bar{p}$, with $B_k^\uparrow := B_k \cap \{p^s \geqslant \bar{p}\}$ and $B_k^\downarrow := B_k \cap \{p^s < \bar{p}\}$.

Figure 1 illustrates bands together with an example path of beliefs from time zero until a point where all consumers stop buying the product in the absence of a good review. Solid arrows show how beliefs drift over time in the absence of reviews, dotted arrows show how beliefs jump when a bad review is disclosed, and $\bar{\bar{p}}$ is such that $f^n(\bar{\bar{p}}, p^s) = \bar{p}$ for arbitrary p^s.

On the other hand, to measure strategic consumers' demand we use their "patience" $\tau(p)$ – the time it takes strategic consumers to quit the market in the absence of any reviews. Formally, for any k and all $p \in B_k$, let

$$\tau(p) := \inf\{t|p_t \in B_k^\downarrow,\, p_0 = p\},$$

where the evolution of p_t is given by (4).[17] Let $\tau(p) = +\infty$ if p_t never reaches $B_k^\downarrow$ from p.

Since "deleting all future bad reviews" is an available strategy, $\tau(p)$ gives a lower bound on time for which a seller can keep strategic consumers in the market starting at p. As we will see in Section 5.5, the low-type seller can actually do no better than this, meaning that $\tau(p)$ is the exact measure of residual demand from strategic consumers faced by the low-type seller.

Two important properties of $\tau(p)$ are worth noting. Firstly, it follows directly from (4) (see

[17]To clarify, $\tau(p)$ is not the "calendar time" of the game when it reaches $B_k^\downarrow$, but rather the time it takes beliefs to drift from p to $B_k^\downarrow$. If the origin state p was reached at some time t then $B_k^\downarrow$ will be reached at time $t + \tau(p)$ in the absence of posted reviews. This interpretation of time is used throughout the paper, and calendar time is ignored throughout.

Lemma 4 in the Appendix) that for any $p \in B_k^{\uparrow}$ for some k, $\tau(p)$ can be expressed as

$$\tau(p) = -\int_{\bar{p}}^{p^s} \frac{1}{\lambda z\,(1-z)\cdot \pi(p^n, z) D(p^n, z)} dz.$$

Secondly, we claim that $\tau(p)$ is actually finite for interior p^s:

Lemma 1. *In any equilibrium, $\tau(p) < +\infty$ and, thus, $D(p) < 0$ for all p with $p^s < 1$.*

The formal proof is contained in the Appendix, although the intuition behind it is simple. Assume by way of contradiction that there exists p such that $\tau(p) = +\infty$. Then once in state p, the seller is able to retain strategic consumers forever by deleting all future bad reviews. On the other hand, $\tau(p) = +\infty$ implies that there exists some state $\hat{p}$ with $D(\hat{p}) \geqslant -\varepsilon$ and $\tau(\hat{p}) \geqslant \frac{1}{\varepsilon}$ for any small $\varepsilon > 0$. The former property implies that some type of the seller should be willing to reveal a bad review at $\hat{p}$ (since otherwise $D(\hat{p}) = -q$ by (4)), and that $f^s(\hat{p}) < \hat{p}^s$ by the martingale property of beliefs.[18] The latter fact – that $\tau(\hat{p})$ is effectively infinite – implies that revealing a bad review should also retain strategic consumers for an arbitrarily long time (i.e., $\tau(f(\hat{p})) \approx +\infty$), as otherwise neither type of the seller would have any incentives to reveal (since a bad review cannot attract naive consumers). However, then we arrive at a point $f(\hat{p})$ with $f^s(\hat{p}) < \hat{p}^s$ and arbitrarily large $\tau(f(\hat{p}))$, so the same argument can be applied again. By iterating the procedure we are bound to eventually arrive at a state with $p^s < \bar{p}$ where the strategic consumers no longer buy the product. This leads to a contradiction, since in that state $\tau(p) = 0$.

5.4 Main Results

The remainder of Chapter 5 is devoted to characterizing the equilibria of the game. Sections 5.5 to 5.7 provide a detailed characterization, while the current section summarizes the main results and provides a condensed version of the intuition behind them.

The first main result, Theorem 1, states that if strategic consumers are buying the product, then any bad review they observe when ready to buy the product will weakly improve their belief in product quality. In particular, this implies that if a strategic consumer is willing to buy the product after observing current time, then reading bad reviews cannot change her mind. We dub this result "reversal", since strategic consumers' reaction to bad reviews is reversed from its natural direction – instead of decreasing p^s, any bad review manages to increase it.

Theorem 1. *In any equilibrium of the game: if $p \in R$ and $p^s \geqslant \bar{p}$ then $f^s(p) \geqslant p^s$.*

Condition $p^s \geqslant \bar{p}$ ensures that strategic consumers' opinion actually matters. This is a sufficient condition for reversal in our model but not a necessary one (as we see below, $f^s(p) > p^s$ for $p \in B_1^{\downarrow} \cap R$, even though by definition $p^s < \bar{p}$ for all $p \in B_1^{\downarrow}$).

While Theorem 1 may seem trivial at first – "if a seller reveals bad reviews then it must be profitable for him to do so" – the devil, as per the tradition, is in the details. In particular, it is not obvious that "more profitable" corresponds to $f^s(p) > p^s$, since the latter property does not

[18] Given that p^s strictly increases after a good review, it has to go down either after a bad review, or in the absence of reviews. We show that $D(p) = -q$ is required for $f^s(p) = p^s$ (see Lemma 3 in the Appendix), while if $D(p) = -\varepsilon > -q$, then it should be that $f^s(p) > p^s$.

guarantee $\tau(f(p)) > \tau(p)$. Before discussing the intuition behind Theorem 1 we state one of its corollaries.

Corollary 1 is an addendum to the main result. It states that in any equilibrium, the high-type seller is less likely to conceal any bad review than the low-type seller.

Corollary 1. *In any equilibrium of the game: if $p \in R$ and $p^s \geqslant \bar{p}$ then $r^H(p) > r^L(p)$.*

This corollary is an immediate consequence of equation (2) and Theorem 1. While formally trivial, this corollary is valuable in that it describes the mechanism at work in Theorem 1: reversal is achieved via the low-type seller deleting more bad reviews than the high type. The fact that a bad review was *not* deleted is then a strong signal of high quality, which in the end outweighs the inherently negative information contained in the review.

The reasoning behind Theorem 1 proceeds in two steps. First we show that in any equilibrium the low-type seller must be indifferent between revealing and deleting a bad review at any $p \in R$ (see Lemma 2 below). Then we show that any strategy profile that satisfies this indifference also necessarily satisfies the first statement of the Theorem.

Arguably the more interesting part of the proof is the second step: from indifference to $f^s(p) \geqslant p^s$. This result comes through two main channels: the direct effect and the expectancy effect. The *direct effect* states that if $p \in B_1$, i.e., $p^n \geqslant \bar{p} > f^n(p)$ – naive consumers are close to quitting the market and one more bad review drives them out, – then the seller's decision to disclose a bad review should be rewarded by higher demand from strategic consumers. This comes from equilibrium reasoning: if a bad review is disclosed then it is beneficial to do so for some type of the seller, meaning that if the seller loses naive consumers, demand from strategic consumers should increase. This higher demand requirement then translates to an increase in reputation requirement.

The *expectancy effect* is more subtle and can be seen as ripples on the water, propagating the original effect away from B_1 into the $B_{2+}^{\uparrow}$ region. By the martingale property of beliefs, values of $f^s(p) - p^s$ and $D(p)$ are negatively associated for any given p.[19] Therefore, the situation in B_1 creates very high *expectancy* for strategic consumers; either outcome affects their belief significantly. Any bad review that is revealed improves it by a lot, but in the absence of reviews this belief deteriorates rapidly. In particular, high expectancy makes strategic consumers impatient: for a given p, more expectancy in the near future leads to lower patience $\tau(p)$, which is disliked by the seller. Therefore, in order to incentivize the seller to reveal bad reviews in $B_2^{\uparrow}$ – and expose himself to this state of high expectancy, – the seller should be rewarded with a reputation premium for doing so. This premium, in turn, increases expectancy above baseline in $B_2^{\uparrow}$, and the whole reasoning unravels to bands $B_k^{\uparrow}$ with $k > 2$. Noteworthy is the fact that if expectancy in $B_1^{\uparrow}$ is high enough to start this chain reaction, then *strictly* positive reputation premia are required in $B_{2+}^{\uparrow}$ to incentivize the seller to reveal a bad review – even though this does not lead to immediate loss of naive consumers' demand.

Our second main result, Theorem 2, demonstrates existence of equilibria in which bad reviews are revealed in a payoff-relevant way. It claims that there exist equilibria that

1. are non-trivially different from the fully censored equilibrium in terms of payoff, and
2. admit strict reversal: $f^s(p) > p^s$ for all $p \in R$.

[19] We use "negatively associated" as an informal term; its exact meaning is given by Lemma 3 in the Appendix.

We construct one family of equilibria that exhibit both features, but one can easily construct an equilibrium that has one of the above features and not the other. Hereinafter "existence" will refer to the existence of equilibria that satisfy both properties above.

Theorem 2 (Existence)**.** *In the set of all equilibria for any given parameter values*

1. *if $\mu \in [0, \frac{1}{2}]$, then all equilibria are payoff-equivalent to the fully censored equilibrium;*
2. *if $\mu \in (\frac{1}{2}, 1)$, then there exist equilibria with $R \neq \emptyset$, which have $f^s(p) > p^s$ for all $p \in R$, and which are not payoff-equivalent to the fully censored equilibrium.*

The first statement of Theorem 2 follows from Proposition 2 below, which implies that if $\mu < \frac{1}{2}$, then $R \cap B_1 = \emptyset$. Therefore, naive consumers' demand cannot be affected by any sequence of bad reviews in equilibrium – bad reviews can only be revealed in B_0 and B_{2+}, where p^n and $f^n(p)$ are always on the same side of $\bar{p}$. In other words, bad reviews can never work as a costly signal because they are never actually costly in terms of driving naive consumers out of the market. Strategic consumers then ignore bad reviews altogether. In the end, while some bad reviews may be revealed in equilibrium, they do not have any payoff-relevant effects.

The second statement of Theorem 2 is ex ante not trivial. Basic models of disclosure (such as Grossman [1981] and Jung and Kwon [1988]) predict that revealing bad news to a strategic audience is always suboptimal. It is straightforward that revealing bad news to a purely naive audience is also suboptimal. However, Theorem 1 shows that the presence of naive consumers in the market affects strategic consumers' reaction to bad news, rendering it positive. The main message of Theorem 2 is that this reversal is enough to warrant the revelation of bad news. Furthermore, it is easy to think of a setting in which bad news *are* revealed but in an irrelevant way. For example, in the model of Grossman [1981], the lowest type is indifferent between revealing his information and not. The value of Theorem 2, and its second part in particular, is saying that bad news can be revealed in a payoff-relevant way. Finally, recall that we only allow $\mu \in [0, 1)$ in the model. It is argued in Section 5.5.1 that in case $\mu = 1$, all equilibria are payoff-equivalent to the fully censored equilibrium. Therefore, naive consumers are necessary for bad reviews to be revealed in a payoff-relevant way, but they should not be the dominant group in the market.

The main idea behind existence is as follows. Reversal – which is satisfied by all equilibria as per Theorem 1, and hence is a necessary condition for equilibrium – requires that the high-type seller reveals more bad reviews than the low type (Corollary 1). Therefore, the high type must be weakly more willing to reveal bad reviews at all $p \in R$. We show that this condition can be satisfied in $B_1^{\downarrow}$ (i.e., there are equilibria with $B_1^{\downarrow} \cap R \neq \emptyset$) whenever $\mu > \frac{1}{2}$ because the high type faces higher rate of arrival of good reviews and is thus less afraid to lose naive consumers than the low type. This preference then extends to $B_1^{\uparrow}$ to at least some extent, and propagates to B_{2+} as well.

The remainder of Chapter 5 characterizes the game's equilibria in greater detail. A reader who is not interested in these details is invited to skip to Chapter 6. Section 5.5 argues that the low-type seller has to always be indifferent between revealing and deleting bad reviews, and demonstrates the implications that this indifference has for equilibrium strategy profiles and belief dynamics. Theorem 1 relies on Section 5.5 only. Section 5.6 explores the incentives of the high-type seller conditional on low type's indifference. Section 5.7 describes an example of the equilibrium that satisfies the conditions of Theorem 2.

5.5 Characterization: Low Type's Preferences

The first big step in understanding the equilibria of the game relates to incentives of the low-type seller. In particular, we show that for any bad review that can be revealed in equilibrium, the low-type seller must be indifferent between revealing this bad review and deleting it.

Lemma 2. *In any equilibrium, all $r^L(p) \in [0,1]$ are optimal at all $p \in R$. Consequently, deleting all future bad reviews is an optimal continuation strategy for the low-type seller at all p. Furthermore:*

1. $\tau(p) = \tau(f(p))$ *for all* $p \in B_0^{\uparrow} \cap R$,
2. $e^{-r\tau(p)} = \frac{1-\mu}{\mu} + e^{-r\tau(f(p))}$ *for all* $p \in B_1^{\uparrow} \cap R$,
3. $\tau(p) = \tau(f(p))$ *for all* $p \in B_{2+}^{\uparrow} \cap R$.

The intuition behind this result is best understood from reasoning by contradiction. Fix arbitrary $p \in R$. If out of the two actions (deleting a bad review and not) at p the low type only ever does one but not the other, then "the other" becomes a strong positive signal for strategic consumers – so strong that the low-type seller should always find it strictly optimal to pick the "other" action. The details of the argument differ for the two actions, but the essence boils down to the reasoning above.

The behavioral strategy $r^L(p) = 0$ (deleting a bad review at p for sure) is weakly optimal for low-type seller at any $p \in R$ by Lemma 2 and strictly optimal at any $p \notin R$ due to the assumption that $f^s(p) = 0$ for $p \notin R$. Therefore, deleting all bad reviews is trivially a weakly optimal continuation strategy. In this case $V^L(p)$ equals the discounted profit from deleting all future bad reviews:

$$V^L(p) = \frac{1}{r}\left[(1-\mu)\cdot \mathbb{I}(p^n \geqslant \bar{p}) + \mu \cdot \left(1 - e^{-r\tau(p)}\right)\right]. \tag{5}$$

In particular, notice that $V^L(p)$ only depends on $\tau(p)$ and the indicator $\mathbb{I}(p^n \geqslant \bar{p})$.

Finally, given the optimality of deleting all future bad reviews, the equalities in Lemma 2 follow directly by ensuring that $V^L(p) = V^L(f(p))$ for all $p \in R$. The patience of strategic consumers should increase in such a way as to exactly compensate for the loss of naive consumers from revealing a given bad review. In particular, it should be unchanged if the purchasing behavior of naive consumers is unaffected by the revelation.

We now move on to exploring the implications of Lemma 2 for belief dynamics. We essentially unravel the game by backward induction on the state space, analyzing different regions separately.

5.5.1 Band B_0

In $B_0 = \{p | p^n \in [0, \bar{p})\}$, naive consumers are too pessimistic about the product quality to make a purchase. If in addition $p^s < \bar{p}$ (i.e., $p \in B_0^{\downarrow}$), then the same applies to strategic consumers, and the market collapses – no purchases are made and no reviews are written. Region $B_0^{\downarrow}$ is thus an absorbing state and serves as a starting point for the "unraveling" of the state space.

If $p^s \geqslant \bar{p}$ (that is, $p \in B_0^{\uparrow}$), then only strategic consumers buy the product. Since p^n is frozen absent any reviews, only two escapes are possible from $B_0^{\uparrow}$: either a good review is posted and consumers' beliefs jump to $p^n = p^s = 1$, in which case the seller's strategy becomes irrelevant and all consumers stay in the market forever, or the strategic consumers become too pessimistic and

the game arrives at the region $B_0^{\downarrow}$ described above (from Lemma 1, we know this happens in finite time). This structure allows us to characterize continuation equilibria of the game starting from any state $p \in B_0^{\uparrow}$.

Proposition 1 states that disclosure of a bad review should not affect the belief of strategic consumers in $B_0^{\uparrow}$. Whenever naive consumers have quit the market, the seller can no longer signal his credibility to the strategic consumers by sacrificing naive consumers' demand.

Proposition 1. *Strategy profile* $\left(r^H(p), r^L(p)\right)$ *constitutes an equilibrium if and only if* $f^s(p) = p^s$ *for all* $p \in B_0^{\uparrow} \cap R$.

From Lemma 2 we already know that since revealing a bad review at $p \in B_0^{\uparrow} \cap R$ does not affect the purchasing behavior of naive consumers, it should also have no effect on strategic consumers' patience: $\tau(p) = \tau(f(p))$. It is relatively straightforward that an equilibrium with $f^s(p) = p^s$ for all $p \in B_0^{\uparrow} \cap R$ would satisfy this requirement. The value of Proposition 1 hence lies in showing that the converse is also true: strategic consumers' belief *has* to stay unaffected in order to warrant $\tau(p) = \tau(f(p))$, since in any other case the equilibrium cannot be sustained. This also implies that all equilibria are payoff-equivalent in B_0.

Corollary 2. *All continuation equilibria starting from any given* $p \in B_0$ *are payoff-equivalent to the fully censored continuation equilibrium.*

This Corollary follows from the fact that at any given p, drift speed $D(p)$ is the same whether $p \notin R$ or $p \in R$ and $f^s(p) = p^s$ – if bad reviews are irrelevant for strategic consumers then it does not matter whether they are revealed or not.

It is worth noting that Proposition 1 and Corollary 2 apply to the whole state space (i.e., all histories) in case $\mu = 1$ when all consumers are strategic. While this case is purposefully not included in the model setup (all following results only apply to $\mu \in [0, 1)$), this is mostly for sake of narrative clarity. The fact that Proposition 1 applies globally when $\mu = 1$ means that in the absence of naive consumers, bad reviews are completely irrelevant under censorship. Even though some bad reviews may be revealed in that case (since $f^s(p) = p^s$ requires $r^L(p) < 1$ for any $p \in R$ as per (2)), those that *are* revealed carry no useful information whatsoever and have no effect on [strategic] consumers' behavior. Furthermore, revealing bad reviews in that setting has no value to either the seller or the consumers, so all equilibria are payoff-equivalent to the fully censored equilibrium.[20] Therefore, all following results rely on nonzero market presence of naive consumers.

5.5.2 Band B_1

Continuing to band $B_1 = \{p | p^n \in [\bar{p}, \bar{\bar{p}})\}$, one should notice that starting from B_1 the beliefs may only jump to $p^n = p^s = 1$ after a good review, band B_0 after a bad review, or stay in B_1 forever absent any reviews. Therefore, having full knowledge of continuation equilibria in B_0, we can describe the continuation equilibria that start in B_1. The parts of most interest are given by the two following propositions. Proposition 2 states that bad reviews are only revealed in B_1 if there are sufficiently many strategic consumers in the market ($\mu \geqslant \frac{1}{2}$) and they are sufficiently close to quitting the market (p^s is low enough). Proposition 3 then concludes that whenever a bad review *is* disclosed in B_1, it strictly improves strategic consumers' belief p^s.

[20] The key to observing this equivalence is noting that $D(p) = -q$ both if $p \notin R$ and if $p \in R$ and $f^s(p) = p^s$.

Proposition 2. *For any $p \in B_1$: $p \in R$ only if $\mu \geqslant \frac{1}{2}$ and $p^s < p^*$, where*

$$p^* = \frac{\bar{p}\mu^{\frac{\lambda}{r}}}{\bar{p}\mu^{\frac{\lambda}{r}} + (1-\bar{p})(1-\mu)^{\frac{\lambda}{r}}}.$$

Proposition 3. *In any equilibrium of the game $f^s(p) > p^s$ for all $p \in B_1 \cap R$.*

By Lemma 2 the low-type seller should be indifferent between revealing and deleting a bad review at any $p \in B_1 \cap R$. He can retain naive consumers in the market forever starting from any p with $p^n \geqslant \bar{p}$ and can never bring them back to the market starting from any p with $p^n < \bar{p}$. Revealing a bad review at some $p \in B_1$ and losing naive consumers forever is then always worse than deleting it and retaining naive consumers forever – even if revealing the review brings strategic consumers to the market and retains them forever. Therefore, any reason to allow bad reviews in B_1 only arises if $\mu \geqslant \frac{1}{2}$, i.e., if strategic consumers are more prevalent in the market than naive consumers. This gives the first condition in Proposition 2.

The second condition in Proposition 2 comes from the fact that at some points $p \in B_1^{\uparrow}$ patience $\tau(p)$ is so large that even a jump to the most optimistic belief $f^s(p) = 1$ – which grants the seller sales to strategic consumers from now until eternity – is not sufficient to compensate the seller for the loss of naive consumers. This leads to an upper bound on p^s at which bad reviews may be disclosed.

The fact that any revealed bad review trades off naive consumers' demand for that of strategic consumers is the basic idea behind Proposition 3: $\tau(p)$ should increase following disclosure, which implies that p^s should increase as well. This implication is not as trivial as may seem at first because the speeds at which the belief of strategic consumers drifts toward $\bar{p}$ in B_0 and B_1 are not the same in general.

5.5.3 Band B_{2+}

Analogous to before, from B_2 the beliefs may only escape to $p^n = p^s = 1$ after a good review, into B_1 after a bad review, or else stay in B_2 if no reviews are posted. Therefore, we can apply our knowledge of continuation equilibria in B_1 to explore the continuation equilibria originating in B_2 and then unravel to include B_k with $k > 2$. This will conclude our analysis of the implications of the low type's incentives, since the set $\cup_{k \geqslant 0} B_k$ covers the whole state space, so any equilibrium of the game is a continuation equilibrium starting in one of these bands.

Proposition 4. *In any equilibrium of the game $f(p) \geqslant p^s$ for all $p \in B_{2+}^{\uparrow} \cap R$.*

While Proposition 4 may look very similar to Proposition 3, the reasoning behind it is more subtle. Recall from Lemma 2 that the low-type seller must be indifferent between disclosing a bad review and deleting it in all $p \in B_{2+} \cap R$, so it must be that $\tau(p) = \tau(f(p))$. Consider a pair of points $\tilde{p} \in B_2^{\uparrow}$ and $f(\tilde{p}) \in B_1^{\uparrow}$ and assume for purposes of conveying intuition that all states p that follow states $\tilde{p}$ and $f(\tilde{p})$ (in the absence of reviews, i.e., in the sense of equation (4)) belong to R. By the martingale property of beliefs, values of $f^s(p) - p^s$ and $D(p)$ are negatively associated. We use the informal term "expectancy" to denote the common factor underlying both values: high

expectancy is associated with high drift speed $|D(p)|$ (i.e., very negative $D(p)$) and strong reversal $f^s(p) - p^s$, and vice versa.[21]

We know from Proposition 3 that $f^s(p) - p^s > 0$ for all $p \in B_1 \cap R$, so expectancy is high and hence $D(p)$ is strongly negative in B_1. This means that strategic consumers are relatively impatient – $\tau(f(\tilde{p}))$ is "small." Then suppose by contradiction that $f^s(p) < p^s$ for all $p \in B_2 \cap R$. This would imply that expectancy is weak in B_2. Consequently, drift speed $|D(p)|$ is low, so strategic consumers are patient. Together with $f^s(\tilde{p}) < \tilde{p}^s$, this leads to $\tau(f(\tilde{p})) < \tau(\tilde{p})$ – a contradiction of indifference. The formal proof shows that in order to restore the indifference it must be the case that $f^s(p) \geqslant p^s$ for all $p \in B_2^{\uparrow} \cap R$ (i.e., it cannot even be that $f^s(p) - p^s$ is positive for some p and negative for other). Iterating the same argument over bands, we then get that $f^s(p) \geqslant p^s$ for all $p \in B_{2+}^{\uparrow} \cap R$.

Unlike in B_1, we cannot make any hard statements about the area below the cutoff, $B_{2+}^{\downarrow}$. The reasoning for $p \in B_{2+}^{\uparrow} \cap R$ relies on the fact that the cutoff $\bar{p}$ is always reached in finite time from any $p^s < 1$. Under the cutoff there is no such terminal point. All states under the cutoff are inherently similar – they all can warrant the status quo forever, where the strategic consumers are out of the market until a good review arrives and the naive consumers buy the product forever. Therefore, as long as the seller is guaranteed to arrive at a state with $\tau(p) = 0$ (i.e., $p^s < \bar{p}$), he does not care about the exact $f^s(p)$. This means that there is no problem in having arbitrary jumps of p^s. On the other hand, this also means that there are no impediments to constructing specific equilibria in which $f^s(p) > p^s$ for all $p \in B_{2+}^{\downarrow} \cap R$.

5.6 Characterization: High Type's Preferences

This section investigates the high-type seller's preferences conditional on the low type's indifference. In doing so we retrace the same path over bands that we followed in the second step of the proof of Theorem 1. In $B_0^{\uparrow}$ Corollary 2 states that all continuation equilibria at any $p \in B_0^{\uparrow}$ are payoff-equivalent: when all bad reviews are completely uninformative, it does not make much difference whether any of them are revealed on an equilibrium path or not.

The key insight into the high type's incentives lies in $B_1^{\downarrow}$. In any state $p \in B_1^{\downarrow} \cap R$, the seller with a bad review in hand faces the following choice. Deleting the review can retain naive consumers in the market forever but cannot attract strategic consumers. Revealing the review, on the other hand, brings strategic consumers to the market for some time $\tau(f(p))$, but drives the naive consumers away. Figure 2 demonstrates this trade-off graphically, showing "expected sales per consumer" as a function of time for the two strategies outlined above. Time zero on the graph corresponds to states p or $f(p)$ respectively. This graph is valid for all $p \in B_1 \cap R$, so one should remember that $\tau(p) = 0$ for all $p \in B_1^{\downarrow}$.

The low-type seller must be indifferent between deleting and revealing a bad review at any $p \in B_1 \cap R$, meaning that expected sales in the absence of future reviews should be equal in the two scenarios.[22] Visually, it means that the areas under the two intertemporal demand curves in Figure 2 should be equal (after discounting future sales appropriately). The graph makes it obvious why $\mu \geqslant \frac{1}{2}$ is necessary to render the low type indifferent in B_1 (and thus generate a strategy profile

[21] Although drift speed $|D(p)|$ and degree of reversal $f^s(p) - p^s$ are not connected one-to-one, what matters for the argument is that $D(p) \geqslant (>) - q$ if and only if $f^s(p) - p^s \leqslant (<) 0$ (see Lemma 3).

[22] It is enough to consider the case when no reviews arrive after p because the low type can never receive good reviews and always weakly prefers to delete bad reviews so deleting all future bad reviews is an optimal strategy.

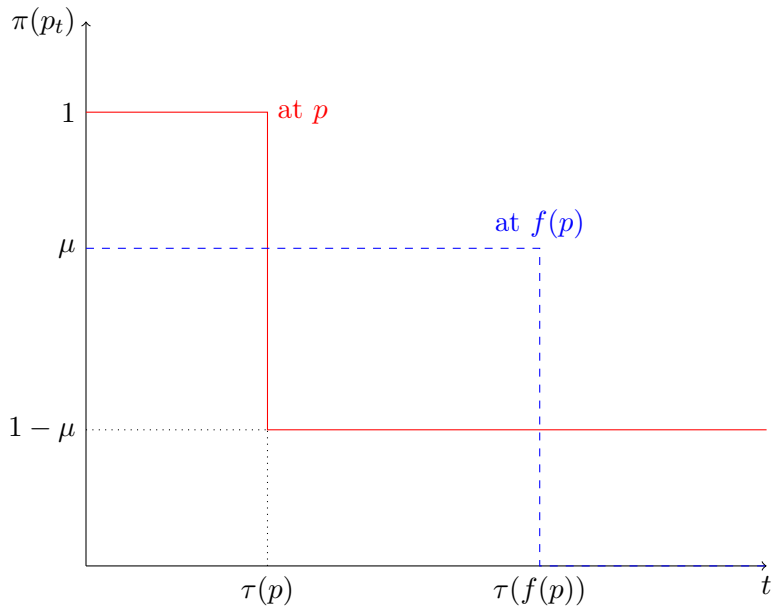

Figure 2: Intertemporal demand starting from some $p \in B_1 \cap R$ and $f(p)$.

with nonempty $B_1 \cap R$) – otherwise deleting the review and staying at p is strictly better.

We next argue that the high type prefers to reveal at $p \in B_1 \cap R$, conditional on the low type's indifference. To see this, note that the only difference between the payoffs of the two types is the option value of receiving a good review for the high-type seller. Because a good review generates the best possible continuation payoff (all consumers buying forever), the high type prefers that this good review arrives sooner rather than later. The rate of arrival of reviews is exactly proportional to sales per consumer. Therefore, conditional on total expected discounted sales being the same in both scenarios (to satisfy the low type's indifference), the high-type seller prefers to frontload sales. In other words, he wants to sell as much product as possible early on in an attempt to generate a good review as early as possible. By looking at Figure 2 it is easy to see that if $\tau(p) = 0$ – which is the case for all $p \in B_1^{\downarrow}$, – then revealing a bad review and jumping to $f(p)$ generates a more frontloaded demand schedule than deleting a bad review and staying at p. The high type prefers to reveal the bad review because it makes the first good review arrive sooner on average.

At the same time, this can be seen as a costly signaling story. The high-type seller strictly prefers to reveal a bad review and lose naive consumers because this is less costly for him than for the low type. In particular, the high type knows that with positive probability he will receive a good review in the future, which will bring naive consumers back to the market. On the other hand, the fact that revealing a bad review is less costly for the high type than for the low type makes this signal credible for strategic consumers, who then react positively to bad reviews.

One can see from Figure 2 that the reasoning for $B_1^{\downarrow}$ presented above also extends by continuity to $p \in B_1^{\uparrow} \cap R$ as long as $\tau(p)$ is low enough. However, it does not need to extend all the way to p^* as given by Proposition 2, meaning that the high type's incentives may provide a tighter bound on p^s for which bad reviews can be revealed in B_1.

Finally, this preference to reveal bad reviews extends from B_1 to B_{2+} (as long as $B_1 \cap R$ is non-trivial). The intuition is as follows. Deleting or revealing a bad review in any $p \in B_{2+} \cap R$ has no immediate effect: it affects neither the naive consumers' decision to purchase the product, nor

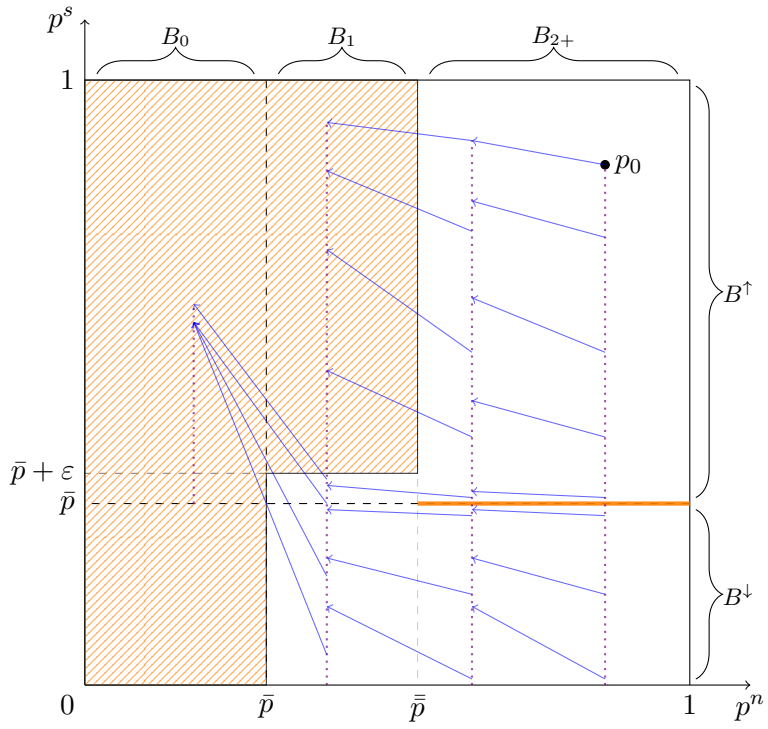

Figure 3: An example of equilibrium with strict reversal.

the strategic consumers' patience $\tau(p)$ (which again follows from Lemma 2). However, revealing a bad review – and sufficiently many bad reviews after it – will bring the high-type seller to some $p' \in B_1$. With positive probability he can then receive another bad review in some $p'' \in B_1 \cap R$ and reveal it, which, as we have established above, is a strictly preferred option. Thus revealing a bad review at $p \in B_{2+} \cap R$ is strictly better than deleting it because it gives the high-type seller a chance to eventually arrive at B_0, which he strictly prefers to staying in B_1 and, by analogy, B_{2+}.

5.7 Existence Example

The argument in 5.6 implies that the strategy profile akin to the one represented in Figure 3 (the formal construction is in the Appendix) could constitute an equilibrium. In particular, this is exactly the equilibrium that is constructed in the proof of Theorem 2. The orange shaded region, which includes B_0, most of $B_1^\uparrow$, and the line $\{p|p^n \geqslant \bar{\bar{p}}, p^s = \bar{p}\}$, is the set $(0,1) \times (0,1) \backslash R$ where no bad reviews are ever revealed. The white region is the revelation set $R = \{B_1|p^s < \bar{p}+\varepsilon\} \cup \{B_{2+}|p^s \neq \bar{p}\}$ for some ε. The argument above has shown that equilibrium conditions (low type's indifference and high type's preference to reveal) can be satisfied for all p within such R.[23] The purple dotted lines show the set of states p that are on equilibrium path given p_0. The blue arrows create a "phase diagram," pointing from p to $f(p)$ for some selected $p \in R$.

We now want to show that the equilibrium with the revelation set R described above can satisfy the two properties we are seeking: it generates payoffs that are different from the fully censored equilibrium and it generates strict reversal in all $p \in R$. We will start with the latter. We have $B_0 \cap R = \emptyset$, and Proposition 3 already gives strict reversal for all $p \in B_1 \cap R$. Therefore, it is

[23] Incentives for $p \notin R$ are trivial due to our assumption that off the equilibrium path $p^\gamma = 0$ for either consumer type γ.

only left to show strict reversal for B_{2+}. In the case of $p \in B_{2+}^{\downarrow}$ any action profile can satisfy the equilibrium conditions as long as $f^s(p) < \bar{p}$, so we can easily construct it in such a way that $f^s(p) > p^s$. Finally, the channel through which reversal works in $B_{2+}^{\uparrow}$ has been described in section 5.5.3. Importantly, that channel relied on high expectancy (i.e., high drift speeds $|D(p)|$) in $B_1^{\uparrow}$, which in turn requires $B_1^{\uparrow} \cap R \neq \emptyset$. Therefore, it is important to include the ε-slice of $B_1^{\uparrow}$ in our revelation set R to generate strict reversal in $B_{2+}^{\uparrow}$.

The other property – payoff-nonequivalence – is easy to deduct given the strict reversal. In fully censored equilibrium $D(p) = -q$ for all p. In the equilibrium we have constructed, $f^s(p) > p^s$ for almost all $p \in B_{2+}^{\uparrow}$, which by Lemma 3 implies that $D(p) < -q$ for all p. Using representation (9) of $\tau(p)$, we immediately obtain that for any given $p \in B_{2+}$ (with $p^s > \bar{p}$), strategic consumers' patience $\tau(p)$ is lower in the equilibrium we have constructed than in the fully censored equilibrium. From Lemma 2 and the consequent optimality of deleting all bad reviews for the low-type seller, one can then conclude that $V^L(p)$ is also lower for such p in the equilibrium of Figure 3 than in the fully censored equilibrium.

6 Discussion and Extensions

This section presents additional observations resulting from our model, which are not directly related to Theorems 1 and 2. We then consider some extensions of the baseline model.

6.1 Seller's Profit

One topic that persists throughout the model is multiplicity of equilibria, which differ in terms of R – the set of states at which bad reviews are revealed on the equilibrium path. It is then natural to ask which types of players prefer which equilibria. Theorem 2 says that in case $\mu \leqslant \frac{1}{2}$ all equilibria are payoff-equivalent. Therefore, from this point onward assume that $\mu \in (\frac{1}{2}, 1)$. Proposition 5 below addresses the question of seller's profit. The next subsection discusses issues related to consumers' welfare.

In general, multiple equilibria may exist with the same revelation set R, and payoff comparison across such equilibria is ambiguous. Therefore, we employ the following equilibrium selection.

Definition 2. *An equilibrium* $(r^L(p), r^H(p))$ *is called* semi-separating *if* $r^H(p) = 1$ *for all* $p \in R$.

This class of equilibria is non-empty, since our constructive proof of the second part of Theorem 2 presents one such equilibrium.[24] As discussed above, Corollary 1 directly implies that at any $p \in R$ we have $r^H(p) > r^L(p) \geqslant 0$, so the high-type seller should at least weakly prefer to disclose a bad review at p. The refinement above then only rules out the case when the high-type seller is exactly indifferent *and* deletes bad reviews with positive probability.

Proposition 5. *Suppose that* $\mu \in (\frac{1}{2}, 1)$ *and consider two semi-separating equilibria with revelation sets* R' *and* $R'' \subset R'$*, respectively. Then the low-type seller weakly prefers equilibrium with* R'' *to equilibrium with* R' *at all* p.

[24] In fact, if there exists some equilibrium with a given R, then there exists a semi-separating equilibrium with that R.

Larger R means more bad reviews are revealed in equilibrium, but it also leads to higher expectancy, making strategic consumers less patient. The latter implies the larger is R, the smaller is $\tau(p)$ for all $p \in B^{\uparrow}$, which in turn makes the low-type seller strictly worse off at those states.

As for the high-type seller, he [weakly] benefits from revealing bad reviews, so he prefers equilibria with larger $B^{\downarrow}_{1+} \cap R$ conditional on $f(p)$ being the same in both equilibria for all $p \in B^{\downarrow}_{1+} \cap R$. However, in $B^{\uparrow}_{1+}$ the two effects described above – less-patient consumers given larger R but more opportunities to reveal a bad review – work in the opposite directions, so the high type's final preferences are ambiguous.

6.2 Asymptotic Learning

The expected utility of a consumer arriving at time t would depend on reviews revealed up to t and is thus tough to measure, even in expectation. However, it is possible to make limit statements. This subsection examines whether the seller's type is learned by the consumers asymptotically as $t \to +\infty$.

Assume that $p_0 > \bar{p}$ to avoid triviality. We say that the seller's type is asymptotically learned by consumers of type γ if the probability that the purchase decision made by consumer of type γ is correct if it approaches 1 as $t \to +\infty$.[25] Asymptotic learning is trivially connected to the welfare of consumers who arrive sufficiently late: if the seller's type is learned, then consumers have full information and thus make efficient purchasing decisions. Not perfectly identifying either type of seller is associated with losses from either buying a low-quality product or not buying a high-quality product.

The bad news for consumers is that learning both sellers' types, by the design of our information structure, is impossible in any equilibrium. To see this, suppose there exists some time t such that absent a good review, both types of consumers stop buying the product by time t with positive probability. With probability $e^{-\lambda q t} > 0$, a high-type seller receives no good review by time t, meaning that with positive probability consumers stop purchasing a high-quality product. On the other hand, if no such time t exists, then consumers never stop buying a low-quality product. Therefore, in any equilibrium at most one type of seller can be identified by all consumers.

In the absence of good reviews, strategic consumers always stop purchasing the product in finite time, so they always reveal a low-type seller. Therefore, they can reveal a high-type seller if and only if naive consumers stay in the market forever so that a good review eventually arrives. In this case the naive consumers also reveal a high-type seller. This happens only if sufficiently many bad reviews are deleted (which is the case in all equilibria if $\mu < \frac{1}{2}$). Conversely, if sufficiently many bad reviews are revealed in equilibrium (R is sufficiently dense), then absent a good review, naive consumers also stop buying the product almost surely as $t \to +\infty$. In this case they reveal a low-type seller, but neither group of consumers is guaranteed to reveal a high-type seller.

6.3 General Information Structures

Following the literature on experimentation, we have adopted the "conclusive good news" structure in our model, so that any good review is a conclusive evidence of $\theta = H$. However, it has

[25] Correct decision is purchasing the product if and only if it is of high quality. Asymptotic mislearning is related to herding on suboptimal alternatives, see Banerjee [1992], Bikhchandani et al. [1992].

been noted that in experimentation models some interesting results disappear with the transition to *"conclusive bad news"* case (see, e.g., Keller and Rady [2015], Halac and Kremer [2017]). In the context of our model this information structure would mean that both types of sellers can generate good reviews but only the low-type seller can receive bad reviews. The following proposition says that in this case no bad reviews are ever revealed.

Proposition 6. *In all equilibria under "conclusive bad news" $R = \emptyset$.*

The intuition behind the proposition is trivial. Under conclusive bad news, any bad review reveals to all future consumers that $\theta = L$, meaning they have no reason to buy the product. Therefore, revealing any bad review is a weakly dominated strategy, and a short proof shows that it is actually strictly dominated.

Of course, possible information structures are not exhausted by the conclusive news cases. Another setting we explore below is one where both good and bad reviews are inconclusive. In particular, consider a *"general"* setting, defined as follows: the low-quality product yields utility 1 with probability q_+^L and utility 0 with probability $q_-^L = 1 - q_+^L$. The respective probabilities for the high-quality product are q_+^H and $q_-^H = 1 - q_+^H$, with $q_+^H > q_+^L$. Let $\bar{p}$ be such that

$$\bar{p}q_+^H + (1 - \bar{p})q_+^L = c.$$

Denote bad and good reviews in this setting as $l \in \{-, +\}$ respectively. Let $r_-^\theta(p)$ and $r_+^\theta(p)$ denote the probability with which seller of type θ reveals a bad review and a good review, respectively, in state p. Let $R_l := \{p|(r_l^H(p), r_l^L(p)) \neq (0, 0)\}$ for $l \in \{-, +\}$. Let $f_l^\gamma(p)$ denote the belief of type-γ consumer who observes a review $l \in \{-, +\}$ posted in state p.

Note that even though this "general" setting is binary, for purposes of our result any setting with more than two reviews/utility levels can be reduced to this general setting by banning (i.e., setting $R_l = \emptyset$ for) all but two reviews. For simplicity we also assume that $q_+^H \cdot q_-^H \geqslant q_+^L \cdot q_-^L$ (so that $f_+^n(f_-(p)) \geqslant p^n$), but this is not a vital assumption.

The following proposition says that in this setting we can still construct an equilibrium in which all revealed bad reviews (and good reviews alike) improve the seller's reputation among strategic consumers – and again bad reviews are revealed in a nonempty set of states in a payoff-relevant way.

Proposition 7. *If $\mu \in \left(\frac{1}{2}, 1\right)$, then there exists an equilibrium in the general setting such that*

1. *$f_l^s(p) > p^s$ for all $p \in R_l$ and all $l \in \{-, +\}$;*
2. *$R_- \neq \emptyset$;*
3. *this equilibrium is payoff-distinct from fully censored equilibrium.*

Fully censored equilibrium in this case can mean either one with $R_- = R_+ = \emptyset$ (i.e., one in which all good and bad reviews are censored), or one with $R_- = \emptyset$ and same R_+ as in the equilibrium under consideration. The latter definition ensures that payoff-nonequivalence is driven by differences in R_- and not R_+.[26]

[26] One can also compare the equilibrium constructed in Proposition 7 to the one with $R_- = \emptyset$ and $R_+ = [0, 1]^2$ with similar results.

The equilibrium constructed in the proof is somewhat more restrictive than that in Theorem 2. In particular, the construction involves $R_- = B^{\downarrow}_{1+}$ and $R_+ = B^{\uparrow}_{-1}$ (where $B_{-1} = \{(p^n, p^s) \in B_0 | (f^n_-)^{-1}(p^n) \geqslant \bar{p}\}$). The important part is that the equilibrium constructed in the proof of Proposition 7 still exhibits relevant economic forces. In particular, now both seller types can bring naive consumers back to the market after driving them out, but this is still cheaper for the high-type seller because he faces a higher rate of arrival of good reviews. Consequently, the high type is more willing to lose naive consumers in the first place, which enables bad reviews' signaling function for strategic consumers.

6.4 Fake Reviews

Suppose that in addition to reviews written by consumers, the seller is able to post fake reviews of his choice. As we show below, our main result (Theorem 2) survives in this case.

Adopt the general setting presented in the previous subsection. Suppose that now the seller also receives opportunities to post any fake review he wants (good or bad) in addition to releasing consumers' real reviews. Future consumers cannot distinguish real reviews and fake reviews. Fake review opportunities arrive with some finite Poisson intensity λ_ϕ, which serves as a proxy for the cost of posting a fake review.[27] This rate λ_ϕ can be arbitrarily high.

For simplicity we impose the same assumptions on fake review opportunities as we do on censorable reviews. Most importantly, opportunities are perishable: given that an opportunity has arrived in some state p, the seller has to decide whether to exercise it immediately, otherwise the opportunity vanishes. The seller also cannot delete fake reviews that he posted in the past.

Let $\phi^\theta_l(p) \in [0,1]$ denote the probability with which seller of type θ fakes review $l \in \{-,+\}$ in state p given that the opportunity. An obvious restriction is $\phi^\theta_-(p) + \phi^\theta_+(p) \leqslant 1$ for any θ, p.

A type-θ seller's strategy in the fake reviews setting is then given by $\{r^\theta_l, \phi^\theta_l\}_{l\in\{-,+\}}$. Rational consumers' beliefs are updated as

$$\frac{f^s_l(p)}{1-f^s_l(p)} = \frac{p^s}{1-p^s} \cdot \frac{\lambda q^H_l r^H(p) + \lambda_\phi \phi^H_l(p)}{\lambda q^L_l r^L(p) + \lambda_\phi \phi^L_l(p)} \tag{6}$$

after review $l \in \{-,+\}$, and as

$$\dot{p^s} = p^s(1-p^s) \cdot \left(\lambda \sum_{l\in\{-,+\}} \left[q^H_l \left(1 - r^H_l(p)\right) - q^L_l \left(1 - r^L_l(p)\right) \right] - \lambda_\phi \sum_{l\in\{-,+\}} \left[\phi^H_l(p) - \phi^L_l(p) \right] \right) \tag{7}$$

in the absence of reviews. We assume that naive consumers ignore the possibility of fake reviews in the same way that they ignore censorship; hence their belief p^n is still frozen in the absence of reviews, and their reaction to review l is given by

$$\frac{f^n_l(p)}{1-f^n_l(p)} = \frac{p^s}{1-p^s} \cdot \frac{q^H_l}{q^L_l}. \tag{8}$$

To show that our main result survives in this setting, we take the equilibrium constructed in the

[27] We interpret the low opportunity arrival rate λ_ϕ as high posting cost, which makes the seller reluctant to post fake reviews very frequently, and vice versa.

proof of Proposition 7 and show that an analogous strategy profile is an equilibrium in fake reviews setting.

Proposition 8. *If $\mu \in \left(\frac{1}{2}, 1\right)$, then there exists an equilibrium in the general setting with fake reviews such that*

1. *$f_l^s(p) > p^s$ for all $p \in R_l$ and all $l \in \{-, +\}$;*
2. *$R_- \neq \emptyset$;*
3. *this equilibrium is payoff-distinct from fully censored equilibrium.*

Fully censored equilibrium here has the same possible meanings as in Proposition 7. The equilibrium constructed in the proof features $\phi_+^\theta(p) = 1$ whenever $p \in R_+$, i.e., both types of seller post fake positive reviews at every opportunity. This dilutes the positive signal contained in good reviews but does not eliminate it completely: $f_+^s(p) > p^s$ but $f_+^s(p) \to p^s$ as $\lambda_\phi \to \infty$.

More interestingly, the equilibrium also features $\phi_-^H(p) = 1$ for all $p \in R_-$: the high type strictly prefers to post fake negative reviews for his own product.[28] This is because, as in the baseline model, the low-type seller is always indifferent between revealing bad reviews and deleting them, while the high type extracts a strictly positive value from signaling through bad reviews (at least in $B_{1+}^{\downarrow}$). The high-type seller writes fake bad reviews only so that he can impress strategic consumers with them. Strategic consumers then do indeed improve their opinion about product quality, even despite (and actually thanks to) the fact that they are fully aware that bad reviews they observe are likely fake and do not stem from any consumer's actual experience.

7 Conclusion

This paper demonstrates that the presence of naive consumers in the market may incentivize the seller to reveal bad reviews even in the presence of an opportunity to costlessly delete them. We show that bad reviews in this setting can be used as a signaling device by the seller with a high-quality product. Revealing bad reviews hurts sales to naive consumers, which he can regain through good reviews more easily than a seller with a low-quality product. This extra information contained in the fact that a bad review was not deleted makes strategic consumers perceive bad reviews more favorably than in the absence of censorship. Furthermore, this between-the-lines information outweighs the inherent negativity of the review, making strategic consumers improve their opinion about the product upon observing a bad review.

Important simplifying assumptions incorporated in the model include the seller's monopoly in the market and his inability to set the price freely, which are in some sense contradictory. Whether the effects demonstrated in this paper survive under competition and/or free pricing of the product is a possible direction for future work.

[28] To clarify, the two features – $\phi_+^\theta(p) = 1$ whenever $p \in R_+$ and $\phi_-^H(p) = 1$ whenever $p \in R_-$ – can coexist in the constructed equilibrium because $R_- \cap R_+ = \emptyset$.

References

D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar. Fast and slow learning from reviews. working paper, 2017.

V. V. Acharya, P. DeMarzo, and I. Kremer. Endogenous information flows and the clustering of announcements. *American Economic Review*, 101(7):2955–2979, December 2011.

K. K. Aköz, C. E. Arbatli, and L. Çelik. Manipulation through biased product reviews. working paper, 2017.

G.-M. Angeletos, C. Hellwig, and A. Pavan. Signaling in a global game: Coordination and policy traps. *Journal of Political Economy*, 114(3):452–484, June 2006.

A. V. Banerjee. A simple model of herd behavior. *Quarterly Journal of Economics*, 107(3):797–817, August 1992.

D. P. Baron. Electoral competition with informed and uninformed voters. *American Political Science Review*, 88(1):33–47, March 1994.

F. Baumann and A. Rasch. Injunctions against false advertising. working paper, 2017.

J. Berger, A. T. Sorensen, and S. J. Rasmussen. Positive effects of negative publicity: When negative reviews increase sales. *Marketing Science*, 29(5):815–827, September-October 2010.

T. Besley and A. Prat. Handcuffs for the grabbing hand? Media capture and government accountability. *American Economic Review*, 96(3):720–736, June 2006.

A. Beyer and R. Dye. Reputation management and the disclosure of earnings forecasts. *Review of Accounting Studies*, 17(4):877–912, December 2012.

S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, October 1992.

A. L. Brown, C. F. Camerer, and D. Lovallo. To review or not to review? Limited strategic thinking at the movie box office. *American Economic Journal: Microeconomics*, 4(2):1–26, May 2012.

Y.-K. Che and J. Hörner. Optimal design for social learning. *Quarterly Journal of Economics*, 2018. forthcoming.

Y.-K. Che and K. Mierendorff. Optimal sequential decision with limited attention. working paper, 2016.

Y. Chen. Communication with two-sided asymmetric information. working paper, 2009.

Y. Chen and D. Yang. The impact of media censorship: Evidence from a field experiment in china. working paper, 2017.

K.-S. Chung and P. Eső. Persuasion and learning by countersignaling. *Economics Letters*, 121(3): 487–491, December 2013.

C. Corona and R. S. Randhawa. The value of confession: Admitting mistakes to build reputation. *Accounting Review*, 93(3):133–161, 2018.

V. P. Crawford. Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *American Economic Review*, 93(1):133–149, 2003.

C. Dellarocas. Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593, October 2006.

M. Deversi, A. Ispano, and P. Schwardmann. Spin doctors: Vague messages in disclosure games. working paper, 2018.

D. Dranove and G. Z. Jin. Quality disclosure and certification: Theory and practice. *Journal of Economic Literature*, 48(4):935–963, December 2010.

R. A. Dye. Disclosure of nonproprietary information. *Journal of Accounting Research*, 23(1):123–145, Spring 1985.

W. Dziuda. Strategic argumentation. *Journal of Economic Theory*, 146(4):1362–1397, July 2011.

W. Dziuda and C. Salas. Communication with detectable deceit. working paper, 2017.

C. Edmond. Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4):1422–1458, October 2013.

G. Egorov, S. Guriev, and K. Sonin. Why resource-poor dictators allow freer media: A theory and evidence from panel data. *American Political Science Review*, 103(4):645–668, November 2009.

J. C. Ely. Beeps. *American Economic Review*, 107(1):31–53, January 2017.

H. Eraslan and S. Ozerturk. Information gatekeeping and media bias. working paper, 2017.

E. Eyster and M. Rabin. Cursed equilibrium. *Econometrica*, 73(5):1623–1672, September 2005.

E. Eyster and M. Rabin. Naive herding in rich-information settings. *American Economic Journal: Microeconomics*, 2(4):221–243, November 2010.

E. Eyster and M. Rabin. Extensive imitation is irrational and harmful. *Quarterly Journal of Economics*, 129(4):1861–1898, November 2014.

N. Feltovich, R. Harbaugh, and T. To. Too cool for school? signalling and countersignalling. *RAND Journal of Economics*, 33(4):630–649, Winter 2002.

M. Frick and Y. Ishii. Innovation adoption by forward-looking social learners. working paper, 2016.

G. Gratton, R. Holden, and A. Kolotilin. When to drop a bombshell. *Review of Economic Studies*, 2018. forthcoming.

G. M. Grossman and E. Helpman. Electoral competition and special interest politics. *Review of Economic Studies*, 63(2):265–286, April 1996.

S. J. Grossman. The informational role of warranties and private disclosure about product quality. *Journal of Law and Economics*, 24(3):461–483, December 1981.

Y. Guo and E. Shmaya. The interval structure of optimal disclosure. working paper, 2018.

I. Guttman. The timing of analysts' earnings forecasts. *Accounting Review*, 85(2):513–545, March 2010.

I. Guttman, I. Kremer, and A. Skrzypacz. Not only what but also when: A theory of dynamic voluntary disclosure. *American Economic Review*, 104(8):2400–2420, August 2014.

M. Halac and I. Kremer. Experimenting with career concerns. working paper, 2017.

R. Harbaugh and T. To. False modesty: When disclosing good news looks bad. working paper, 2017.

R. Harbaugh, J. Maxwell, and K. Shue. Consistent good news and inconsistent bad news. working paper, 2017.

D. Hauser. Promotion, censorship, and reputation for quality. working paper, 2018.

S. Heinsalu. Good signals gone bad: Dynamic signalling with switching efforts. *Journal of Mathematical Economics*, 73:132–141, 2017.

N. Inostroza and A. Pavan. Persuasion in global games with application to stress testing. working paper, 2017.

A. Ispano. Information acquisition and the value of bad news. *Games and Economic Behavior*, 110: 165–173, July 2018.

A. Ispano and P. Schwardmann. Competition over cursed consumers. working paper, 2018.

G. Jin. Competition and disclosure incentives: An empirical study of hmos. *RAND Journal of Economics*, 36(1):93–112, Spring 2005.

G. Jin, M. Luca, and D. Martin. Is no news (perceived as) bad news? an experimental investigation of information disclosure. working paper, 2018.

W.-O. Jung and Y. K. Kwon. Disclosure when the market is unsure of information endowment of managers. *Journal of Accounting Research*, 26(1):146–153, Spring 1988.

N. Kartik and R. P. McAfee. Signaling character in electoral competition. *American Economic Review*, 97(3):852–870, June 2007.

G. Keller and S. Rady. Breakdowns. *Theoretical Economics*, 10(1):175–202, January 2015.

G. Keller, S. Rady, and M. Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, January 2005.

S. Kovbasyuk and G. Spagnolo. Memory and markets. working paper, 2017.

I. Kremer, Y. Mansour, and M. Perry. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, October 2014.

P. Kurlat and F. Scheuer. Signaling to experts. working paper, 2017.

L. Li, S. Tadelis, and X. Zhou. Buying reputation as a signal of quality: Evidence from an online marketplace. working paper, 2016.

X. Li and L. M. Hitt. Self-selection and information role of online product reviews. *Information Systems Research*, 19(4):456–474, December 2008.

T. Liu and P. Schiraldi. New product launch: Herd seeking or herd preventing? *Economic Theory*, 51(3):627–648, November 2012.

M. Luca and J. Smith. Strategic disclosure: The case of business school rankings. *Journal of Economic Behavior & Organization*, 112:17–25, April 2015.

M. Luca and G. Zervas. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427, December 2016.

I. Marinovic and F. Varas. No news is good news: Voluntary disclosure in the face of litigation. *RAND Journal of Economics*, 47(4):822–856, 2016.

E. Maslowska, E. C. Malthouse, and S. F. Bernritter. Too good to be true: the role of online reviews' features in probability to buy. *International Journal of Advertising*, 36(1):142–163, 2017.

D. Mayzlin. Promotional chat on the internet. *Marketing Science*, 25(2):155–163, 2006.

D. Mayzlin, Y. Dover, and J. Chevalier. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–2455, August 2014.

D. Nikiforov. On the belief manipulation and observational learning. Masters Thesis, 2015.

D. Orlov, A. Skrzypacz, and P. Zryumov. Persuading the regulator to wait. working paper, 2016.

L. Pontryagin. *Ordinary Differential Equations*. Addison-Wesley, 1962.

D. Quigley and A. Walther. Inside and outside information. working paper, 2017.

J. Renault, E. Solan, and N. Vieille. Optimal dynamic information provision. *Games and Economic Behavior*, 104:329–349, July 2017.

P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood. The value of reputation on ebay: A controlled experiment. *Experimental Economics*, 9(2):79–101, June 2006.

M. Shadmehr and D. Bernhardt. State censorship. *American Economic Journal: Microeconomics*, 7(2):280–307, May 2015.

J. D. Sheth. *Naivety about hidden information: An experiment*. PhD thesis, 2018. working paper.

L. Smith and P. Sørensen. Pathological outcomes of observational learning. *Econometrica*, 68(2): 371–398, March 2000.

L. Smith and P. Sørensen. Observational learning. *New Palgrave Dictionary of Economics Online Edition*, pages 29–52, 2011.

J. Sobel. A theory of credibility. *Review of Economic Studies*, 52(4):557–573, October 1985.

Y. Sun. A dynamic model of censorship. working paper, 2018.

S. H. Teoh and C. Y. Hwang. Nondisclosure and adverse disclosure as signals of firm value. *Review of Financial Studies*, 4(2):283–313, April 1991.

M. Thordal-Le Quement. Contrarian persuasion. working paper, 2014.

W. Zhong. Optimal dynamic information acquisition. working paper, 2017.

Appendix

All statements below fix some strategy profile $(r^L(p), r^H(p))_{p\in[0,1]^2}$, which in turn produces functions $D(p)$ and $f^s(p)$. Some statements further require this strategy profile to constitute an equilibrium.

Lemmas 1 and 2 are used heavily throughout the Appendix. They are monolithic in essence, but it proved more convenient to stagger their proofs for different bands, since they use different supplementary results.

Lemma 3. *1. $D(p) \in [-1, -q)$ if and only if $f^s(p) > p^s$.*

2. $D(p) = -q$ if and only if either $f^s(p) = p^s$ or $r^H(p) = r^L(p) = 0$.

3. $D(p) \in (-q, 1-q]$ if and only if $f^s(p) < p^s$.

Proof. To show the first claim, observe that $D(p) < -q$ is equivalent to

$$\begin{aligned}(1-q)\cdot(1-r^H(p)) - (1-r^L(p)) &< -q\\ \Leftrightarrow (1-q)\cdot r^H(p) - r^L(p) &> 0\\ \Leftrightarrow \left(\frac{f^s(p)}{1-f^s(p)}\right)\cdot\left(\frac{p^s}{1-p^s}\right)^{-1} \equiv (1-q)\cdot\frac{r^H(p)}{r^L(p)} &> 1.\end{aligned}$$

Two other claims can be obtained by reversing the inequalities or equating both sides. Finally, if $r^H(p) = r^L(p) = 0$, then (4) directly gives that $D(p) = (1-q) - 1 = -q$. □

Lemma 4. *For any $k \geqslant 0$ the following hold:*

1. *For all $p = (p^n, p^s) \in B_k^{\uparrow}$, if there exists $\tilde{p} = (p^n, \tilde{p}^s)$ with $\tilde{p}^s \in [\bar{p}, p^s]$ and $D(\tilde{p}) \geqslant 0$, then $\tau(p) = +\infty$. Otherwise $\tau(p)$ can be represented as*

$$\tau(p) = -\int_{\bar{p}}^{p^s} \frac{1}{\lambda z(1-z)\cdot\pi(p^n, z)D(p^n, z)}dz. \tag{9}$$

2. *For any $p = (p^n, p^s) \in B_k^{\uparrow}$, if $D(p^n, \tilde{p}^s) < 0$ for all $\tilde{p}^s \in [\bar{p}, p^s]$ and $\tau(p) < +\infty$, then $\tau(p^n, \cdot)$ is differentiable in its second argument at p^s.*[29]

[29] At p with $p^s = \bar{p}$ by the derivative of $\tau(p^n, \cdot)$ we understand its right derivative.

3. *If $D(p) \leqslant -\varepsilon < 0$ for all $p \in B_k^{\uparrow}$, then $\tau(p)$ is finite for all $p \in B_k^{\uparrow}$.*

4. *Suppose $g(p) : B_k^{\uparrow} \to [\bar{p}, 1]$ is defined indirectly as $\tau\left(f^n(p^n), g(p)\right) = \psi\left(\tau(p)\right)$ for some differentiable and strictly increasing function ψ, and $\tau(p)$ is finite for any $p \in B_k^{\uparrow} \cup B_{k-1}^{\uparrow}$ with $p^s < 1$, strictly increasing and differentiable in p^s on $[\bar{p}, 1)$.*[30] *Then $g(p)$ is a strictly increasing and differentiable function of p^s. In particular, we have the following representation:*

$$\ln\left(\frac{g(p)}{1-g(p)}\right) = J(p) + \ln\left(\frac{p^s}{1-p^s}\right)$$

where $J(p)$ is a differentiable function of p^s.

Proof. 1. If there exists $\tilde{p} = (p^n, \tilde{p}^s)$ with $\tilde{p}^s \in [\bar{p}, p^s]$ and $D(\tilde{p}) \geqslant 0$, then p_t never reaches $B_k^{\downarrow}$, so $\tau(p) = +\infty$ by definition. Now let p_t^s denote the solution to (4) with the initial condition $p_0^s = p^s$. If $D(p^n, \tilde{p}^s) < 0$ for all $\tilde{p}^s \in [\bar{p}, p^s]$, p_t^s is a strictly decreasing function of t. Therefore, there exists an inverse function $t(p_t^s)$ measuring the time it takes for belief to drift from the initial value p^s to p_t^s. Its derivative is given by

$$\frac{dt(p_t^s)}{dp_t^s} = \left(\lambda p_t^s(1-p_t^s) \cdot \pi(p^n, p_t^s) D(p^n, p_t^s)\right)^{-1},$$

and $t(p^s) = 0$. Therefore, $t(p_t^s) = \int_{p^s}^{p_t^s} \frac{1}{\lambda z(1-z)\cdot\pi(p^n,z)D(p^n,z)} dz$. As $D(\bar{p}) < 0$, the threshold is crossed in zero time. Then substituting $p_t^s = \bar{p}$ we get the result.[31]

2. If $D(p^n, \tilde{p}^s) < 0$ for all $\tilde{p}^s \in [\bar{p}, p^s]$, then representation (9) is valid. Taking the derivative with respect to p^s we get

$$\frac{d\tau(p)}{dp^s} = -\left(\lambda p^s(1-p^s) \cdot \pi(p^n, p^s) D(p^n, p^s)\right)^{-1}. \tag{10}$$

As long as $0 < \bar{p} \leqslant p^s < 1$ and $D(p^n, p^s) < 0$, the derivative is finite and positive.

3. In case $D(p) \leqslant -\varepsilon < 0$ the improper integral in (9) converges for any p and therefore $\tau(p) < +\infty$.

4. Differentiability of $g(p)$ follows directly from the differentiability and monotonicity of a composition and an inverse function. Differentiability of $J(p)$ is then straightforward as $\ln\left(\frac{g(p)}{1-g(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right)$ is a sum of differentiable functions and is therefore differentiable. □

Band B_0

Lemma 5. *1. $D(p) \geqslant -q + \varepsilon$ for some $\varepsilon \in (0, q]$ implies $\ln\left(\frac{f^s(p)}{1-f^s(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right) \leqslant \ln(1-\varepsilon)$.*

2. $D(p) \leqslant -q - \varepsilon$ for some $\varepsilon \in (0, 1-q]$ implies $\ln\left(\frac{f^s(p)}{1-f^s(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right) \geqslant \ln(1+\varepsilon)$.

Proof. We prove only the first claim, the second one is analogous. $D(p) \geqslant -q + \varepsilon$ implies

$$-(1-q) \cdot r^H(p) + r^L(p) \geqslant \varepsilon$$

and further

$$\ln\left(\frac{f^s(p)}{1-f^s(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right) = \ln\left(\frac{(1-q)\cdot r^H(p)}{r^L(p)}\right) \leqslant \ln\left(1 - \frac{\varepsilon}{r^L(p)}\right) \leqslant \ln(1-\varepsilon). \qquad \square$$

[30] If $p \in B_0$, then we let $B_{k-1} = B_k = B_0$.

[31] This proof does not imply that the integral converges. Hence even if $D(p^n, \tilde{p}^s) < 0$ for all $\tilde{p}^s \in [\bar{p}, p^s]$, it may still be that $\tau(p) = +\infty$.

Proof of Lemma 1 for $p \in B_0^{\uparrow}$. Suppose there exists $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_0^{\uparrow}$ with $\tau(\tilde{p}) = +\infty$. Then consider points $p_{inf,1} := (\tilde{p}^n, p^s_{inf,1})$ and $p_{inf,2} := (f^n(\tilde{p}^n), p^s_{inf,2})$, where $p^s_{inf,1} = \inf\{p^s | \tau(\tilde{p}^n, p^s) = +\infty\}$ and $p^s_{inf,2} = \inf\{p^s | \tau(f^n(\tilde{p}^n), p^s) = +\infty\}$.[32] We start by showing that

$$\ln\left(\frac{p^s_{inf,1}}{1-p^s_{inf,1}}\right) - \ln\left(\frac{p^s_{inf,2}}{1-p^s_{inf,2}}\right) \geqslant -\ln\left(1-\frac{q}{2}\right). \tag{11}$$

By Lemma 4 there can be three (mutually non-exclusive) sub-cases to consider.

Case 1: $D(p_{inf,1}) \geqslant 0$. Then $\tau(p_{inf,1}) = +\infty$, and $r^L(p_{inf,1}) > 0$.[33] Therefore, a low-type seller must weakly prefer to disclose a bad review, and thus $\tau(f(p_{inf,1})) = +\infty$. Then $p^s_{inf,2} \leqslant f^s(p_{inf,1})$ by definition of $p^s_{inf,2}$, and $\ln\left(\frac{f^s(p_{inf,1})}{1-f^s(p_{inf,1})}\right) - \ln\left(\frac{p^s_{inf,1}}{1-p^s_{inf,1}}\right) \leqslant \ln(1-q)$ by Lemma 5, which together imply (11).

Case 2: $D(p_{inf,1}) < 0$ and there exists a sequence of points $\{\tilde{p}^s_k\}$ such that $\tilde{p}^s_k \downarrow p^s_{inf,1}$ and $D(\tilde{p}_k) > -\frac{1}{k}$, where $\tilde{p}_k := (\tilde{p}^n, \tilde{p}^s_k)$. Then for any $\varepsilon > 0$ and sufficiently high K we have $D(\tilde{p}_K) > -\frac{q}{4}$, $r^L(\tilde{p}_K) > 0$, and $\ln\left(\frac{\tilde{p}^s_K}{1-\tilde{p}^s_K}\right) - \ln\left(\frac{p^s_{inf,1}}{1-p^s_{inf,1}}\right) < \varepsilon$. As $\tau(\tilde{p}_K) = +\infty$ and $r^L(\tilde{p}_K) > 0$, we must have $\tau(f(\tilde{p}_K)) = +\infty$, and therefore $p_{inf,2} \leqslant f(\tilde{p}_K)$. Finally, by Lemma 5 we then have $\ln\left(\frac{f^s(\tilde{p}_K)}{1-f^s(\tilde{p}_K)}\right) - \ln\left(\frac{\tilde{p}^s_K}{1-\tilde{p}^s_K}\right) \leqslant \ln\left(1-\frac{3q}{4}\right)$. It is then true that

$$\ln\left(\frac{p^s_{inf,1}}{1-p^s_{inf,1}}\right) - \ln\left(\frac{p^s_{inf,2}}{1-p^s_{inf,2}}\right) > \ln\left(\frac{\tilde{p}^s_K}{1-\tilde{p}^s_K}\right) - \varepsilon - \ln\left(\frac{f^s(\tilde{p}_K)}{1-f^s(\tilde{p}_K)}\right) \geqslant -\ln\left(1-\frac{3q}{4}\right) - \varepsilon.$$

The last term is greater than $-\ln\left(1-\frac{q}{2}\right)$ for sufficiently small ε.

Case 3: $D(p_{inf,1}) < 0$ and there exists a sequence of points $\{\tilde{p}^s_k\}$ such that $\tilde{p}^s_k \uparrow p^s_{inf,1}$ and $k < \tau(\tilde{p}_k) < +\infty$, where $\tilde{p}_k := (\tilde{p}^n, \tilde{p}^s_k)$. As $\tau(p_{inf,1}) = +\infty$ in this sub-case and $\tau(\tilde{p}_k) < +\infty$, for any k there exists $\hat{p}_k = (\tilde{p}^n, \hat{p}^s_k)$ with $\hat{p}^s_k \in [\tilde{p}^s_k, p^s_{inf,1}]$ such that $\tau(\hat{p}_k) > k$ and $D(\hat{p}_k) > -\frac{1}{k}$. Note that $\hat{p}_k \to p_{inf,1}$ as $k \to +\infty$. Now suppose (11) does not hold. Fix some arbitrary $\delta > 0$. For any $\delta > 0$ we have $\tau(f^n(\tilde{p}^n), p^s_{inf,2} - \delta)) < +\infty$, so we can find $k > \frac{4}{q}$ such that $\tau(\hat{p}_k) > \tau(f^n(\tilde{p}^n), p^s_{inf,2} - \delta))$. By Lemma 5 we know that $\ln\left(\frac{f^s(\hat{p}_k)}{1-f^s(\hat{p}_k)}\right) - \ln\left(\frac{\hat{p}^s_k}{1-\hat{p}^s_k}\right) \leqslant \ln\left(1-\frac{3q}{4}\right)$. As $r^L(\hat{p}^s_k) > 0$, we must have $\tau(f(\hat{p}_k)) \geqslant \tau(\hat{p}_k)$, and therefore by the monotonicity of $\tau(f^n(\tilde{p}^n), p^s)$ in its second argument we must have $f^s(\hat{p}_k) > p^s_{inf,2} - \delta$. However,

$$\ln\left(\frac{f^s(\hat{p}_k)}{1-f^s(\hat{p}_k)}\right) - \ln\left(\frac{p^s_{inf,2}}{1-p^s_{inf,2}}\right) < \ln\left(1-\frac{3q}{4}\right) - \ln\left(1-\frac{q}{2}\right) < 0,$$

which implies that $f^s(\hat{p}_k) < p^s_{inf,2}$, and by taking sufficiently small δ we achieve a contradiction.

Having shown (11), consider the sequence $\{p_{inf,k}\}$ where $p_{inf,k} := \left((f^n)^{k-1}(\tilde{p}^n), p^s_{inf,k}\right)$ and $p^s_{inf,k} = \inf\{p^s | \tau\left((f^n)^{k-1}(\tilde{p}^n), p^s\right) = +\infty\}$. Equation (11) then implies that $p^s_{inf,k} < \bar{p}$ for all $k > M := \left\lceil \frac{\ln\left(\frac{p_{inf,1}}{1-p_{inf,1}}\right)}{\ln(1-\frac{q}{2})} \right\rceil$, i.e., we have $p_{inf,k} \in B_0^{\downarrow}$, and there exists $\varepsilon_k > 0$ such that $p \in B_0^{\downarrow}$ for all $p = (p^n_{inf,k}, p^s)$ with $p^s \in [p^s_{inf,k}, p^s_{inf,k} + \varepsilon_k)$. However, by definition we have $\tau(p) = 0$ for all $p \in B_0^{\downarrow}$, which brings us to a contradiction with the definition of $p_{inf,M}$. □

Proof of Lemma 2 for $p \in B_0^{\uparrow}$. Proofs for this and other regions proceed by contradiction: we show that the low-type seller can neither have strict preference towards revealing a review ($r^L(p) = 1$), nor towards deleting a review ($r^L(p) = 0$).

Suppose that at some $p \in B_0^{\uparrow} \cap R$ a low-type seller strictly prefers to reveal a bad review, i.e., $r^L(p) = 1$. Then $D(p) \geqslant 0$ and $\tau(p) = +\infty$, which contradicts Lemma 1 for $B_0^{\uparrow}$. If $r^L(p) = 0$ and $r^H(p) > 0$ instead,

[32] As the set is non-empty and bounded from below by $\bar{p}$, the infimum exists.

[33] The latter is true because if $r^L(p_{inf,1}) = 0$, then $D(p_{inf,1}) \leqslant -q$.

then revealing a bad review brings the maximal continuation profit to a low-type seller, while deleting it yields strictly less if no new bad review arrives in time $\tau(p)$, which is finite by Lemma 1 for all $p \in B_0^{\uparrow}$, so the probability of this happening is strictly positive. That contradicts $r^L(p) = 0$. As $r^L(p) < 1$ for all $p \in B_0^{\uparrow}$, we have that a low-type seller weakly prefers to conceal a bad review at every point in $B_0^{\uparrow}$. Therefore, the value of a low-type seller is equal to the value he gets by deleting all further bad reviews: $V^L(p) = \int_0^{\tau(p)} e^{-rt} \cdot \mu dt$. As $V^L(p) = V^L(f(p))$, we must then have $\tau(p) = \tau(f(p))$. □

For further proofs we introduce a new object: the *average drift* at state $p = (p^n, p^s)$ is defined as

$$\bar{D}(p) := \frac{1}{\lambda\pi(p)\tau(p)} \left(\ln\left(\frac{p^s}{1-p^s}\right) - \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)\right).$$

By Lemma 4 $\tau(p)$ is differentiable in p^s, and by Lemma 1 for $B_0^{\uparrow}$, in any equilibrium $\tau(p) < +\infty$ for all $p \in B_0^{\uparrow}$. Therefore, in any equilibrium $\bar{D}(p)$ is well defined in $B_0^{\uparrow}$ and is differentiable with respect to p^s for any $p^s < 1$. Lemma 2 for $B_0^{\uparrow}$ also states that $\tau(p) = \tau(f(p))$, and therefore function $J(p)$ is well-defined for all $p \in B_0^{\uparrow}$ by Lemma 4.

Lemma 6. *1. Suppose there exists a point $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_0^{\uparrow}$ such that $\bar{D}(\tilde{p}) \leqslant -q - \varepsilon$ for some $\varepsilon \in (0, 1-q]$. Then there exists a point $\hat{p} = (\tilde{p}^n, \hat{p}^s)$ with $\hat{p}^s \in [\bar{p}, \tilde{p}^s]$ such that $\bar{D}(\hat{p}) \leqslant -q - \varepsilon$ and $J(\hat{p}) \geqslant \ln(1+\varepsilon)$.*

2. Suppose there exists a point $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_0^{\uparrow}$ such that $\bar{D}(\tilde{p}) \geqslant -q + \varepsilon$ for some $\varepsilon \in (0, q)$. Then there exists a point $\hat{p} = (\tilde{p}^n, \hat{p}^s)$ with $\hat{p}^s \in [\bar{p}, \tilde{p}^s]$ such that $\bar{D}(\hat{p}) \geqslant -q + \varepsilon$ and $J(\hat{p}) \leqslant \ln(1-\varepsilon)$.

Proof. We only show the first statement; the second is proved analogously. Consider a set $S := \{p^s \in [\bar{p}, \tilde{p}^s] | J(\tilde{p}^n, p^s) \geqslant \ln(1+\varepsilon)\}$. First, it is nonempty, as otherwise by Lemma 5 we have $D(p) > -q - \varepsilon$ for all p with $p^s \in [\bar{p}, \tilde{p}^s]$, which violates $\bar{D}(\tilde{p}) \leqslant -q - \varepsilon$.[34] Second, S is closed (as $J(p)$ is continuous in p^s) so its upper contour sets are closed in p^s. Finally, S is trivially bounded from above by $\tilde{p}^s$. Therefore, there exists $\hat{p}^s := \sup S \in S$. Moreover, for all $p^s > \hat{p}^s$ we have $J(\tilde{p}^n, p^s) < \ln(1+\varepsilon)$ and, therefore, $D(\tilde{p}^n, p^s) > -q - \varepsilon$, which implies $\bar{D}(\tilde{p}^n, \hat{p}^s) \leqslant -q - \varepsilon$. The second property of $\hat{p}^s$ follows directly from the definition of S. □

Proof of Proposition 1. First note that any strategy profile that generates $f^s(p) = p^s$ for all $p \in B_0^{\uparrow} \cap R$ constitutes an equilibrium. Indeed, by Lemma 3 $f^s(p) = p^s$ implies $D(p) = -q$ for all p, and therefore $\tau(p) = \tau(f(p))$ for all $p \in B_0^{\uparrow}$, making both types of sellers indifferent between disclosing and concealing a bad review.

Proof of the converse is separated into two steps. First we show that if there exists $p \in B_0^{\uparrow} \cap R$ such that $J(p) \neq 0$, then there exists a point $\tilde{p}$ such that either $\bar{D}(\tilde{p}) \leqslant -q - \varepsilon$ and $J(\tilde{p}) \geqslant \ln(1+\varepsilon)$, or $\bar{D}(\tilde{p}) \geqslant -q + \varepsilon$ and $J(\tilde{p}) \leqslant \ln(1-\varepsilon)$.[35] Then we achieve a contradiction in both of these cases.

Step 1. Suppose there exists $p \in B_0^{\uparrow}$ such that $J(p) \neq 0$. If $\bar{D}(p) \neq -q$, then the claim is valid by Lemma 6. Now suppose that $\bar{D}(p) = -q$. Then as $J(p) \neq 0$, it must be that $\bar{D}(f(p)) \neq -q$ and we can apply Lemma 6 to $f(p)$.

Step 2. Suppose there exists p_1 such that $\bar{D}(p_1) \leqslant -q - \varepsilon$ and $J(p_1) \geqslant \ln(1+\varepsilon)$. Denote $K := 1 + \ln(1+\varepsilon) \cdot \tau(p_1)^{-1}$. As $\tau(f(p_1)) = \tau(p_1)$ and $D(p) \geqslant -1$, it must be that $\bar{D}(f(p_1)) \leqslant K \cdot (-q - \varepsilon)$. Then by Lemma 6 there exists $p_2 = (p_2^s, p_2^n)$ with $p_2^s \in [\bar{p}, f^s(p_1)]$ and $p_2^n := f^n(p_1^n)$ such that $\bar{D}(p_2) \leqslant K(-q-\varepsilon)$

[34] If $D(p) \neq -q$, then $p \in R$ and (2) imply $J(p) = \ln\left(\frac{f^s(p)}{1-f^s(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right)$.

[35] Note that $J(p) = 0$ implies $f^s(p) = p^s$.

and $J(p_2) \geqslant \ln(1 + K(q+\varepsilon) - q) > \ln(1+\varepsilon)$. Iterating this procedure $M := [-\log_K(q+\varepsilon)] + 1$ times we achieve a point p_M such that $\bar{D}(p_M) \leqslant K^M(-q-\varepsilon) < -1$, which is impossible.

Alternatively, suppose there exists p_1 such that $\bar{D}(p_1) \geqslant -q+\varepsilon$ and $J(p_1) \leqslant \ln(1-\varepsilon)$. Then as $\tau(f(p_1)) = \tau(p_1)$ and $J(p_1) < 0$, it must be that $\bar{D}(f(p_1)) > -q+\varepsilon$. Then by Lemma 6 there exists $p_2 = (p_2^s, p_2^n)$ with $p_2^s \in [\bar{p}, f^s(p_1)]$ and $p_2^n := f^n(p_1^n)$ such that $\bar{D}(p_2) \geqslant -q+\varepsilon$ and $J(p_2) \leqslant \ln(1-\varepsilon)$. At the same time, $\ln\left(\frac{p_2^s}{1-p_2^s}\right) - \ln\left(\frac{p_1^s}{1-p_1^s}\right) < \ln(1-\varepsilon)$. Iterating this procedure $M := \left[\left(\ln\left(\frac{\bar{p}}{1-\bar{p}}\right) - \ln\left(\frac{p_1^s}{1-p_1^s}\right)\right) \cdot \frac{1}{\ln(1-\varepsilon)}\right] + 1$ times we achieve a point $p_M = (p_M^s, p_M^n)$ such that $p_M^s < \bar{p}$ and $\tau(p_M) = \tau(p_1)$, – a contradiction. □

Proof of Corollary 2. Proposition 1 and Lemma 3 imply that $D(p) = -q$ for all $p \in B_0$ in any equilibrium. Therefore, (9) states that $\tau(p)$ for any given p must be the same in any equilibrium. Representation (5) then implies that the same is true for $V^L(p)$. The high type's value $V^H(p)$ is also the same in any equilibrium, since it can be written for $p \in B_0$ as

$$\begin{aligned} V^H(p) &= \int_0^{\tau(p)} e^{-rt}\left(\mu + (1-\mu)\cdot(1-e^{-\lambda q \mu t})\right) dt + \int_{\tau(p)}^{+\infty} e^{-rt}\left(1 - e^{-\lambda q \mu \tau(p)}\right) dt \\ &= \frac{\mu\,(r+\lambda q)}{r(r+\lambda q \mu)} \cdot \left(1 - e^{-(r+\lambda q \mu)\tau(p)}\right). \end{aligned}$$

Finally, consumers' behavior and, hence, payoffs are always the same at a given p in any equilibrium. Therefore, for a given $p \in B_0$ all players' payoffs are the same in any equilibrium. □

Band B_1

Lemma 7. *If $\mu < \frac{1}{2}$, then $B_1 \cap R = \emptyset$.*

Proof. If $p \in B_1$ then the low-type seller can guarantee himself

$$V^L(p) \geqslant \frac{1-\mu}{r} + (1 - e^{-r\tau(p)}) \cdot \frac{\mu}{r} > \frac{1-\mu}{r}$$

by deleting all future reviews and retaining naive consumers forever and strategic consumers for time $\tau(p)$. Disclosing any bad review makes naive consumers quit the market until a good review arrives (which is never for a low-type seller), so

$$V^L(f(p)) = (1 - e^{-r\tau(f(p))}) \cdot \frac{\mu}{r} \leqslant \frac{\mu}{r}.$$

As one can see, if $\mu < \frac{1}{2}$, then $V^L(f(p)) < V^L(p)$, hence the low-type seller is never willing to disclose a bad review. □

Proof of Lemma 2 for $p \in B_1$. Whenever $\mu < \frac{1}{2}$, by Lemma 7 we have $B_1 \cap R = \emptyset$ so the statement is trivially true. Thus from now on assume $\mu \geqslant \frac{1}{2}$. We divide the proof into two parts corresponding to two subregions of B_1.

Case 1: $p \in B_1^{\downarrow}$. There it must be that $(1-q)\cdot r^H(p) > r^L(p)$, as by sacrificing the pool of naive consumers any seller must gain the pool of strategic consumers for at least some period of time, so $f^s(p) > \bar{p} > p^s$. In particular, this implies that $r^L(p) = 1$ is not possible in any equilibrium.

As for the second case, suppose instead that $r^L(p) = 0$ and $r^H(p) > 0$. Then any single bad review reveals a high-type seller and trades off the pool of naive consumers for the whole pool of strategic consumers forever. Either group under the respective scenario stays on the marker forever, and the other group joins after a good review. Thus $r^L(p) = 0$ is optimal for the low-type seller only if $\mu = \frac{1}{2}$. In that case the

low-type seller is indifferent between disclosing a bad review and concealing it. If, however, $\mu > \frac{1}{2}$, then the combination of $r^L(p) = 0$ and $r^H(p) > 0$ is impossible, and thus $r^L(p) > 0$.

Case 2: $p \in B_1^{\uparrow}$**.** If $\mu = \frac{1}{2}$, then a strategy profile constitutes an equilibrium in $B_1^{\uparrow}$ if and only if $r^H(p) = r^L(p) = 0$ for points with $p^s > \bar{p}$, and $r^L(p^n, \bar{p}) = 0$. At $\bar{p}$ the low-type seller can then retain one and only one of two types of consumers on the market while another is driven out forever, and he is therefore indifferent between revealing and deleting (but in equilibrium deletes all bad reviews).

Thus for the remainder of the proof we assume that $\mu > \frac{1}{2}$ and consider $p \in B_1^{\uparrow} \cap R$. If $r^L(p) = 1$, then $D(p) \geqslant 0$, so staying silent at p gives the maximum possible continuation payoff to any seller. At the same time, following disclosure any seller loses naive consumers for at least some time and therefore gets strictly less, – a contradiction with $r^L(p) = 1$.

We are left to show that the low-type seller cannot strictly prefer to delete a bad review. Suppose by way of contradiction that there exists $\tilde{p} \in B_1^{\uparrow}$ such that $r^L(\tilde{p}) = 0$ and $r^H(\tilde{p}) > 0$. Then $f^s(\tilde{p}) = 1$. It implies that for all $p \in \{(\tilde{p}^n, p^s) | p^s > \tilde{p}^s\}$ it must be that $r^H(p) = r^L(p) = 0$, as otherwise a seller should receive a strictly higher payoff by disclosing at p than at $\tilde{p}$, which is impossible given $f^n(p) < \bar{p}$. Therefore, if such $\tilde{p}$ exists, then it is unique and $\tau(f(\tilde{p})) = +\infty$. Moreover, non-disclosure is on path in this and all future states, thus deleting this and all future bad reviews must be weakly better for the low-type seller than disclosing a bad review at p and all further bad reviews afterwards:

$$\begin{aligned} \int_0^{+\infty} e^{-rt}(1-\mu)dt + \int_0^{\tau(p)} e^{-rt}\mu dt &\geqslant \int_0^{\tau(f(p))} e^{-rt}\mu dt \\ \Leftrightarrow \frac{1-\mu}{\mu} + e^{-r\tau(f(p))} &\geqslant e^{-r\tau(p)}. \end{aligned} \tag{12}$$

On the other hand, the high-type seller's value from disclosing a bad review at p is

$$\begin{aligned} V^H(f(p)) &= \int_0^{\tau(f(p))} e^{-rt}\left(\mu + (1-\mu)\cdot(1-e^{-\lambda q \mu t})\right)dt + \int_{\tau(f(p))}^{+\infty} e^{-rt}\left(1 - e^{-\lambda q \mu \tau(f(p))}\right)dt \\ &= \frac{\mu(r+\lambda q)}{r(r+\lambda q \mu)} \cdot \left(1 - e^{-(r+\lambda q\mu)\tau(f(p))}\right). \end{aligned} \tag{13}$$

The value that the high-type seller gets from deleting a bad review at p is at least as large as the value from deleting all bad reviews from p onwards:

$$\begin{aligned} V^H(p) &\geqslant \int_0^{\tau(p)} e^{-rt}dt + \int_{\tau(p)}^{+\infty} e^{-rt}\left(1 - \mu e^{-\lambda q(\mu\tau(p) + (1-\mu)t)}\right)dt \\ &= \frac{1}{r} \cdot \left(1 - \frac{r\mu}{r + \lambda q(1-\mu)} e^{-(r+\lambda q)\tau(p)}\right) \\ &\geqslant \frac{1}{r} - \frac{\mu}{r + \lambda q(1-\mu)} \cdot \left(\frac{1-\mu}{\mu}\right)^{1+\frac{\lambda q}{r}}, \end{aligned}$$

where the last inequality follows from (12) after recalling that $\tau(f(p)) = +\infty$. As $r^H(p) > 0$, it must be that

$V^H\left(f(p)\right) \geqslant V^H(p)$, which implies:

$$\frac{\mu\left(r+\lambda q\right)}{r(r+\lambda q\mu)} - \frac{1}{r} + \frac{\mu}{r+\lambda q\left(1-\mu\right)} \cdot \left(\frac{1-\mu}{\mu}\right)^{1+\frac{\lambda q}{r}} \geqslant 0$$

$$\Leftrightarrow \frac{1+\frac{\lambda q}{r}\mu}{1+\frac{\lambda q}{r}(1-\mu)} \cdot \left(\frac{1-\mu}{\mu}\right)^{\frac{\lambda q}{r}} \geqslant 1 \tag{14}$$

Note that (14) holds with equality for $\mu = \frac{1}{2}$ and that $\left(1+\frac{\lambda q}{r}x\right)(1-x)^{\frac{\lambda q}{r}}$ is a decreasing function of x for all $x \in (0,1)$. This means that (14) is violated whenever $\mu > \frac{1}{2}$, so there does not exist any $p \in B_1^{\uparrow} \cap R$ with $r^L(p) = 0$.

Finally, as $r^L(p) < 1$ for all $p \in B_1$ and a low-type seller is indifferent between disclosing and concealing a bad review at all $p \in B_0^{\uparrow} \cup B_1$, (12) holds with equality for all $p \in B_1^{\uparrow}$. □

Proof of Lemma 1 for $p \in B_{1+}^{\uparrow}$. We prove the claim only for $B_1^{\uparrow}$. Induction to all further bands is straightforward. Assume the contrary. Then there exists $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_1^{\uparrow}$ with $\tau(\tilde{p}) = +\infty$. Consider a state $\tilde{p}_{inf} := (\tilde{p}^n, \tilde{p}^s_{inf}) \in B_1^{\uparrow}$ where $\tilde{p}^s_{inf} = \inf\{p^s | \tau(\tilde{p}^n, p^s) = +\infty\}$. According to Lemma 4, there can be three sub-cases. Either $D(\tilde{p}_{inf}) \geqslant 0$, or $D(\tilde{p}_{inf}) < 0$ and there exists a sequence $\tilde{p}^s_k$ converging to $\tilde{p}^s_{inf}$ either from below or from above such that $D(\tilde{p}^n, \tilde{p}^s_k) \to 0$.

If $D(\tilde{p}_{inf}) \geqslant 0$ or $\tilde{p}^s_k$ converges to $\tilde{p}^s_{inf}$ from above, there exists $\hat{p}$ such that $\tau(\hat{p}) = +\infty$ (i.e., no disclosure at $\hat{p}$ grants the maximal continuation payoff) and $D(\hat{p}) > -q$, with the latter implying that $\hat{p} \in R$. By deleting all bad reviews the seller can retain both groups of consumers in the market forever starting from $\hat{p}$. However, we know that $V^{\theta}(f(\hat{p}))$ is strictly smaller than the maximal possible payoff for seller of type θ, since this is true for any $p \in B_0$ with $p^s < 1$. Revealing a bad review at $\hat{p}$ is thus strictly suboptimal for either type of the seller, which contradicts $\hat{p} \in R$.

If $\tilde{p}^s_k$ converges to $\tilde{p}^s_{inf}$ from below, then for any $\varepsilon > 0$ and any $C > 0$ there exists $\hat{p}$ such that $D(\hat{p}) > -\varepsilon$ and $\tau(\hat{p}) > C$. The latter property is satisfiable, as improper integral in (9) diverges, and therefore for any $C > 0$ there exists some k such that $\tau(\tilde{p}^n, \tilde{p}^s_{inf} - \frac{1}{k}) > C$. As for the former, we know that $\tau(\tilde{p}^n, \tilde{p}^s_{inf}) - \tau(\tilde{p}^n, \tilde{p}^s_{inf} - \frac{1}{k}) = +\infty$, and therefore there exists $\hat{p}^s \in [\tilde{p}^s_{inf} - \frac{1}{k}, \tilde{p}^s_{inf}]$ such that $D(\tilde{p}^n, \hat{p}^s) > -\varepsilon$. As the seller's value $V^{\theta}(p)$ in any state $p \in B_0^{\uparrow}$ with $p^s < 1$ is strictly less than the maximal one and as C can be made arbitrarily large, we can find C large enough that the value of disclosure is strictly less than the value of staying silent. Since $\hat{p} \in R$ as long as $\varepsilon < q$, we achieve a contradiction. □

Proof of Proposition 2. It is shown in Lemma 7 that if $\mu < \frac{1}{2}$, then $r^L(p) = r^H(p) = 0$ for all $p \in B_1 \cap R$ is the only possible equilibrium strategy profile. To show the second condition, recall from Lemma 2 that a low-type seller must be indifferent between revealing a bad review at $p \in B_1^{\uparrow}$ and not, and that his indifference condition can be written as

$$\frac{1-\mu}{\mu} + e^{-r\tau(f(p))} = e^{-r\tau(p)}. \tag{15}$$

As $\tau(f(p)) \leqslant +\infty$, it should be that $\tau(p) \leqslant \frac{1}{r}\ln\frac{\mu}{1-\mu}$. Therefore, as $D(p) \geqslant -1$, we have $\ln\frac{p^*}{1-p^*} - \ln\frac{\bar{p}}{1-\bar{p}} \leqslant \frac{\lambda}{r}\ln\frac{\mu}{1-\mu}$ which gives the result. □

Proof of Proposition 3. The claim was already established for $B_1^{\downarrow}$ in the proof of Lemma 2 for B_1. We are left to show it for $B_1^{\uparrow}$. As Lemma 4 shows, we can construct a mapping g such that

$$\frac{1-\mu}{\mu} + e^{-r\tau(g(p))} = e^{-r\tau(p)}, \tag{16}$$

so $g(p) = f^s(p)$ for all $p \in R$. Further, $g(p)$ can be represented as

$$\ln\left(\frac{g(p)}{1-g(p)}\right) = J(p) + \ln\left(\frac{p^s}{1-p^s}\right) \tag{17}$$

for some function $J(p)$ which is differentiable in p^s. Taking the derivative of both sides of (16) with respect to p^s, we get

$$e^{-r\tau(p)} \cdot \frac{d\tau(p)}{dp^s} = e^{-r\tau(g(p))} \cdot \frac{d\tau(g(p))}{dg(p)} \frac{dg(p)}{dp^s}. \tag{18}$$

As is shown in Lemma 4, $\frac{d\tau(p)}{dp^s} = \left(\lambda p^s \left(1-p^s\right) \pi(p) D(p)\right)^{-1}$. Differentiating (17) we get

$$\frac{dg(p)}{dp^s} = \frac{e^{-J(p)} + p^s(1-p^s)\frac{dJ(p)}{dp^s}e^{-J(p)}}{\left[p^s + (1-p^s)e^{-J(p)}\right]^2}.$$

Plugging the three derivatives, we get that (18) corresponds to

$$e^{-r\tau(p)} \cdot \mu q = e^{-r\tau(g(p))} \cdot (-D(p)) \cdot \left[1 + p^s(1-p^s) \cdot \frac{dJ(p)}{dp^s}\right].$$

Plugging (16) into the expression above we get

$$(-D(p)) \cdot \left[1 + p^s(1-p^s) \cdot \frac{dJ(p)}{dp^s}\right] = q \cdot \left[\mu + (1-\mu) \cdot e^{r\tau(g(p))}\right]. \tag{19}$$

Consider state $(p^n, \bar{p}) \in B_1$. We know $\tau(p^n, \bar{p}) = 0$, therefore (16) implies $\tau(g(p^n, \bar{p})) > 0$, which in turn means $J(p^n, \bar{p}) > 0$. For any $p \in B_1^{\uparrow}$ we have $\tau(g(p)) > \tau(g(p^n, \bar{p})) > 0$, therefore there exists $\varepsilon > 0$ such that the RHS of (19) is larger than $q + \varepsilon$. If additionally $\frac{dJ(p)}{dp^s} < 0$, (19) implies $D(p) < -q - \varepsilon$, and consequently $J(p) \geqslant \ln(1+\varepsilon)$ by Lemma 5. It then follows from continuity of $J(p)$ in p^s that $J(p) > 0$ for all $p \in B_1^{\uparrow}$. For there to exist $p \in B_1^{\uparrow}$ such that $J(p) \leqslant 0$ there should exist $\tilde{p}$ such that $J(\tilde{p}) \in (0, \ln(1+\varepsilon))$ and $\frac{dJ(\tilde{p})}{dp^s} < 0$, which is ruled out by the argument above. □

Lemma 8. *If $\mu \geqslant \frac{1}{2}$, then for any set $\tilde{R} \subseteq B_1^{\downarrow}$ there exists an equilibrium with $B_1^{\downarrow} \cap R = \tilde{R}$.*

Proof. As Lemma 2 states, for $\mu \geqslant \frac{1}{2}$ the low-type seller is indifferent between disclosing a bad review and concealing it at all $p \in B_1^{\downarrow} \cap R$. This indifference is given by (15), and using $\tau(p) = 0$ for all $p \in B_1^{\downarrow} \cap R$ as well as the fact that $\tau(f(p)) = \frac{1}{\lambda q \mu}\left[\ln\left(\frac{f^s(p)}{1-f^s(p)}\right) - \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)\right]$ it can be rewritten as[36]

$$\left(\frac{\bar{p}}{1-\bar{p}} \cdot \frac{1-f^s(p)}{f^s(p)}\right)^{\frac{r}{\lambda q \mu}} = 2 - \frac{1}{\mu}$$

$$\Leftrightarrow \frac{f^s(p)}{1-f^s(p)} = \frac{p^s}{1-p^s} \cdot \frac{(1-q)\cdot r^H(p)}{r^L(p)} = \frac{\bar{p}}{1-\bar{p}}\left(2 - \frac{1}{\mu}\right)^{-\frac{\lambda q \mu}{r}}.$$

Next we consider incentives of a high-type seller. Since $r^H(p) > 0$, he should weakly prefer to reveal a bad review. We further show that this is always true whenever $\mu \geqslant \frac{1}{2}$ (and the preference is strict if $\mu > \frac{1}{2}$), and therefore $r^H(p) = 1$ constitutes an equilibrium in $B_1^{\downarrow}$. The value from revealing a bad review can be computed by plugging (15) and $\tau(p) = 0$ into (13) to obtain

$$V^H\left(f(p)\right) = \left(\frac{1}{r} - \frac{1-\mu}{r+\lambda q \mu}\right) \cdot \left(1 - \left(2 - \frac{1}{\mu}\right)^{1+\frac{\lambda q \mu}{r}}\right). \tag{20}$$

[36] As $D(p) = -q$ in $B_0^{\uparrow}$, the expression for $\tau(f(p))$ follows from (9) and the fact that $\pi(p) = \mu$ for $p \in B_0^{\uparrow}$.

The value of staying silent at p is no greater than supremum over all T of expected payoffs from staying silent until T and then receiving and disclosing a bad review exactly at T (with $T = +\infty$ allowed as an option to stay silent forever). The remainder of this proof shows that this amount is smaller than $V^H(f(p))$, which finalizes the argument. The supremum is equal to

$$\bar{V} = \sup_T \int_0^T e^{-rt}\left[1 - \mu \cdot e^{-\lambda q(1-\mu)t}\right] dt + e^{-rT}\left(e^{-\lambda q(1-\mu)T} \cdot V^H\left(f(p_T)\right) + \left(1 - e^{-\lambda q(1-\mu)T}\right) \cdot \frac{1}{r}\right).$$

By simplifying the expression above we obtain

$$\bar{V} = \sup_T \left(\frac{1}{r} - \frac{\mu}{r + \lambda q(1-\mu)}\right) \cdot \left(1 - e^{-(r+\lambda q(1-\mu))T}\right) + e^{-(r+\lambda q(1-\mu))T} \cdot V^H\left(f(p_T)\right),$$

which is a convex combination of $\left(\frac{1}{r} - \frac{\mu}{r+\lambda q(1-\mu)}\right)$ and $V^H(f(p_T))$. The latter is given by (20) and is thus independent of T. Therefore, to finalize the argument we need to show that

$$V^H(f(p)) \geqslant \left(\frac{1}{r} - \frac{\mu}{r + \lambda q(1-\mu)}\right),$$

which would mean that staying silent at p is weakly worse than revealing a bad review at p, and the equality is attained only when $\mu = \frac{1}{2}$. Indeed, the condition above is equivalent to

$$\frac{1}{1 + \frac{\lambda q(1-\mu)}{r}} \cdot \left(2 - \frac{1}{\mu}\right) \geqslant \left(2 - \frac{1}{\mu}\right)^{1 + \frac{\lambda q \mu}{r}}. \tag{21}$$

For $\mu = \frac{1}{2}$ the inequality is trivially satisfied with equality. And for $\mu \in (\frac{1}{2}, 1)$ we have

$$\left(1 - \frac{1-\mu}{\mu}\right)^{\frac{\lambda q \mu}{r}} < e^{-\frac{\lambda q(1-\mu)}{r}} < \frac{1}{1 + \frac{\lambda q(1-\mu)}{r}},$$

which concludes the argument. □

Band B_{2+}

Proof of Lemma 2 for $p \in B_{2+}$. Suppose not: there exists some $p \in R$ at which the low-type seller has strict preference. Depending on the direction of this preference, two cases are possible:

Case 1: $r^L(p) = 0$, $r^H(p) > 0$. Then $f^s(p) = 1$ and $f^n(p) \geqslant \bar{p}$, so by revealing this bad review and deleting all future ones the seller can guarantee himself the maximal possible continuation payoff. Therefore, deleting bad review at p cannot be strictly better than leaving it – a contradiction.

Case 2: $r^L(p) = 1$, $r^H(p) \leqslant 1$. It implies $D(p) \geqslant 0$. If $p^s \geqslant \bar{p}$, then this contradicts Lemma 1 for $p \in B_{1+}^{\uparrow}$. If, however, $p^s < \bar{p}$, then by Lemma 3 $D(p) \geqslant 0$ implies that $f^s(p) < p^s$ for bad reviews revealed at p, and therefore $f^s(p) < \bar{p}$. The low-type seller's value from revealing a bad review in $B_1^{\downarrow}$ is equal to the value of deleting all future bad reviews starting from $f(p)$. Deleting a bad review in $B_2^{\downarrow}$ can guarantee at least the same value by case of deleting all bad reviews. This means that despite we've assumed $r^L(p) = 1$, the low-type seller is indeed indifferent between disclosure and concealment at p.

We have shown that the low-type seller's value at any $p \in B_{2+}$ is equal to that from deleting all bad reviews starting from p, and the value of disclosure at p is equal to the value he gets deleting all bad reviews

starting from $f^s(p)$. Thus the indifference condition of the low-type seller results in

$$\int_0^{\tau(p)} e^{-rt}dt + (1-\mu)e^{-r\tau(p)} \int_0^{+\infty} e^{-rt}dt = \int_0^{\tau(f(p))} e^{-rt}dt + (1-\mu)e^{-r\tau(f(p))} \int_0^{+\infty} e^{-rt}dt,$$

which can be further reduced to

$$\tau(f(p)) = \tau(p). \tag{22}$$

This concludes the proof. □

Proof of Proposition 4. We show the claim for $B_2^{\uparrow}$. Induction to $B_k^{\uparrow}$ with $k > 2$ is straightforward.

As Lemma 4 shows, we can construct mapping g such that $\tau(g(p)) = \tau(p)$, and for some function $J(p)$ which is continuous in p^s we have:

$$\ln\left(\frac{g(p)}{1-g(p)}\right) = J(p) + \ln\left(\frac{p^s}{1-p^s}\right).$$

Now suppose that there exists $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_2^{\uparrow}$ such that $f^s(\tilde{p}) < \tilde{p}^s$. Then $g(\tilde{p}) = f^s(\tilde{p})$, and therefore $J(\tilde{p}) < 0$. As $J(p)$ is a continuous function of p^s and $J(\tilde{p}^n, \bar{p}) = 0$, there exists $\hat{p}^s < \tilde{p}^s$ such that $J(\tilde{p}^n, \hat{p}^s) = 0$ and $J(\tilde{p}^n, p^s) < 0$ for all $p^s \in (\hat{p}^s, \tilde{p}^s]$. Thus $g(\tilde{p}^n, p^s) \leqslant p^s$ for all $p^s \in [\hat{p}^s, \tilde{p}^s]$. Therefore, by Lemmas 3 and 4 we must have $D(\tilde{p}^n, p^s) \geqslant -q$ for all $p^s \in [\hat{p}^s, \tilde{p}^s]$. However, $D(p) \leqslant -q$ for all $p \in B_1^{\uparrow}$ which violates $\tau(g(\tilde{p})) - \tau(g(\hat{p})) = \tau(\tilde{p}) - \tau(\hat{p})$ given representation (9). □

Proofs of the Main Results

Proof of Theorem 1. The statement of the Theorem follows directly from Propositions 1, 3 and 4. □

Proof of Corollary 1. From Theorem 1 and expression (2) we have

$$(1-q) \cdot r^H(p) \geqslant r^L(p).$$

Therefore, if $r^H(p) = 0$, then we must have $r^L(p) = 0$ and $p \notin R$. If $r^H(p) > 0$, then

$$r^H(p) > (1-q) \cdot r^H(p) \geqslant r^L(p)$$

which proves the claim. □

Proof of Theorem 2. To prove the first part note that first, by Corollary 2 all continuation equilibria are payoff-equivalent in $B_0^{\uparrow}$. Next, if $\mu < \frac{1}{2}$, then Lemma 7 implies that $D(p) = -q$ for all $p \in B_1$, and therefore all continuation equilibria are payoff-equivalent in B_1 as well. As $B_1^{\downarrow} \cap R = \emptyset$ by Lemma 7 and p^s can never cross $\bar{p}$ from below, seller's value $V^{\theta}(p)$ for $p \in B_{2+}^{\downarrow}$ is equal in any equilibrium to the value of keeping naive consumers in the market forever. Finally, in any equilibrium $D(p) = -q$ for all $p \in B_{2+}^{\uparrow}$: by Theorem 1 and Lemma 3, $D(p) \leqslant -q$, and if there exists an equilibrium and $p \in B_{2+}^{\uparrow}$ with $D(p) < -q$, then $J(p) > 0$ by Lemma 3, which violates $\tau(f(p)) = \tau(p)$ as $D(p) = -q$ for all $p \in B_1$. This implies that $\tau(p)$ is constant across equilibria, which together with the above gives payoff-equivalence in $B_{2+}^{\uparrow}$.

If $\mu = \frac{1}{2}$, then $D(p) = -q$ for all $p \in B_1^{\uparrow} \cap R$ with $p^s > \bar{p}$. Since on any path of play the game only passes through one state in $B_1^{\uparrow}$ with $p^s = \bar{p}$ (which is the only state in $B_1^{\uparrow}$ where $D(p) < -q$ is possible), and drift there is still negative, $\tau(p)$ in any equilibrium must be the same as in case $\mu < \frac{1}{2}$ (where $B_1^{\uparrow} \cap R = \emptyset$) for all $p \in B_1^{\uparrow}$. The same logic as above can then establish that $D(p) = -q$ for all $p \in B_{2+}^{\uparrow}$. Finally, in case

$\mu = \frac{1}{2}$ it may be that $B_1^{\downarrow} \cap R \neq \emptyset$, but both types of seller are in any such p indifferent between revealing and deleting a bad review, and therefore receive the same payoff as if $B_1^{\downarrow} \cap R = \emptyset$.

The remainder of the proof is devoted to constructing an equilibrium that satisfies the requirements of the second part of Theorem 2. We propose a strategy profile and show that it satisfies all equilibrium conditions.

Construct the strategy profile as follows. Let $B_0^{\uparrow} \cap R = \emptyset$, and for all $p \in B_1^{\downarrow}$ build the strategy profile $\left(r^H, r^L\right)$ in such a way that $B_1^{\downarrow} \cap R = B_1^{\downarrow}$, – the latter is possible by Lemma 8.

For $\mu > \frac{1}{2}$ the inequality in (21) is strict for all $p \in B_1^{\downarrow}$, so by continuity of preferences of the high-type seller there exists $\varepsilon_1 > 0$ such that he strictly prefers to reveal at all $p \in \{B_1^{\uparrow} | p^s \in [\bar{p}, \bar{p} + \varepsilon_1)\}$, i.e., these states can belong to R in equilibrium. In all such states it must be that $r^H(p) = 1$, and $r^L(p)$ is then defined implicitly by (15). The latter can be reduced to the following differential equation for $J(p)$:

$$\left(1 - (1-q)e^{-J(p)}\right) \cdot \left[1 + p^s(1-p^s) \cdot \frac{dJ(p)}{dp^s}\right] = q \cdot \left[\mu + (1-\mu) \cdot \left(\frac{1-\bar{p}}{\bar{p}} \cdot \frac{p^s}{1-p^s}\right)^{\frac{r}{\lambda q \mu}} \cdot e^{\frac{r}{\lambda q \mu} J(p)}\right] \quad (23)$$

with an initial condition $J(p^n, \bar{p}) = -\frac{\lambda q \mu}{r} \ln\left(2 - \frac{1}{\mu}\right)$.[37] Then $r^L(p)$ can be obtained from $J(p) = \ln(1-q) - \ln r^L(p)$. By the existence theorem (see Pontryagin [1962], chapter 4, §21) a solution to (23) exists in some neighborhood of $(p^n, \bar{p})$, i.e., there exists $\varepsilon_2 > 0$ such that $J(p)$, and consequently $r^L(p)$, is well-defined for all $p = (p^n, p^s)$ with $p^s \in [\bar{p}, \bar{p} + \varepsilon_2)$. Take $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$ and set $r^L(p)$ for all $p \in \{B_1^{\uparrow} | p^s < \bar{p} + \varepsilon\}$ as prescribed by the procedure above. At the remaining states $p \in \{B_1^{\uparrow} | p^s \geqslant \bar{p} + \varepsilon\}$ set $r^H(p) = r^L(p) = 0$.

The strategy profile in B_{2+} is constructed as follows. For any $p \in B_{2+}^{\downarrow}$ let $r^H(p) = 1$ and $r^L(p) = (1-q) \cdot \left(\frac{p^s}{1-p^s} \cdot \frac{1-\bar{p}}{\bar{p}}\right)^{\frac{1}{2}}$ so that $J(p) = \frac{1}{2} \cdot \left(\ln\left(\frac{\bar{p}}{1-\bar{p}}\right) - \ln\left(\frac{p^s}{1-p^s}\right)\right) > 0$, meaning $\bar{p} > f^s(p) > p^s$. In $B_{2+}^{\uparrow}$ let $r^H(p) = r^L(p) = 0$ for $p \in \{B_{2+}^{\uparrow} | p^s = \bar{p}\}$. Let $r^H(p) = 1$ for all $p \in \{B_{2+}^{\uparrow} | p^s > \bar{p}\}$. We compute $r^L(p)$ inductively over bands, where the induction statement is "$r^L(p)$ is constructed for all $p \in B_k^{\uparrow}$ and it is such that $D(p) \leqslant -q$". This is true by construction for $k = 1$, which starts the induction. Suppose it holds for $k-1$. For $p \in B_k^{\uparrow}$ we construct $r^L(p)$ so that (22) holds. In particular, consider a change of variable $z = \ln\left(\frac{p^s}{1-p^s}\right)$ and let $J(p^n, z)$ represent, with abuse of notation, the respective transformation of $J(p)$, i.e., $J(p^n, z) = \ln(1-q) - \ln r^L\left(p^n, \frac{e^z}{1+e^z}\right)$. Then taking the derivatives of both sides of (22) with respect to z, we obtain the following differential equation for $J(p^n, z)$:

$$\left(1 - (1-q)e^{-J(p^n,z)}\right) \cdot \left[1 + \frac{dJ(p^n, z)}{dz}\right] = -D\left(f^n(p), z + J(p^n, z)\right) \quad (24)$$

with the initial condition $J\left(p^n, \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)\right) = 0$.[38]

We next show that a solution to (24) exists and is nonnegative for all $z \geqslant \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)$. Suppose that there exists $p = (p^n, p^s) \in B_k^{\uparrow}$ such that $J\left(p^n, \ln\left(\frac{p^s}{1-p^s}\right)\right) = -\varepsilon < 0$. As a solution to an ODE, $J(p^n, z)$ is a continuous function of z. Therefore, there exists $\tilde{p}^s \in (\bar{p}, p^s)$ such that $J\left(p^n, \ln\left(\frac{\tilde{p}^s}{1-\tilde{p}^s}\right)\right) = \max\{-\frac{1}{2}\varepsilon, \frac{1}{2}\ln(1-q)\}$. Then

$$\left.\frac{dJ(p^n, z)}{dz}\right|_{z=\ln\left(\frac{\tilde{p}^s}{1-\tilde{p}^s}\right)} = \left.\frac{-D(f^n(p), z + J(p^n, z))}{1 - (1-q)e^{-J(p^n,z)}}\right|_{z=\ln\left(\frac{\tilde{p}^s}{1-\tilde{p}^s}\right)} - 1 > \frac{q}{q} - 1 = 0.$$

[37] This initial condition is such that $J(p)$ is continuous at $(p^n, \bar{p})$.

[38] The RHS of (24) is not smooth in $J(p^n, z)$, but is piecewise smooth. Therefore, as a solution to (24) we take a composition of two solutions which are pasted together using continuity.

Therefore, as we increase z from $\ln\left(\frac{\tilde{p}^s}{1-\tilde{p}^s}\right)$, $J(p^n, z)$ could never fall below $-\frac{\varepsilon}{2}$, while we have assumed $J\left(p^n, \ln\left(\frac{p^s}{1-p^s}\right)\right) = -\varepsilon$ – a contradiction. As $\varepsilon > 0$ was taken arbitrarily, it shows that $J(p^n, z) \geqslant 0$ for all $z \geqslant \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)$. We next can take arbitrary solution to (24) in the neighborhood of its initial condition, the existence of which is ensured by the existence theorem (see Pontryagin [1962], chapter 4, §21). It can be extended for all $z \geqslant \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)$ if and only if $J(p^n, z) < +\infty$ for all such z (see Pontryagin [1962], chapter 4, §24) which is true as

$$\left|\frac{dJ(p^n, z)}{dz}\right| < \frac{1}{q} - 1 = \frac{1-q}{q} < +\infty.$$

Consequently, by Lemma 3 we obtain that $D(p) \leqslant -q$ for all $p \in B_k^{\uparrow}$, which concludes this part of the proof.

We next show that the constructed strategy profile constitutes an equilibrium. We first show that the low-type seller is indifferent whether to reveal a bad review or to conceal it at all $p \in B_{2+} \cap R$. If $p \in B_{2+}^{\downarrow}$, then by construction $0 < r^L(p) < 1$. From Lemma 8 we also know that $0 < r^L(p) < 1$ for $p \in B_1^{\downarrow} \cap R$. Then the value of a low-type seller in any $p \in B_{1+}^{\downarrow} \cap R$ is equal to the value he receives in case he deletes all future bad reviews: $V^L(p) = \frac{1-\mu}{r}$. Therefore, a low-type seller is indeed indifferent between disclosing a bad review and deleting it for any $p \in B_{2+}^{\downarrow} \cap R$. For $p \in B_{1+}^{\uparrow} \cap R$ the indifference directly follows from the way $r^L(p)$ is constructed and the fact that $r^L(p) < 1$.[39]

By construction, the high-type seller strictly prefers to reveal bad reviews at all $p \in B_1 \cap R$. We proceed by showing that the high-type seller weakly prefers to reveal a bad review at all $p \in B_{2+}^{\downarrow} \cap R$. Concealing a review at $p \in B_2^{\downarrow} \cap R$ cannot yield him a payoff higher than if he could choose time T at which a bad review will arrive and will be revealed:

$$\begin{aligned} V^H(p) &\leqslant \max_{T>0}\left\{\int_0^T e^{-(r+\lambda q(1-\mu))t}\cdot(1-\mu)\cdot\left(1+\frac{\lambda q}{r}\right)dt + e^{-(r+\lambda q(1-\mu))T}\cdot V^H\left(f\left(p_T\right)\right)\right\} \\ &= \max_{T>0}\left\{\left(1-e^{-(r+\lambda q(1-\mu))T}\right)\cdot\left(\frac{1}{r}-\frac{\mu}{r+\lambda q\left(1-\mu\right)}\right) + e^{-(r+\lambda q(1-\mu))T}\cdot V^H\left(f\left(p_T\right)\right)\right\} \\ &\leqslant \max_{T>0} V^H\left(f\left(p_T\right)\right) = V^H\left(f\left(p\right)\right) \end{aligned}$$

where process p_t is given by (4) with initial condition $p_0 = p$. The last inequality holds because

$$V^H(p) \geqslant \int_0^{+\infty} e^{-rt}\left[e^{-\lambda q(1-\mu)t}\cdot(1-\mu) + \left(1-e^{-\lambda q(1-\mu)t}\right)\right]dt = \left(\frac{1}{r}-\frac{\mu}{r+\lambda q(1-\mu)}\right)$$

for all $p \in B_1^{\downarrow}$ since the high-type seller can delete all future bad reviews. The last equality holds because $V^H(f(p_T))$ is independent of T. Indeed, distributions of arrival times of the next buying consumer are the same for all $p \in B_1^{\downarrow}$. Therefore, $V^H(p)$ is the same for all $p \in B_1^{\downarrow}$. The resulting inequality $V^H(p) \leqslant V^H(f(p))$ implies that the high-type seller weakly prefers to reveal a bad review at all $p \in B_2^{\downarrow}$. The argument above can be extended by induction to all further bands in order to obtain that $V^H(p) \leqslant V^H(f(p))$ for all $p \in B_{2+}^{\downarrow}$.

We are left to show that the high type at least weakly prefers to reveal a bad review in $B_{2+}^{\uparrow}$. We show the argument for $B_2^{\uparrow}$, and the argument for $B_k^{\uparrow}$ with higher k then follows by induction. Fix some point $p = (p^n, p^s) \in B_2^{\uparrow} \cap R$. The high-type seller's value in case he decides to conceal a bad review at p is bounded

[39] The latter is true as $J(p) < +\infty$ for all $p \in B_{1+}^{\uparrow} \cap R$.

from above by his payoff when he can receive and reveal a bad review at any time T of his choice:

$$V^H(p) \leqslant \max\left\{\max_{T\leqslant\tau(p)}\int_0^T e^{-rt}\left(1+\frac{\lambda q}{r}\right)dt + e^{-rT}\cdot V^H(f(p_T)),\ \int_0^{\tau(p)} e^{-rt}\left(1+\frac{\lambda q}{r}\right)dt + e^{-r\tau(p)}\cdot V^H(p^n,\bar{p})\right\}$$

where we use that $p^s_{\tau(p)} = \bar{p}$. On the one hand, since deleting all bad reviews is always feasible for the high-type seller, we have

$$V^H(f(p)) \geqslant \int_0^{\tau(f(p))} e^{-rt}\left(1+\frac{\lambda q}{r}\right)dt + e^{-r\tau(f(p))}\cdot V^H(f^n(p),\bar{p}) \geqslant \int_0^{\tau(p)} e^{-rt}\left(1+\frac{\lambda q}{r}\right)dt + e^{-r\tau(p)}\cdot V^H(p^n,\bar{p})$$

where the second inequality follows because by construction $\tau(p) = \tau(f(p))$, and $V^H(f^n(p),\bar{p}) \geqslant V^H(p^n,\bar{p})$ as shown above.[40] On the other hand, for any $T \leqslant \tau(p)$ we can write

$$V^H(f(p)) \geqslant \int_0^{T} e^{-rt}\left(1+\frac{\lambda q}{r}\right)dt + e^{-rT}\cdot V^H(f(p_T))$$

because the high-type seller can reveal no reviews during $[0,T]$, and because if $p_T \in R$, then the process given by (4) with starting point $f(p)$ reaches value $f(p_T)$ at exactly time T (this is since $\tau(p) = \tau(f(p))$ for all $p \in B_2^{\uparrow} \cap R$), while if $p_T \notin R$, then $V^H(f(p_T)) = 0$. Everything said above implies that $V^H(f(p)) \geqslant V^H(p)$ for all $p \in B_{2+}^{\uparrow}$, which concludes the proof that strategy profile is an equilibrium.

Finally, to conclude the proof of Theorem 2 we need to show that the equilibrium has the desired properties. We start with the fact that $f^s(p) > p^s$ for all $p \in R$. By construction the strategy profile already implies $f^s(p) > p^s$ for all $p \in \left(B_0 \cup B_1 \cup B_{2+}^{\downarrow}\right) \cap R$. We next establish the claim for $p \in B_2^{\uparrow}$. From (24) we know that $J(p) \geqslant 0$ for all $p \in B_{2+}^{\uparrow}$. Suppose by way of contradiction that there exists some $\tilde{p} = (\tilde{p}^n, \tilde{p}^s) \in B_2^{\uparrow}$ such that $f^s(\tilde{p}) = \tilde{p}^s$, i.e., $J(\tilde{p}) = 0$. Assume first that $f(\tilde{p}) \in R$ which by Lemma 3 implies $D(f(\tilde{p})) < -q$. Then there exists $\varepsilon > 0$ such that $D(f^n(\tilde{p}), p^s) \leqslant -q-\varepsilon$ for all $p^s \in [f^s(\tilde{p}) - \frac{\varepsilon}{4}, f^s(\tilde{p})]$ as, by construction, $D(p)$ is continuous in p^s.[41] At the same time, we have that $J(p) < J(\tilde{p}) + \frac{\varepsilon}{2} = \frac{\varepsilon}{2}$ for all $p = (\tilde{p}^n, p^s)$ with $p^s \in [\tilde{p}^s - \frac{\varepsilon}{4}, \tilde{p}^s]$. By converse of Lemma 5 this implies that $D(p) > -q - \frac{\varepsilon}{2}$ for those p. Therefore $\tau(\tilde{p}) - \tau(\tilde{p}^n, \tilde{p}^s - \frac{\varepsilon}{4}) > \tau(f(\tilde{p})) - \tau(f^n(\tilde{p}), f^s(\tilde{p}) - \frac{\varepsilon}{4})$. Consequently, $J(\tilde{p}^n, \tilde{p}^s - \frac{\varepsilon}{4}) < 0$, – a contradiction. Now assume $f(\tilde{p}) \notin R$, that is $D(f(\tilde{p})) = -q$. Then (24) can be solved explicitly. Its general solution satisfies

$$(1-q)(z + J(p^n, z)) + q\ln\left(1 - e^{J(p^n,z)}\right) = C \tag{25}$$

where C is a constant pinned down by the boundary condition for z_0 where $z_0 = \inf\{z | f(p^n, z) \notin R\}$ and $J(p^n, z_0) > 0$ is given as a solution to (24) for $z \in \left[\ln\left(\frac{\bar{p}}{1-\bar{p}}\right), z_0\right]$ with initial condition $J\left(\ln\left(\frac{\bar{p}}{1-\bar{p}}\right)\right) = 0$. Therefore, C is well-defined and finite. As we have assumed $J(\tilde{p}) = 0$ for some $\tilde{p}$, substituting it into (25) we achieve $C = -\infty$, – a contradiction.

All said above shows that $J(p) > 0$ for all $p \in B_2^{\uparrow}$. By Lemma 3 it implies $D(p) < -q$ for all $p \in B_2^{\uparrow}$, and the argument then extends to further bands straightforwardly.

To see that this equilibrium is not payoff-equivalent to an equilibrium with $R = \emptyset$, note that, for instance,

[40] Values at the cutoff are equal to respective values under the cutoff since the latter are constant, and total payoff is insensitive to alterations of flow payoff in a single point (i.e., the fact that strategic consumers are buying in $p^s = \bar{p}$ does not affect payoffs).

[41] Otherwise there exists a sequence of points $\{p^s_k\}$ such that $p^s_k \to f^s(\tilde{p})$ and $D(f^n(\tilde{p}), p^s_k) \to -q$ as $k \to +\infty$ which contradicts the continuity of $D(p)$ in $B_1^{\uparrow}$.

the equilibrium constructed above has $D(p) < -q$ for $p \in \{B_1^{\uparrow} | p^s \in [\bar{p}, \bar{p} + \varepsilon)\}$, as opposed to $D(p) = -q$ in the fully censored equilibrium, meaning that $\tau(p)$ is smaller in the former for all $p \in B_1^{\uparrow}$. Noticing that $\tau(p)$ directly enters the low-type seller's value in $B_1^{\uparrow}$ concludes the argument. □

Proofs for Section 6

Proof of Proposition 5. By Lemma 2 the low-type seller is indifferent between revealing a bad review and deleting it at all $p \in B_{1+} \cap R$. Therefore, $V^L(p) = \frac{1-\mu}{r}$ for all $p \in B_{1+}^{\downarrow}$ irrespective of equilibrium. For $p \in B_{1+}^{\uparrow}$ we have

$$V^L(p) = \frac{1-\mu}{r} + (1 - e^{-r\tau(p)}) \cdot \frac{\mu}{r} = \frac{1 - \mu e^{-r\tau(p)}}{r}.$$

Therefore, to show the claim we need to establish that larger R implies pointwise weakly smaller $\tau(p)$. The claim holds for $B_0^{\uparrow}$ (larger $B_0^{\uparrow} \cap R$ has no effect on $\tau(p)$ for $p \in B_0^{\uparrow}$). Proceed by induction and show that if the claim holds for $B_{k-1}^{\uparrow}$, then it also holds for $B_k^{\uparrow}$. For any $p \in B_k^{\uparrow}$ we show that if $\tau'(p) = \tau''(p)$, then $\frac{d\tau'(p)}{dp} \leqslant \frac{d\tau''(p)}{dp}$ where objects indexed by single and double primes denote respective objects in the two equilibria under consideration with R' and $R'' \subset R'$ respectively. Three cases are possible for every p with $\tau'(p) = \tau''(p)$:

1. If $p \notin R'$, then $D'(p) = D''(p) = -q$.
2. If $p \in R' \backslash R''$, then $D'(p) \leqslant -q = D''(p)$, where the first inequality follows from Theorem 1 and Lemma 3.
3. If $p \in R''$, then $\tau'(f(p)) \leqslant \tau''(f(p))$ implies that $J'(p) \geqslant J''(p)$, which in turn means that $D'(p) \leqslant D''(p)$ because both equilibria are semi-separating.

Therefore, $D'(p) \leqslant D''(p)$ for all $p \in B_k^{\uparrow}$. Since $\tau(p^n, \bar{p}) = 0$ for all p^n, (9) implies that $\tau'(\tilde{p}) \leqslant \tau''(\tilde{p})$ for all $p \in B_k^{\uparrow}$. □

Proof of Proposition 6. As the seller of a high quality product never receives any bad review, after any bad review beliefs jump to $f^s(p) = f^n(p) = 0$ and no future consumers ever buy the product again. Revealing a bad review thus grants the worst continuation payoff, and is therefore strictly dominated by deleting it for any seller who can guarantee non-zero continuation payoff which is true if either $p^n \geqslant \bar{p}$ or $p^s > \bar{p}$. □

Proof of Proposition 7. First let us introduce some extra notation for the general setting. Let $B_{-1} = \{(p^n, p^s) \in B_0 | (f_-^n)^{-1}(p^n) \geqslant \bar{p}\}$ and $B_{-k} = \{(p^n, p^s) | \left((f_-^n)^{-1}(p), p^s\right) \in B_{-k+1}\}$ for $k > 1$.[42] By analogy with B_k for $k > 0$, B_{-k} measure distance between p^n and $\bar{p}$: if $p \in B_{-k}$ for $k > 0$, then k less bad reviews would be required to bring naive consumers back to the market.

Let us also refresh the expressions for belief updating for the general case. Rational consumers' beliefs are updated in the general setting as:

$$\frac{f_+^s(p)}{1 - f_+^s(p)} = \frac{p^s}{1 - p^s} \cdot \frac{q_+^H r_+^H(p)}{q_+^L r_+^L(p)}; \qquad \frac{f_-^s(p)}{1 - f_-^s(p)} = \frac{p^s}{1 - p^s} \cdot \frac{q_-^H r_-^H(p)}{q_-^L r_-^L(p)} \tag{26}$$

after good and bad reviews respectively, and as

$$\dot{p}^s = \lambda p^s (1 - p^s) \cdot \left[q_+^H \left(1 - r_+^H(p)\right) + q_-^H \left(1 - r_-^H(p)\right) - q_+^L \left(1 - r_+^L(p)\right) - q_-^L \left(1 - r_-^L(p)\right)\right] \tag{27}$$

[42] Here function f^n is meant in the sense of $[0,1] \to [0,1]$ (i.e., $f^n(p^n)$) since, as we remember, $f^n(p)$ does not depend on p^s.

in the absence of reviews. Naive consumers' reaction to good and bad reviews respectively is given by:

$$\frac{f_+^s(p)}{1-f_+^s(p)} = \frac{p^s}{1-p^s}\cdot\frac{q_+^H}{q_+^L}; \qquad\qquad \frac{f_-^s(p)}{1-f_-^s(p)} = \frac{p^s}{1-p^s}\cdot\frac{q_-^H}{q_-^L}.$$

We construct the equilibrium as follows. For good reviews let $R_+ = B_{-1}^{\uparrow}$ and $r_+^\theta(p) = 1$ for either θ and all $p \in R_+$. For bad reviews let $R_- = \cup_{k\geqslant 1} B_k^{\downarrow}$ and $r_-^H(p) = 1$ for all $p \in R_-$. Let $r_-^L(p)$ for $p \in B_{2+}^{\downarrow}$ be constructed as in Theorem 2. Finally, $r_-^L(p)$ for $p \in B_1^{\downarrow}$ is constructed below.

We construct $r_-^L(p)$ for $p \in B_1^{\downarrow}$ in such a way as to make the low-type seller indifferent between revealing a bad review and not. In such construction, $V^L(p) = \frac{1-\mu}{r}$ for any $p \in B_1^{\downarrow}$ (and actually all $p \in B_{1+}^{\downarrow}$ given the remainder of the construction), so deleting all future bad reviews is optimal. On the other hand, for any $p \in B_{-1}^{\uparrow}$ we have

$$\begin{aligned} V^L(p) &= \int_0^{\tau(p)} e^{-rt}\left[e^{-\lambda\mu q_+^L t}\cdot\mu + \left(1-e^{-\lambda\mu q_+^L t}\right)\cdot 1\right]dt + e^{-r\tau(p)}\left(1-e^{-\lambda\mu q_+^L\tau(p)}\right)\cdot\frac{1}{r} = \\ &= \left(1-e^{-(r+\lambda\mu q_+^L)\tau(p)}\right)\cdot\left(\frac{1}{r}-\frac{1-\mu}{r+\lambda\mu q_+^L}\right). \end{aligned}$$

To clarify, this expression describes payoff from selling to strategic consumers until $\tau(p)$ and to all consumers after a good review arrives if this happens before $\tau(p)$. The latter is valid because condition $q_+^H\cdot q_-^H \geqslant q_+^L\cdot q_-^L$ ensures that revealing one additional good review in any $p \in B_{-1}^{\uparrow}$ brings naive consumers back to the market.

Given the strategies defined above, $D(p) = -(q_+^H - q_+^L) < 0$ for all $p \in B_{-1}^{\uparrow}$, hence

$$\tau(p) = \frac{1}{\lambda\mu(q_+^H - q_+^L)}\left(\ln\left(\frac{p^s}{1-p^s}\right) - \ln\left(\frac{\bar{p}}{1-\bar{p}}\right)\right) < \infty \tag{28}$$

for all $p = (p^n, p^s) \in B_{-1}^{\uparrow}$. Furthermore, $\tau(p)$ is continuous and strictly increasing in p^s, so $V^L(p)$ is continuous and strictly increasing in p^s as well. Finally, $\tau(p) \to \infty$ as $p^s \to 1$ and $\tau(p) \to 0$ as $p^s \to \bar{p}$, therefore $V^L(p)$ spans the whole interval $\left[0, \frac{1}{r} - \frac{1-\mu}{r+\lambda\mu q_+^L}\right]$ across $p \in B_0^{\uparrow}$.

Fix some $p \in B_1^{\downarrow}$. Let $\hat{p} \in B_{-1}^{\uparrow}$ be such that $V^L(\hat{p}) = \frac{1-\mu}{r}$. It exists for reasons described above: $\frac{1-\mu}{r} < \frac{1}{r} - \frac{1-\mu}{r+\lambda\mu q_+^L}$ whenever $\mu > \frac{1}{2}$. Finally, let $r_-^L(p)$ for $p \in B_1^{\downarrow}$ be such that $f_-(p) = (f_-^n(p^n), \hat{p}^s)$ (closed-form expression for $r_-^L(p)$ can be obtained from (26)).

The construction above trivially implies $f_-^s(p) > \bar{p} > p^s$ for all $p \in B_1^{\downarrow}$. It also generates $f_+(p) > p^s$ for all $p \in R_+$. Construction in Theorem 2 also implies that $f_-^s(p) > p^s$ for all $p \in B_{2+}^{\downarrow}$. This verifies the first property in the Proposition. The second property is trivial – R_- is nonempty for the strategy profile constructed above. Therefore, to conclude the proof we need to verify two things: that the constructed strategy profile constitutes an equilibrium and that this equilibrium is payoff-distinct from fully censored equilibrium in any meaning of the latter.

We start by verifying that the strategy profile above constitutes an equilibrium. First, either type of the seller at least weakly prefers to reveal good reviews at all $p \in R_+$. This is because $f_+(p) \in B_1^{\uparrow}$ so $D(f(p)) = 0$ and $\tau(f(p)) = \infty$.[43] Simply speaking, revealing a good review moves seller to an absorbing state in which he can retain both naive and strategic consumers in the market forever. This attains the maximal payoff, so is always at least weakly optimal.

Low-type seller is by construction indifferent between deleting and revealing bad reviews at all $p \in B_1^{\downarrow}$.

[43] In case $q_+^H\cdot q_-^H < q_+^L\cdot q_-^L$ which we do not consider in this proposition, one would need to either ensure that prior p_0 is such that $f_+^n(f_-(p)) \geqslant \bar{p}$ for all $p \in B_1^{\downarrow}$ on equilibrium path, or to verify that the argument to follow holds even if more than one good review is required to achieve $B_1^{\uparrow}$ from any $p \in B_{-1}^{\uparrow}$.

This indifference extends to $B^{\downarrow}_{2+}$. If in any $p \in B^{\downarrow}_2$ the low-type seller chooses to delete a bad review, he can achieve a payoff of $\frac{1-\mu}{r}$ by deleting all future bad reviews as well. At the same time, revealing a bad review at p (or any future state) grants him $V^L(f(p)) = \frac{1-\mu}{r}$ which is exactly the same payoff. The argument can be iterated further to show that the low type is indifferent at all $B^{\downarrow}_{2+}$.

The only equilibrium property left to verify is the high type's preference. Suppose that the high-type seller is currently in some state $p \in B^{\downarrow}_1$. If he deletes all future bad reviews, then his payoff equals $\frac{1-\mu}{r}$. If, however, he has a bad review in hand and reveals it, then he arrives at some $f(p)$ with $f^s(p) = \hat{p}^s$ and receives

$$V^H(f(p)) = \int_0^{\tau(f(p))} e^{-rt}\left[e^{-\lambda\mu q^H_+ t}\cdot\mu + \left(1-e^{-\lambda\mu q^H_+ t}\right)\cdot 1\right]dt + e^{-r\tau(f(p))}\left(1-e^{-\lambda\mu q^H_+\tau(f(p))}\right)\cdot\frac{1}{r} =$$
$$= \left(1-e^{-(r+\lambda\mu q^H_+)\tau(f(p))}\right)\cdot\left(\frac{1}{r}-\frac{1-\mu}{r+\lambda\mu q^H_+}\right).$$

Given that $q^H_+ > q^L_+$ and that the low type's indifference dictates $\left(1-e^{-(r+\lambda\mu q^L_+)\tau(f(p))}\right)\cdot\left(\frac{1}{r}-\frac{1-\mu}{r+\lambda\mu q^L_+}\right) = \frac{1-\mu}{r}$, trivially $V^H(f(p)) > \frac{1-\mu}{r}$. Doing the usual argument with the high-type seller solving a relaxed problem in which he has a choice of when to reveal the bad review (used in proofs of Lemma 8 and Theorem 2), we can arrive at the conclusion that he strictly prefers to reveal a bad review at p. Using the same argument as in the proof of Theorem 2 we can then show that this strict preference propagates to $B^{\downarrow}_{2+}$. This concludes the proof that the constructed strategy profile is an equilibrium.

Finally, we want to show that the equilibrium above is payoff-distinct from fully censored equilibrium in either sense of the latter (i.e., where $R_- = \emptyset$ and R_+ is either same as above, or also empty). In either case it is enough to consider $V^H(p)$ at any $p \in B^{\downarrow}_1$. In either fully censored equilibrium we have $V^H(p) = \frac{1-\mu}{r}$ because the high-type seller is unable to reveal any reviews. In contrast, in the equilibrium constructed above $V^H(p) > \frac{1-\mu}{r}$ because this inequality is true for all $p \in B^{\uparrow}_0$ and the high-type seller jumps to $B^{\uparrow}_0$ from $B^{\downarrow}_1$ (by receiving and revealing a bad review) with a positive probability in finite time. □

Proof of Proposition 8. We construct the equilibrium in a way analogous to Proposition 7 but accounting for fake reviews. For good reviews let $R_+ = B^{\uparrow}_{-1}$ and $r^\theta_+(p) = \phi^\theta_+(p) = 1$ for either θ and all $p \in R_+$. For bad reviews let $R_- = \cup_{k\geqslant 1} B^{\downarrow}_k$ and $r^H_-(p) = \phi^H_-(p) = 1$ for all $p \in R_-$. For any $p \in B^{\downarrow}_{2+}$ let $r^L_-(p)$ and $\phi^L_-(p)$ be an arbitrary solution of the equation $\ln\left(\frac{f^s_-(p)}{1-f^s_-(p)}\right) - \ln\left(\frac{p^s}{1-p^s}\right) = \frac{1}{2}\cdot\left(\ln\left(\frac{\bar{p}}{1-\bar{p}}\right) - \ln\left(\frac{p^s}{1-p^s}\right)\right)$.[44]

Finally, $r^L_-(p)$ and $\phi^L_-(p)$ for $p \in B^{\downarrow}_1$ are constructed in such a way as to make the low type indifferent between revealing bad reviews and not. Similarly to Proposition 7 in such construction we have $V^L(p) = \frac{1-\mu}{r}$ for any $p \in B^{\uparrow}_1$, while for any $p \in B^{\uparrow}_0$: $V^L(p) = \left(1-e^{-(r+\lambda\mu q^L_+)\tau(p)}\right)\cdot\left(\frac{1}{r}-\frac{1-\mu}{r+\lambda\mu q^L_+}\right)$.

Given the strategies defined above, $D(p) = -(q^H_+ - q^L_+) < 0$ for all $p \in B^{\uparrow}_{-1}$ as in Proposition 7 (since effects of fake positive reviews on $D(p)$ imposed by high and low type cancel each other out). Further, $\frac{f^s_+(p)}{1-f^s_+(p)} = \frac{p^s}{1-p^s}\cdot\frac{\lambda q^H_+ + \lambda_\phi}{\lambda q^L_+ + \lambda_\phi}$ for all $p \in B^{\uparrow}_{-1}$, meaning that $f^s_+(p) > p^s$ so $f_+(p) \in B^{\uparrow}_1$ for all $p \in B^{\uparrow}_{-1}$.

From here the fact that this strategy profile is an equilibrium and all required equilibrium properties can be verified in exactly the same way as in Proposition 7. □

[44] This is analogous to the construction in Theorem 2. It ensures that $f^s_-(p) > p^s$ and $f_-(p) \in B^{\downarrow}$.