Decentralizing Dispute Resolution in Two-Sided Platforms: The Case of Review Blackmail

Yiangos Papanastasiou
Haas School of Business · University of California, Berkeley, yiangos@haas.berkeley.edu
S. Alex Yang
London Business School, sayang@london.edu
Angela Huyue Zhang
The University of Hong Kong - Faculty of Law, angelaz@hku.hk

We study the relative merits of different dispute resolution mechanisms in two-sided platforms, in the context of disputes involving malicious reviews and blackmail. We develop a game-theoretic model of the strategic interactions between a seller firm and a (potentially malicious) consumer. In our model, the seller takes into account the impact of consumer reviews on his future earnings; recognizing this, a malicious consumer may attempt to blackmail the seller by purchasing the product, posting a negative review, and demanding ransom to remove it. Without a dispute resolution mechanism in place, the presence of malicious consumers in the market can lead to a significant decrease in firm profit, especially in settings characterized by high uncertainty about product quality. The introduction of a standard "centralized" dispute resolution mechanism (whereby the firm can report potentially malicious reviews to the host platform, which then judges whether to remove the review) can restore efficiency to some extent, but requires the platform's judgments to be both very quick and highly accurate. We demonstrate that an appropriately-designed "decentralized" mechanism (whereby the firm is allowed to remove reviews without consulting the platform, subject to ex post penalties for wrongdoing) can be much more effective, while simultaneously alleviating—almost entirely—the need for the platform's judgments to be quick. These results suggest that decentralization, when implemented correctly, may be a more efficient approach to dispute resolution.

Key words: platform governance, dispute resolution, blackmail, decentralization, reviews, extortion

1. Introduction

Adjudicating disputes between market participants is one of the core functions performed by online platforms connecting sellers to consumers. It is a difficult function to manage: on one hand, buyers and sellers expect their complaints and disputes to be resolved in a timely and efficient manner; on the other, the sheer volume of interactions occurring inside the platform often means that allocating

the necessary resources to do so is prohibitively costly.[1,2] It is also a consequential function: as a matter of seller support and customer service, a platform's inability to deal with disputes efficiently erodes merchants' and consumers' relationship with the platform, causing loss of goodwill which may lead to decreased marketplace participation.

The standard approach to resolving disputes in online platforms follows the traditional model of a service firm: If a seller or buyer feels aggrieved, she can report the incident to the platform, which then conducts an investigation and decides the matter. However, unlike the traditional model of a service firm, this "centralized" process can be particularly inefficient in the case of online platforms, where the investigation often requires (a potentially lengthy process of) collecting information held privately by the parties involved and reconciling conflicting accounts of the events.

Recognizing the challenges associated with the traditional centralized approach, some platforms have experimented with more decentralized forms of governance, aimed at alleviating the demand for platform resources and enhancing the legitimacy of the adjudication process (e.g., in terms of fairness and transparency). One such approach, pioneered by eBay, involves "crowd-judging" (see Rule and Nagarajan 2010), whereby disputes are adjudicated by a panel of volunteers drawn from the platform's buyers and sellers. More recently, Taobao—the world's largest online retailer by gross merchandise value—has experimented with another form of decentralization, which grants one of the parties involved in the dispute (in this case the seller) the authority to adjudicate the dispute, subject to the possibility of ex post review by the platform and penalties for wrongdoing.

The introduction of the latter mechanism emerged in part as a result of seller complaints regarding "review blackmail."[3] In a typical instance of review blackmail, an opportunistic consumer purchases the seller's product, posts a negative review, and then demands ransom in order to remove it. Over a series of public hearings held by Taobao in 2018, sellers lamented having to allocate significant resources to deal with such attacks on a daily basis, noting that their response was often to pay the ransom for fear of suffering significant damage in their reputation and sales while

[1] For example, eBay handles more than 60 million buyer-seller disputes each year (Rule and Nagarajan 2010).

[2] While automation provides a solution for some forms of dispute, others are more nuanced and cannot be adjudicated without human intervention.

[3] Review blackmail has been a significant problem for online marketplaces based in China for at least a decade (over a six-month period in 2012, Zhang et al. (2020) empirically document some twenty-six thousand online sellers in a single product category were the victims of at least one blackmail attempt). More recently, US-based TripAdvisor has responded to growing concerns regarding review blackmail by introducing a formal procedure through which sellers can report such attempts (TripAdvisor 2018). In the UK, in response to a query by the Competition and Markets Authority, the British Hospitality Association reported that *all* of its members had suffered from "blackmail, malicious or patently false reviews" (Competition and Markets Authority 2015).

their cases were pending resolution by the platform.[4] The rationale behind Taobao's decentralized approach is that, when faced with a blackmail attempt, the seller can now bypass the platform and remove the malicious review on his own; moreover, knowing this, the malicious consumer may be less likely to attempt blackmail in the first place.

Motivated by the above developments, in this paper we investigate the relative merits of different mechanisms for dispute resolution. In particular, we are interested in understanding whether and to what extent such mechanisms may lead to improved outcomes of disputes between platform participants, and if so how these outcomes can be achieved.

To keep our analysis grounded, we focus on the motivating context of review blackmail described above. We develop a stylized model focusing on the strategic interactions between a monopolist seller and a (potentially malicious) consumer. The seller cares about the impact of product reviews on his future earnings. Recognizing this, a malicious consumer may purchase the product, post a negative review, and demand a ransom in exchange for removing it. The seller can respond to a negative review by (i) doing nothing, (ii) paying the ransom request (if such a request occurs), or (iii) utilizing the dispute resolution mechanism made available by the host platform. We consider two types of mechanisms:

i. Centralized: The seller reports a review to the platform and requests its removal. The platform examines the evidence and decides whether to remove the review.
ii. Decentralized: The seller removes the review without consulting the platform. If the platform judges that the removal was unjustified, the review is reinstated and the seller incurs a penalty.

In investigating dispute cases, we assume that the platform's judgments may suffer from inefficiencies relating to speed and accuracy. Our equilibrium analysis focuses on how these inefficiencies and the available dispute resolution mechanism affect the firm's pricing decision, the malicious consumer's ransom request, the firm's course of action in response to blackmail, and the market's resulting belief about the product's quality. The main qualitative insights extracted from our analysis are summarized as follows.

First, in the absence of a dispute resolution mechanism, we find that the presence of malicious consumers in the market can indeed have a significant impact on the seller's profit. Apart from the ransom payouts that occur as part of successful blackmail attempts, we show that the presence of malicious consumers in the market can also result in upwards distortions in equilibrium prices, which reduce the firm's future profit by restricting the production of genuine product reviews. Overall, we observe that the combined impact of these two effects is most pronounced in settings

[4] In a case on record in the Chinese judiciary system, a consumer was able to successfully extort a laptop seller for an amount five times the value of the product (see Guangdong Shenzhen Longhua District Court Criminal Decision 2018, Yue 0309 Xing Chu 862).

where there is significant uncertainty about product quality, and when the prevalence of malicious consumer behavior is relatively low.

Next, with respect to the relative merits of the two types of mechanisms for dispute resolution, our analysis highlights the following:

i. The centralized dispute resolution mechanism can serve as a credible course of action for the seller, either discouraging the malicious consumer from attempting to blackmail the seller or forcing him to lower his ransom demand. However, we find that the effectiveness of this mechanism can be severely limited by the inefficiencies in the platform's investigation process. In particular, we observe that for desirable outcomes to be achieved, the platform's judgments must be both very quick and highly accurate (two objectives which in practice are often at odds). When this is not the case, the mechanism may be taking up platform resources while offering no advantage to the seller; in fact, at intermediate levels of efficiency, we show that the centralized mechanism not only offers no advantage to the seller, but may even place him at a further disadvantage, allowing the malicious consumer to extract a higher ransom. Moreover, even in those cases where the mechanism is highly efficient, the fraction of the firm's profit loss which is recovered by the mechanism can be underwhelming.

ii. The decentralized approach to dispute resolution has the potential to perform much better than the centralized mechanism, while at the same time significantly reducing the need for platform resources. However, our analysis cautions that for this potential to be fulfilled, the penalty for wrongdoing associated with the mechanism must be chosen wisely: a penalty too low results in abuse of the mechanism by the seller (who may then use the mechanism to remove all negative reviews, both genuine and fake), while a penalty too high may deter the seller from using it, given that the platform's own judgments are subject to errors. In contrast, when the penalty is set at an appropriate intermediate level, we find that the mechanism can be quite effective in recovering the seller's profit. Importantly, because the decentralized approach allows the seller to remove fake reviews immediately (thus neutralizing their impact on future profit), the effectiveness of the mechanism relies predominantly on the platform's judgment accuracy, rather than on whether judgments are made in a timely manner.

Although our model focuses on the specific context of review blackmail (which was the motivation for the introduction of Taobao's decentralized mechanism), it is worth noting that the qualitative nature of our results suggest that decentralization may be an efficient approach to resolving a much broader range of disputes in online platforms. For platform participants, decentralization can speed up the adjudication process significantly, thus alleviating inefficiencies associated with potential delays in the platform's centralized investigation; for the platform, the responsibility to adjudicate disputes can be delegated to the participants without introducing undesirable behavior from the

participants, provided the penalties associated with the mechanism are chosen appropriately. In addition, decentralization can free up significant platform resources, allowing the platform to focus its efforts on the accuracy of its ex post reviews without the need to arrive at quick judgments. Finally, the shift to decentralized dispute resolution may also have implications for longer-term operational decisions, such as personnel hiring (e.g., smaller groups of more specialized investigators versus larger groups of nonspecialists).

The rest of this paper is organized as follows. In §2 we discuss existing literature relating to this work. In §3 we describe our model. In §4 we analyze the implications of the centralized dispute resolution mechanism (which also includes the absence of any mechanism as a special case) and in §5 we consider whether and how the addition of the decentralized mechanism can be advantageous. We conclude in §6.

2. Literature Review

This paper contributes to the literature on two-sided platform governance (see Parker et al. 2016). Bakos and Dellarocas (2011) compare online reputation and the traditional litigation-like mechanism for dispute resolution and show that the latter is more efficient in inducing seller effort in a variety of settings. Bolton et al. (2018) conduct experiments to examine the feedback withdrawal option adopted by some online markets and find that this option can be gamed, producing an escalation of conflict. Using a proprietary dataset, Yang and Zhang (2019) empirically assess the effectiveness of crowdsourcing (i.e., using buyers and sellers of a two-sided marketplace as jurors to resolve disputes between platform participants) as a dispute resolution mechanism. This paper adds to this literature by considering another form of decentralized platform governance, where one of the parties involved (i.e., the seller) is allowed to preemptively settle the dispute, subject to ex post review and potential penalties for wrongdoing.

We study this mechanism in the context of review fraud. The prevalence of fraudulent review practices, whereby sellers create or procure fake reviews for themselves or their competitors, has been empirically documented in numerous studies (Mayzlin et al. 2014, Luca and Zervas 2016, Lappas et al. 2016). More relevant to our work is the paper by Zhang et al. (2020), which provides an empirical analysis of the review blackmail phenomenon considered in this paper. Review fraud has motivated various technological and managerial interventions, such as using algorithms to identify abnormal review patterns (e.g., Mukherjee et al. 2012) and limiting reviews to verified buyers (Mayzlin et al. 2014, Lappas et al. 2016). In this paper, we analyze the use of platform mechanisms for dispute resolution, which rely on sellers either reporting review fraud to the platform or proactively removing fraudulent reviews.

This work also contributes to a growing body of work that focuses on the operational implications of misinformation in online platforms and marketplaces. Chen and Papanastasiou (2019) consider how a monopolist firm's ability to fake purchase transactions affects product pricing and social learning outcomes, while Jin et al. (2019) analyze the impact of "sales brushing" (i.e., sales inflation) on the usefulness of product ranking algorithms. Papanastasiou (2020) analyze optimal fact-checking policies for a social media platform dealing with the circulation of fake news. Mayzlin (2006) and Dellarocas (2006) study competing firms' attempts to manipulate online opinion by publishing fake reviews and recommendations. Our work adds a new dimension to this literature, by considering the implications of misinformation generated on the consumer side (with the goal of extorting the seller), as opposed to on the firm side (with the goal of manipulating consumer beliefs).

Finally, at a higher level, this paper adds to the growing literature studying the operations of two-sided platforms and marketplaces. In a crowdfunding context, Zhang et al. (2017) study how the dynamics of the pledging process affect the optimal pledging level and campaign duration; Chakraborty and Swinney (2020) consider whether entrepreneurs can signal the quality of their product through their choice of crowdfunding campaign parameters; and Babich et al. (2020) study how crowdfunding interacts with more traditional financing sources such as venture capital and bank financing. Feldman et al. (2019b) consider whether food-delivery platforms benefit restaurants. Kanoria and Saban (2017) show that search inefficiencies in matching markets can be alleviated by placing restrictions on agents' actions. Papanastasiou et al. (2018) and Bimpikis et al. (2020) analyze how platforms can filter/repackage the presentation of reviews so as to achieve desirable consumer and supplier behavior, respectively.

3. Model Description

Firm and Consumers. We consider a firm selling an experiential product or service through an online platform (e.g., Amazon, Taobao, TripAdvisor). The product's price is denoted by p and the per unit production cost is normalized to zero. The product's quality can be low or high, $q \in \{l, h\}$. If the quality is low ($q = l$), the gross utility derived by a consumer who purchases the product is zero. If the quality is high ($q = h$), the gross utility derived by a consumer of type j is $v_j > 0$ with probability $\theta \in (0, 1)$ and zero otherwise, where v_j is a random variable with cumulative distribution function $F(\cdot)$; let $\bar{F}(\cdot) = 1 - F(\cdot)$.[5] For ease of exposition, we assume throughout that $F(\cdot)$ is a standard uniform cdf. The product's quality is unobservable before purchase, and the prior belief

[5] For example, the probability θ can be related to uncertainty in the product's manufacturing and/or delivery process, while the heterogeneity in valuations v_i can be attributed to the consumers' idiosyncratic preferences.

that the product's quality is high is denoted by $a \in (0,1)$.[6] Following a purchase decision, consumer j posts a publicly-observable review $R \in \{N, P\}$ consisting of his post-purchase experience, where $R = N$ denotes a negative (i.e., zero-valued) experience, and $R = P$ a positive (i.e., v_j-valued) experience. If the consumer chooses not to purchase, then no review is generated; the absence of a review is denoted by $R = 0$. The review signal $R \in \{N, 0, P\}$ is used to update the market's belief about the product's quality from a to a', via Bayes' rule.

To analyze the effects of different dispute resolution mechanisms, we focus on the interactions between the seller and a single consumer, which we now describe. With probability $\beta \in (0,1)$, the consumer is "malicious." If the consumer is malicious, he purchases the product if and only if it is profitable to do so by blackmailing the seller. The blackmail process is modeled as follows: first, the malicious consumer purchases the product and posts a negative review; next, the consumer contacts the seller and demands a ransom $r > 0$ in exchange for removing his review; the seller then chooses whether to (i) accept and pay the ransom, (ii) refuse the ransom request and do nothing, or (iii) refuse the request and report the fake review to the platform via a dispute resolution mechanism (see next section for details).[7] If the consumer is non-malicious (i.e., a regular consumer), he purchases the product if and only if his expected utility from purchase $u_j = a\theta v_j - p$ is nonnegative, and subsequently posts a (truthful) review depending on his experience.

Apart from the payoff associated with his interaction with the consumer described above, after this interaction the seller also extracts a payoff $\pi(a')$, which captures the firm's future payoffs as a function of the market's posterior belief a'. We assume that $\pi(\cdot)$ is nonnegative and strictly increasing on the unit interval $[0,1]$ (that is, a higher market posterior belief following the seller's interaction with the consumer results in higher future profits).[8]

Dispute Resolution. When facing a blackmail attempt, the seller can (i) agree to pay the ransom to have the negative review removed by the malicious consumer, (ii) refuse to pay the ransom and allow the negative review to remain posted, or (iii) refuse to pay the ransom and make use of the dispute resolution mechanism provided by the platform. Motivated by the practical observations discussed in the introduction, we consider the following two types of mechanisms:

[6] As is common in the literature on experience goods, we assume that the seller and the consumer are symmetrically informed about the product's quality (e.g., Feldman et al. 2019a, Papanastasiou and Savva 2017, Yu et al. 2016)

[7] We implicitly assume that the seller will utilize the platform's internal dispute resolution mechanism rather than the public courts. In the vast majority of cases involving review blackmail, the ransom demanded is low relative to the potential costs of taking the case to court.

[8] While not necessary for our analytical results, in our numerical experiments we will choose $\pi(\cdot)$ to be convex, which ensures that market learning is beneficial for the seller in expectation.

(a) Centralized Dispute Resolution ("C"). The seller reports the blackmail attempt to the platform and requests that the negative review be removed. In doing so, the seller incurs a hassle cost $c \geq 0$ for using the mechanism (e.g., for collecting evidence and making the claim).

(b) Decentralized Dispute Resolution ("D"). The platform allows the seller to remove the negative review without reporting to the platform. However, if the firm removes a review which is then deemed by the platform to be non-malicious, the review is reinstated and the firm incurs a penalty $b > 0$.[9]

Note that while the seller always knows whether a negative review was posted by a malicious consumer, he has no way of conveying this information to the platform efficiently, beyond presenting whatever evidence he can collect from the interaction with the consumer, and waiting for the platform to investigate.

We assume that the platform's investigation of disputes suffers from two forms of inefficiency. First, the dispute is investigated and judged immediately following the seller-consumer interaction with probability $\gamma \in [0, 1]$. This may reflect the possibility of the case being either overlooked entirely or incurring significant delays, for example, owing to the platform's total case load at the time of a reported incident; throughout our analysis, we refer to γ as the "timeliness" parameter. Second, if the case involves a malicious review, the review is correctly identified by the platform as malicious with probability $\delta \in [0, 1]$, for example, owing to the level of evidence required by the platform and/or to the skillfulness of the malicious consumer in conducting the ransom request (for simplicity, we assume that a truthful negative review is never misjudged as malicious); δ is referred to throughout as the "accuracy" parameter.

It is useful to note that while parameters γ and δ are treated as independent in our analysis, in practice the two are often inversely related. For instance, ensuring higher accuracy in judging the merit of a claim often involves conducting a lengthier investigation. Alternatively, parameters γ and δ can also be considered in terms of costly platform resources. For example, ensuring that claims are investigated in a more timely fashion might involve hiring a larger number of investigators, while ensuring that investigation outcomes are more accurate may involve hiring more highly skilled investigators.

Equilibrium. The seller and the consumer are risk neutral and make decisions to maximize their expected profit and utility, respectively. The game proceeds in the following steps.

1. The seller chooses the product's price p.
2. The consumer arrives and his type is realized.

[9] Assuming that the seller also incurs a hassle cost for using the decentralized mechanism has no qualitative bearing on our results (see also Figure 5 in §5).

(i) If the consumer is malicious, he observes the price and decides whether to purchase. Following a purchase decision, he posts a fake negative review and chooses a ransom r to be demanded from the seller in exchange for removing the review.

(ii) If the consumer is non-malicious, he observes the price and decides whether to purchase. Following a purchase decision, he posts a truthful review according to his experience with the product.

3. The seller observes the posted review. If there is a ransom request, the firm chooses whether to accept the request, reject it, or utilize the available dispute resolution mechanism. If there is no ransom request, the seller may still choose to utilize the mechanism.
4. If the seller has utilized the mechanism, the platform's investigation occurs (in accordance with the timeliness parameter γ), and the investigation outcome is realized (in accordance with the accuracy parameter δ).
5. The market observes the posted review $R \in \{N, 0, P\}$ and the posterior belief a' is updated via Bayes' rule.

Throughout our analysis, we focus on perfect Bayesian equilibria (PBE) in pure strategies. Note that the interaction between the seller and the consumer which precedes the generated review is not observable to the market. Thus, a PBE in our model requires that the market's posterior belief about the product's quality is consistent with the seller's and the consumer's equilibrium strategies, and that the seller's and the consumer's strategies are optimal given the market's posterior belief.

4. Centralized Dispute Resolution

In this section, we analyze the properties of the centralized dispute resolution mechanism ("C"). Under the centralized mechanism, the seller reports a negative review to the platform, and the platform investigates the claim and decides whether the review should be removed. Recall that the mechanism may exhibit inefficiencies relating to timeliness $\gamma \in [0, 1]$ and accuracy $\delta \in [0, 1]$. Note that in the special case with $\gamma = \delta = 0$, the model reduces to one where no dispute resolution mechanism is available to the seller.

We begin with a straightforward result that follows trivially from our assumption that the platform never misjudges a truthful review as fake.

LEMMA 1. *Under the centralized mechanism, in equilibrium, the seller never disputes a nonmalicious review.*

All proofs are relegated to the Appendix. The seller has nothing to gain from reporting a genuine review, while doing so incurs the hassle cost $c \geq 0$. Accordingly, in the analysis that follows it will suffice to focus on the seller's response when he encounters a malicious consumer.

We solve the game between the seller and the consumer via backwards induction, starting from the last step where the market's posterior belief is formed according to the observed review and the equilibrium strategies of the firm and the consumer.

4.1. Market's Posterior Belief

The market's posterior belief a' determines the seller's terminal payoff $\pi(a')$ and is formed according to the review observation $R \in \{N, 0, P\}$ and the seller's and the consumer's equilibrium strategies. Given that the seller never reports a nonmalicious consumer's review (Lemma 1), it will suffice to characterize the posterior belief as a function of how the seller chooses to deal with a malicious review. There are three possible scenarios: (i) s: the seller *settles* with the malicious customer (i.e., pays the ransom); (ii) c: the seller reports the malicious customer to the *centralized* mechanism; (iii) n the seller does *nothing* and allows the negative review to stay posted. Let

$$a_R^i = P(q = h \mid i, R)$$

denote the posterior belief when the seller's response to malicious reviews is $i \in \{s, c, n\}$ and the observed review is $R \in \{N, 0, P\}$.

LEMMA 2. *The posterior belief a_R^i satisfies $a_P^i = 1$, $a_0^i = a$, and $a_N^i \in [0, a]$, where*

$$\begin{aligned}
a_N^n &= \frac{a}{a + (1-a)\frac{\left(\beta + (1-\beta)(1-\frac{p}{a\theta})\right)}{\left(\beta + (1-\beta)(1-\frac{p}{a\theta})(1-\theta)\right)}}, \\
a_N^c &= \frac{a}{a + (1-a)\frac{(\beta(1-\gamma\delta) + (1-\beta)(1-\frac{p}{a\theta}))}{(\beta(1-\gamma\delta) + (1-\beta)(1-\frac{p}{a\theta})(1-\theta))}}, \\
a_N^s &= \frac{a}{a + (1-a)\frac{1}{1-\theta}}.
\end{aligned}$$

Note first that irrespective of the firm's approach to dealing with a malicious customer, if a positive review is observed, then the posterior belief that the product is of high quality is one (i.e., $a_P^i = 1$). To see this, note that a positive review in our model can only have been generated for a high-quality product, which generates a positive experience with probability β (by contrast, a low-quality product never generates a positive experience). Next, if no review is observed, the posterior belief remains equal to the prior (i.e., $a_0^i = a$). The absence of a review indicates that either a non-malicious consumer has chosen not to purchase the product, or a malicious consumer has chosen to purchase and his review has been subsequently removed (either through a successful extortion attempt, or through the centralized dispute resolution mechanism); in either of the two scenarios, the absence of a review carries no information about the product's quality, so that the posterior belief stays equal to the prior.

Now consider the posterior belief following a negative review (i.e., a_N^i). In this case, the posterior depends on the firm's approach to dealing with malicious reviews. Note that from the expressions of Lemma 2, it follows that $a_N^s < a_N^c < a_N^n$. In particular, if the firm chooses to settle with the malicious consumer (i.e., $i = s$), then a negative review can only have been generated from a regular consumer who had a bad experience with the product—in this scenario, a negative review contains significant information about the product's quality and therefore carries significant weight in the belief update. By contrast, if the firm chooses to ignore the malicious consumer's ransom request (i.e., $i = n$), then a negative review may have been generated by a regular consumer or by a malicious consumer who failed in his extortion attempt—here, the information contained in a negative review is questionable, so that the review does not significantly impact the posterior belief. Finally, if the firm chooses to report the malicious review to the centralized mechanism (i.e., $i = c$), then a negative review may have been generated by a regular consumer or by a malicious consumer whose review the centralized mechanism failed to remove from the system—in this case, the informational content of the review lies between the two aforementioned extremes, as the malicious consumer's review remains in the system with some positive probability less than one.

Recall that the seller's terminal payoff $\pi(\cdot)$ is an increasing function of the posterior belief. Thus, Lemma 2 provides a preview of the potential equilibrium scenarios. When facing a malicious consumer, if the seller chooses to settle, he is able to remove the negative review, thus shifting the posterior to a_0^s, but at the cost of the ransom r. If he chooses to do nothing, the negative review remains in the system, and the posterior belief takes the higher value a_N^n. The centralized dispute resolution mechanism provides a third option for dealing with the malicious review, which comes at a cost c, but whose outcome in terms of the market's posterior belief is uncertain.

4.2. Seller's Response to Blackmail

Given that the seller's approach to dealing with malicious consumers is unobservable to the market, for an equilibrium to be established we require that the seller's approach is optimal given the market's belief about his approach.

To illustrate, suppose that the market's belief is that the seller does nothing in response to ransom requests (i.e., $i = n$). To establish the conditions under which $i = n$ is indeed an equilibrium strategy, we consider whether the seller has a profitable deviation. If the seller adopts approach $i = n$, his payoff gain is $\pi(a_N^n)$. If, instead, he deviates to strategy $i = s$ (while the market believes his approach to be n), his payoff gain is $\pi(a_0^n) - r = \pi(a) - r$, where r is the (equilibrium) ransom request. The difference between the two is then

$$\Delta^{s|n} = \pi(a) - \pi(a_N^n) - r$$

Similarly, if the seller deviates to strategy $i = c$, the difference in payoff gains is

$$\begin{aligned}\Delta^{c|n} &= [\gamma\delta\pi(a_0^n) + (1-\gamma\delta)\pi(a_N^n)] - \pi(a_N^n) - c, \\ &= \gamma\delta\,[\pi(a) - \pi(a_N^n)] - c.\end{aligned}$$

Then, for $i = n$ to be an equilibrium strategy, we require that both $\Delta^{s|n}$ and $\Delta^{c|n}$ are nonpositive. This occurs when

$$\begin{aligned}\Delta^{s|n} \le 0 &\iff \pi(a_N^n) \ge \pi(a) - r, \text{ and} \\ \Delta^{c|n} \le 0 &\iff \pi(a_N^n) \ge \pi(a) - \frac{c}{\gamma\delta}\end{aligned}$$

or, equivalently, when $\pi(a_N^n) \ge \pi(a) - \min\{r, \frac{c}{\gamma\delta}\}$. Note that a_N^n takes values in the interval $[0, a]$, which implies that an equilibrium with $i = n$ does exist for some combinations of our model parameters. A similar process establishes the conditions for the existence of equilibria involving seller strategies $i = s$ and $i = c$. In particular,

PROPOSITION 1. *Suppose a malicious consumer enters the system, posts a negative review, and demands a ransom r. Then:*

(i) An equilibrium with $i = n$ exists if and only if $\pi(a_N^n) \ge \max\{\pi(a) - r, \pi(a) - \frac{c}{\gamma\delta}\}$.

(ii) An equilibrium with $i = s$ exists if and only if $\pi(a_N^s) \le \min\{\pi(a) - r, \pi(a) + \frac{c-r}{1-\gamma\delta}\}$.

(iii) An equilibrium with $i = c$ exists if and only if $\pi(a) + \frac{c-r}{1-\gamma\delta} \le \pi(a_N^c) \le \pi(a) - \frac{c}{\gamma\delta}$.

That is, settling with the malicious consumer (i.e., strategy $i = s$) is an equilibrium provided the ransom request r is sufficiently small, while doing nothing in response to the blackmail attempt (i.e., strategy $i = n$) is an equilibrium when the ransom request is high and the overall efficiency of the centralized mechanism, captured by the product $\gamma\delta$, is low. An equilibrium at strategy $i = c$, which utilizes the centralized mechanism, exists when the mechanism efficiency is high and the ransom request is not too low. We note that Proposition 1 admits the possibility of parameter combinations where more than one equilibria in seller strategies exist. Whenever this is the case, we assume that the equilibrium which maximizes the seller's expected payoff prevails. We next analyze the malicious consumer's purchase-and-blackmail strategy.

4.3. Malicious Consumer's Strategy

The malicious consumer is interested in purchasing the product only in order to profit by blackmailing the seller. Therefore, the malicious consumer in our model purchases if and only if there exists a ransom r which (i) the seller is willing to accept as part of a settlement to have the fake review removed, and also (ii) satisfies $r > p$ yielding positive surplus for the consumer.

PROPOSITION 2. *The malicious consumer's equilibrium strategy is described as follows:*

(i) When $c \geq \gamma\delta\left(\pi(a) - \pi(a_N^n)\right)$, the malicious consumer purchases if and only if $p < \pi(a) - \pi(a_N^n)$. He then posts a negative review and demands a ransom

$$r^* = \pi(a) - \pi(a_N^n). \tag{1}$$

(ii) When $c < \gamma\delta\left(\pi(a) - \pi(a_N^n)\right)$, the malicious consumer purchases if and only if $p < (1 - \gamma\delta)[\pi(a) - \pi(a_N^c)] + c$. He then posts a negative review and demands a ransom

$$r^* = (1 - \gamma\delta)[\pi(a) - \pi(a_N^c)] + c. \tag{2}$$

The first part of the proposition refers to cases where the overall efficiency of the dispute resolution mechanism is relatively low (or, equivalently, the cost of using the mechanism is relatively high). Observe that in these cases, the malicious consumer's purchase-and-blackmail strategy is independent of the mechanism parameters c, γ and δ. That is, the presence of the mechanism has no impact on the seller-consumer interaction. Instead, the malicious consumer estimates the difference in future earnings for the seller between allowing the negative review to stay in the system and having it removed—this is the maximum ransom the malicious consumer can extract from the seller, $r^* = \pi(a) - \pi(a_N^n)$. Having identified the maximum ransom he can extract, the consumer purchases if and only if the product's price is sufficiently low to allow positive surplus (i.e., $p < r^*$).

The second part of the proposition addresses cases where the efficiency of the mechanism is relatively high. In these cases, the presence of the mechanism places a limit on the malicious consumer's ability to extract ransom from the seller. In particular, the malicious consumer recognizes that if his ransom request is too high, the mechanism provides a credible course of action for the seller. To avoid this, the consumer sets the ransom at a level $r^* = (1 - \gamma\delta)[\pi(a) - \pi(a_N^c)] + c$, which accounts for the efficiency of the mechanism as well as the cost to the seller of utilizing the mechanism. As in the previous case, he then purchases if and only if the price is sufficiently low for the transaction to be profitable.

It is worth noting that in both parts of Proposition 2, the equilibrium ransom r^* is decreasing in the proportion of malicious consumers β. To see why this occurs, observe that according to Lemma 2, the posterior beliefs a_N^n and a_N^c both approach a as β increases, because in both scenarios the market interprets a negative review as more likely to have been generated by a malicious consumer, so that the detrimental effect of a negative review is reduced. However, we note that this does not imply that in equilibrium the seller pays a lower ransom (in expectation) as β increases; on the contrary, it is straightforward to show that the expected ransom βr^* is increasing in β.

A closer look at Proposition 2 also reveals the following interesting phenomenon.

COROLLARY 1. *Suppose $p < (1-\gamma\delta)[\pi(a) - \pi(a_N^n)]$ (i.e., in equilibrium, the malicious consumer chooses to purchase). The equilibrium ransom r^* is not monotonically decreasing in the mechanism efficiency $\gamma\delta$.*

One might conjecture that as the centralized mechanism becomes more efficient, the seller might be better equipped to deal with the malicious consumer. However, Corollary 1 establishes that this is not the case. Instead, the equilibrium ransom is constant up to a threshold value of $\gamma\delta$ and is monotonically decreasing above that. More interestingly, the malicious consumer's ability to extract ransom from the seller is maximized at some intermediate of mechanism efficiency (i.e., the equilibrium ransom exhibits a positive "jump"). The implication of this result is that the presence of the centralized dispute resolution mechanism can in fact be detrimental for the seller, allowing the malicious consumer to leverage the availability of the mechanism to improve his "bargaining" position.

The result is illustrated in Figure 1. The key driver of the observed structure is the impact of the mechanism's presence on the market's expectation of how the seller deals with malicious consumers, which in turn manifests through the posterior beliefs a_N^i, for $i \in \{s, c, n\}$. When the efficiency of the mechanism is low, the market expects that the seller will either settle with the malicious consumer, or do nothing in response to the ransom request. By contrast, when the efficiency is relatively high, the market expects the seller to either settle or use the mechanism. At the same time, the malicious consumer sets the ransom accordingly, demanding $r^* = (1-\gamma\delta)[\pi(a) - \pi(a_N^c)] + c$ when $\gamma\delta > \frac{c}{\pi(a)-\pi(a_N^n)}$, and $\pi(a) - \pi(a_N^n)$ otherwise. Recalling that by Lemma 2, we have $\pi(a_N^n) > \pi(a_N^c)$, it can then be deduced that the malicious consumer's ransom request is at its highest when the mechanism efficiency $\gamma\delta$ lies just above the threshold $\frac{c}{\pi(a)-\pi(a_N^n)}$.

4.4. Seller's Pricing Decision

We consider next the seller's pricing problem. Building on the analysis of §4.3, let $\mathcal{P}_{pur}$ be the set of prices at which the malicious consumer chooses to purchase, for a given set of model parameters. The seller's payoff function in the presence of the centralized mechanism can be expressed as

$$\Pi_C(p) = \beta[\pi(a) - \mathbb{1}_{p\in\mathcal{P}_{pur}}(r^*(p) - p)] + (1-\beta)[p\pi(a) + (1-p)[p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)], \quad (3)$$

where $r^*(p)$ is given in Proposition 2. The first term captures the seller's expected profit in the event that the consumer is malicious. In particular, if the seller chooses a price $p \notin \mathcal{P}_{pur}$, then the malicious consumer does not purchase and the seller's payoff is $\pi(a)$, since no review signal is generated. On the other hand, if the seller chooses a price $p \in \mathcal{P}_{pur}$, then the malicious consumer purchases, and the seller agrees to pay the ransom $r^*(p) \geq p$ to have the malicious review removed. The second term is the seller's expected payoff in the event that the consumer is nonmalicious. In

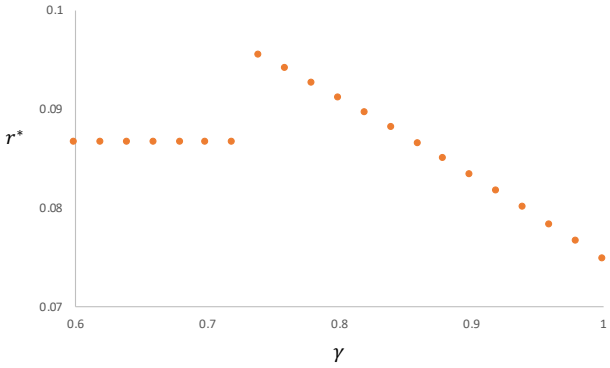

Figure 1 **Equilibrium ransom as a function of the centralized mechanism efficiency. Parameter values:** $a = \theta = 0.5$**,** $\beta = 0.1$**,** $\delta = 0.8$**,** $p = 0.08$**,** $c = 0.05$**.**

this case, the seller's payoff depends on whether the consumer chooses to purchases and, if so, the review she generates after her experience with the product.

Observe that by Lemma 2, the continuation payoff $\pi(a_N^s)$ is independent of the seller's chosen price, which in turn implies that the seller's payoff function $\Pi_C(p)$ is concave for $p \notin \mathcal{P}_{pur}$. Therefore, the seller's problem may be viewed as choosing a price to maximize a concave function, less a penalty equal to $r^*(p) - p$ which applies whenever a price $p \in \mathcal{P}_{pur}$ is chosen. Let p_0 be the unique maximizer of the seller's payoff function ignoring the penalty, that is,

$$p_0 := \arg\max_{p \in [0,1]} \quad \beta\pi(a) + (1-\beta)\left(p\pi(a) + (1-p)[p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)]\right).$$

It follows that if $p_0 \notin \mathcal{P}_{pur}$, then $p^* = p_0$; that is, if the optimal price ignoring the penalty does not belong to the set of prices that induces the malicious consumer to purchase, then this price is globally optimal. While the properties of the set $\mathcal{P}_{pur}$ depend on the functional form of the continuation payoff $\pi(\cdot)$, in general the malicious consumer tends to purchase when the price is relatively low (see also Proposition 2). Accordingly, let us define

$$\bar{p} := \max \mathcal{P}_{pur}. \tag{4}$$

We then have from the above discussion,

PROPOSITION 3. *The following statements hold:*

1. *If* $p_0 \geq \bar{p}$*, then* $p^* = p_0$*.*
2. *If* $p_0 < \bar{p}$*, then* $p^* \in (0, \bar{p}]$*.*

In cases where the seller would prefer to set a relatively high price in the absence of malicious consumers (i.e., $p_0 > \bar{p}$), the presence of such consumers does not affect his pricing decision. By

contrast, when the seller would prefer to set a relatively low price in the absence of malicious consumers, the presence of such consumers introduces a tradeoff for the seller. To explain this tradeoff, we enlist the example of Figure 2. We note first that in this example, we have $p_0 = 0.17$. Observe that the equilibrium price satisfies $p^* = p_0$ when the probability of a malicious consumer β is either zero or very high. In all other cases, the optimal price is strictly higher than p_0; that is, the presence of the malicious consumer causes an upwards price distortion, which is particularly pronounced when β is low-to-intermediate. We now discuss the figure in more detail.

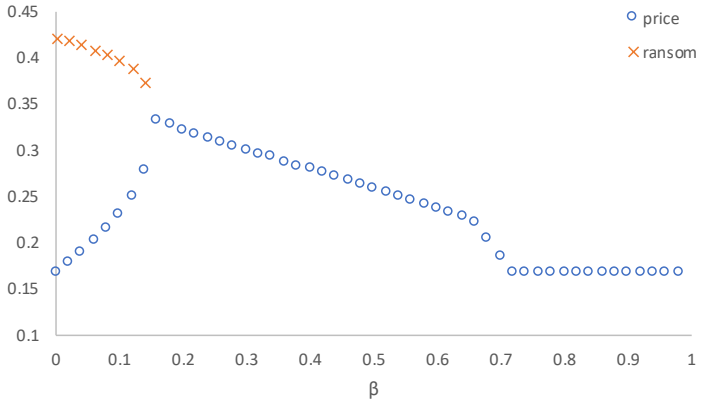

Figure 2 **Equilibrium price and ransom as a function of the probability that the customer is malicious. Parameter values:** $a = 0.5$, $\theta = \gamma = \delta = 0.8$, $c = 0.1$, $\pi(a) = 4a^2$.

Consider first the region $\beta \in [0, 0.15]$, and note that in this region there is a positive ransom request, which implies that in the equilibrium induced by the seller's pricing decision, the malicious consumer chooses to purchase the product. When β is very low, the seller anticipates that he may encounter a malicious consumer, but knows that the probability of this occurring is small. Therefore, the seller largely ignores the presence of malicious consumers in the market, and opts for a price that is close to p_0. At the same time, observe that even though the probability of encountering a malicious consumer is small, whenever such an encounter does occur, the seller is forced to pay a heavy ransom whose magnitude can reach up to two-and-a-half times the product's price in this example. As the probability β increases, the seller adjusts the price upwards so as to reduce the damage from encounters with malicious consumers, recognizing that this scenario now occurs with a higher, albeit still low, probability.

Next, consider the region $\beta \in [0.15, 0.7]$. In this region, the probability of encountering a malicious consumer is sufficiently high so that the seller prefers to avoid having to deal with the malicious

consumer. To do so, the seller must set the price sufficiently high so that the malicious consumer is deterred from purchasing the product, realizing that the seller would prefer to use the platform's dispute resolution mechanism rather than pay a ransom which is high enough to be profitable for the malicious consumer. Therefore, the optimal price in this region is the lowest price at which the malicious consumer is deterred from purchasing. This results in an upwards price distortion, because the latter price is higher than p_0. We point out that this distortion in the seller's pricing decision, while eliminating the threat from malicious consumers, can cause a significant loss in seller profit (see also §4.5 where we quantify this loss), for two reasons. First, the higher price renders a nonmalicious consumer less likely to purchase, which causes an immediate loss in profit. Second, and more importantly, the nonmalicious consumer's lower likelihood of purchase translates into a lower likelihood of a review being generated, which causes a loss in (expected) future profit.

Moving to the higher region of $\beta \in [0.7, 1]$, here the seller knows that there is a high chance of encountering a malicious consumer. However, this turns out to be irrelevant—as a result of the equilibrium belief structure described in Lemma 2, when β is sufficiently high, the price threshold $\bar{p}$ above which the malicious consumer chooses not to purchase falls below p_0, and the first part of Proposition 3 applies with $p^* = p_0$.

4.5. Profit Implications

The preceding sections describe the seller's and the consumer's equilibrium strategies, as well as the market's equilibrium beliefs following the seller-consumer interaction. In this section, we investigate the effectiveness of the centralized dispute resolution mechanism in mitigating the detrimental impact of malicious consumer behavior.

To do so, it is instructive to first evaluate the impact of the malicious consumer's presence on the seller's profit in the absence of a dispute resolution mechanism, at different values of our model parameters. We use Π^*_{no} to denote the firm's optimal profit in the absence of a mechanism and Π^*_{opt} to denote optimal profit in the presence of a perfectly efficient mechanism (we note that the absence of a mechanism can be retrieved from the preceding analysis by setting $\gamma = \delta = 0$, while a perfectly efficient mechanism can be retrieved by setting $\gamma = \delta = 1$ and $c = 0$). The contour plot of Figure 3 summarizes our observations. In particular, we find that the seller's profit loss is particularly pronounced when (i) the seller's future profit potential is sufficiently high (otherwise, the malicious consumer has little power to conduct blackmail), (ii) the probability that the consumer is malicious is low-to-intermediate (this is where the seller's pricing decision is distorted the most and the probability of a ransom payment is significant), and (iii) there is significant uncertainty regarding the product's quality (so that a negative review is most detrimental for the seller's future profit). With the observations of Figure 3 at hand, we focus the rest of our experiments on the parameter regions that are the most problematic in terms of profit loss for the firm.

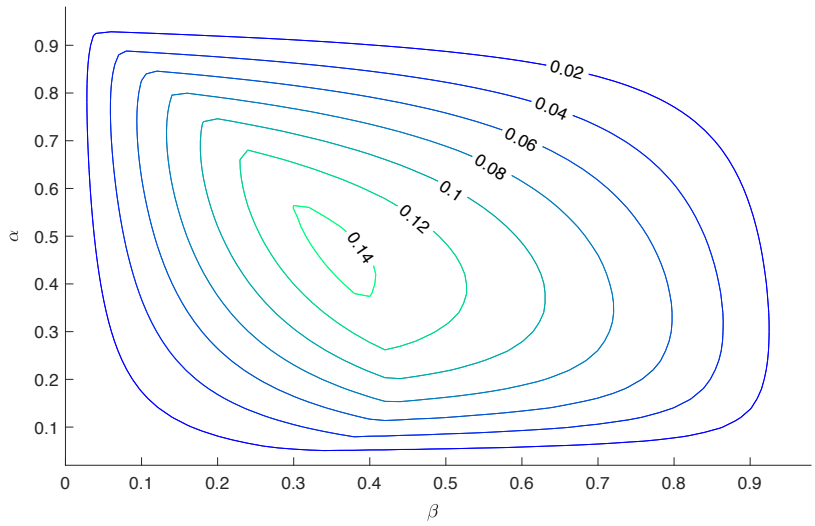

Figure 3 **Efficiency loss in the absence of a dispute resolution mechanism,** $1 - \Pi^*_{no}/\Pi^*_{opt}$**. Parameter values:** $\theta = 0.9$**,** $\pi(a) = 50a^2$**.**

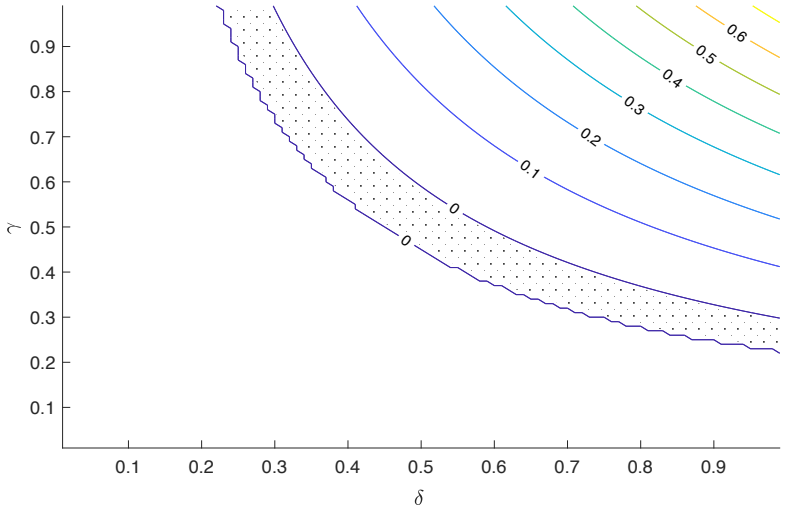

Figure 4 **Efficiency loss recovered by the centralized dispute resolution mechanism,** $(\Pi^*_C - \Pi^*_{no})/(\Pi^*_{opt} - \Pi^*_{no})$**. Parameter values:** $\theta = 0.9$**,** $a = 0.5$**,** $\beta = 0.3$**,** $c = 2$**,** $\pi(a) = 50a^2$**.**

Accordingly, in Figure 4 we evaluate how much of the profit loss incurred by the seller (due to the presence of malicious consumer behavior) can be recovered by the centralized mechanism, at different values of the timeliness γ and accuracy δ of the mechanism. We highlight the following

observations. First, the shape of the contour lines suggest that parameters γ and δ exhibit complementarities in determining the mechanism's effectiveness. Indeed, we note that in the preceding analysis γ and δ feature always as the product $\gamma\delta$, implying that the two are not only complementary, but also interchangeable for the centralized mechanism. Second, observe that for a large region of $\gamma\delta$ combinations, the mechanism is completely ineffective, resulting in a zero increase in seller profit (see the lower-left region of Figure 4). What is more, we note that there exist intermediate values of $\gamma\delta$ (see the shaded region of Figure 4) where the seller is in fact worse off in the presence of the mechanism—this observation is consistent with Corollary 1 (see §4.3), which suggests that a mechanism with intermediate efficiency can hurt the seller, by putting the malicious consumer in a better position to extract ransom. Third, in the region where the mechanism is helpful for the seller, the platform's judgment is required to be both very quick (i.e., high γ) and highly accurate (i.e., high δ) before the mechanism is able to recover a significant portion of the efficiency loss. The latter observation is particularly important given that, in practice, one might expect judgment speed and accuracy to be inversely related (assuming a fixed amount of resources)—indeed, the tradeoff between speed and accuracy in service systems has received significant attention in the existing literature (see Alizamir et al. (2013), Kostami and Rajagopalan (2014), and references therein).

5. Decentralized Dispute Resolution

In this section, we analyze the properties of the decentralized dispute resolution mechanism ("D"). Under this mechanism, the seller can remove a review immediately without consulting the platform; however, if the platform investigates the dispute and finds that the review removal was not warranted, the review is reinstated and a penalty $b \geq 0$ is imposed upon the seller.

The analysis of the decentralized mechanism follows the same qualitative steps as that of the centralized mechanism in §4. However, the general analysis of the decentralized mechanism is significantly more cumbersome, because the seller in this case may choose to apply the decentralized mechanism to remove genuine negative reviews (i.e., in addition to using the mechanism to remove malicious negative reviews), expanding the space of seller strategies to be considered. Thus, before proceeding, it is useful to restrict the scope of our analysis by pointing out that if the penalty b is sufficiently low, the seller will abuse his access to the decentralized mechanism, using the mechanism to remove genuine negative reviews (i.e., even if he knows there is a high chance that the review will be re-posted and the penalty b will be incurred). Therefore, for abuse of the mechanism by the seller to be avoided, the penalty attached to the decentralized mechanism should not be too low. Define $\underline{b} := (1-\gamma)\left[\pi(a) - \pi\left(a_N^s\right)\right]$.

PROPOSITION 4. *Suppose $b \geq \underline{b}$. Then, in equilibrium, the seller does not use the decentralized mechanism to remove nonmalicious reviews.*

In the remainder of this section, we will focus on the more relevant cases of mechanisms satisfying the sufficient condition of Proposition 4.[10]

5.1. Malicious Consumer's Strategy

We pick up the analysis of the decentralized mechanism with the malicious consumer's equilibrium response to a given price p. We note that, as was the case in §4.3, this characterization also encompasses the seller's response to blackmail attempts, to the extent that, in equilibrium, the malicious consumer would only purchase the product and demand a ransom if he anticipates that the seller will accept such a demand.

In addition to the market posterior beliefs a_N^n and a_N^s (see Lemma 2), the result that follows makes use of the belief

$$a_N^d = \frac{a}{a + (1-a)\frac{\left(\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}{\left(\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right)}},$$

which denotes the market's posterior belief conditional on observing a negative review, when the seller uses the decentralized mechanism to deal with malicious consumers. Following a similar logic as for the result of Lemma 2, it can be deduced that $a_N^s < a_N^d < a_N^n$.

PROPOSITION 5. *The malicious consumer's equilibrium strategy is described as follows:*

(i) When $b \geq \frac{1-\gamma(1-\delta)}{1-\delta}\left(\pi(a) - \pi(a_N^n)\right)$, the malicious consumer purchases if and only if $p < \left(\pi(a) - \pi(a_N^n)\right)$. He then posts a negative review and demands a ransom

$$r^* = \left(\pi(a) - \pi(a_N^n)\right). \tag{5}$$

(ii) When $b < \frac{1-\gamma(1-\delta)}{1-\delta}\left(\pi(a) - \pi(a_N^n)\right)$, the malicious consumer purchases if and only if $p < (1-\delta)\left(\gamma\left(\pi(a) - \pi(a_N^d)\right) + b\right)$. He then posts a negative review and demands a ransom

$$r^* = (1-\delta)\left(\gamma\left(\pi(a) - \pi(a_N^d)\right) + b\right). \tag{6}$$

Observe first that when the wrongdoing penalty b is sufficiently high, the mechanism has no impact on the equilibrium interaction between the seller and the malicious consumer. That is, if the seller faces a steep enough penalty when he is judged to have wrongfully removed a review, the decentralized mechanism does not constitute a credible course of action for the seller. Note that

[10] We do so having in mind that, in practice, it would be unlikely for a host platform to set the penalty low enough to allow/incentivize sellers to remove all negative reviews (i.e., genuine in addition to malicious).

the threshold value of the penalty above which the decentralized mechanism becomes irrelevant increases with the accuracy parameter δ, but decreases with the timeliness parameter γ.

The second part of Proposition 5 describes the cases where the decentralized mechanism becomes relevant. Observe that that the price below which the malicious consumer chooses to purchase, as well as the ransom he demands after purchasing, are increasing in the timeliness γ and the mechanism penalty b, and are decreasing in the mechanism accuracy δ.

A direct comparison between Proposition 5 and its counterpart in the case of the centralized mechanism, Proposition 2, reveals the relative merits of the two mechanisms. We note, in particular, the following qualitative differences:

i. Under the centralized mechanism ("C"), an increase in the judgment timeliness γ: (i) renders the mechanism *more* likely to be a credible course of action for the seller, and (ii) results in a *decrease* in the malicious consumer's ransom request (should such a request occur).
ii. Under the decentralized mechanism ("D"), an increase in the judgment timeliness γ: (i) renders the mechanism *less* likely to be a credible course of action for the seller, and (ii) results in an *increase* in the malicious consumer's ransom request.

Given that the high-level structure of Proposition 5 is much like that of Proposition 2 in §4, the seller's pricing problem in the presence of the decentralized mechanism is qualitatively similar to that in the presence of the centralized mechanism. We therefore bypass the analysis of the seller's pricing problem, and consider directly the profit implications of the decentralized mechanism.

5.2. Profit Implications

We now evaluate the effectiveness of the decentralized mechanism in restoring the loss in profit incurred by the seller as a result of malicious consumer behavior. With respect to the implementation of this mechanism, the results of the preceding analysis suggest that (i) when the penalty b is too low, the seller will abuse the decentralized mechanism (i.e., removing negative reviews at will, even when these are nonmalicious), while (ii) when the penalty is too high, the mechanism becomes irrelevant (i.e., the mechanism cannot serve as a credible course of action for the seller). Moreover, we note that at intermediate values of the penalty b, the malicious consumer's ransom request increases with b. These observations suggest that for the mechanism to effective, the penalty b should take an intermediate-to-low value.

In Figure 5, we consider how the decentralized mechanism performs in comparison to the centralized mechanism analyzed in §4. In this experiment, we compare the decentralized mechanism with penalty $b = \underline{b}$ against a centralized mechanism with cost $c = 0$. We use Π_D^* to denote the seller's profit under the decentralized mechanism. Notice that even though the centralized mechanism in this example is costless, the decentralized mechanism dominates, with the exception of cases

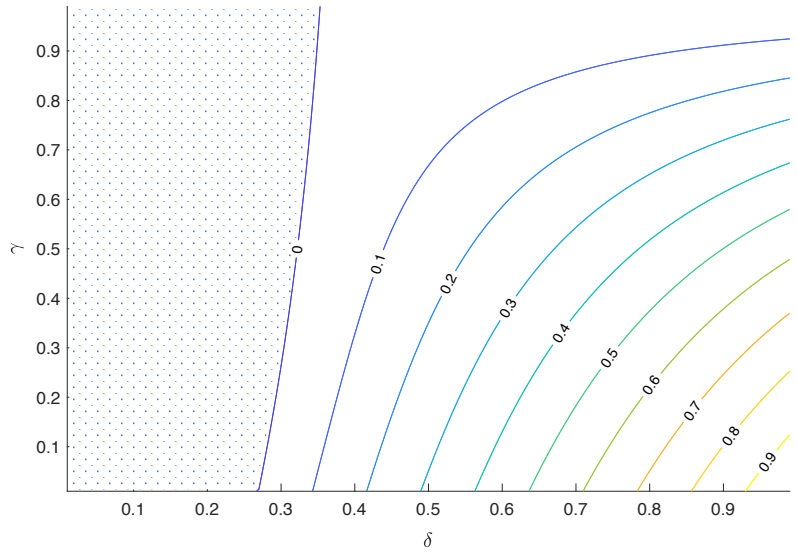

Figure 5 **Profit difference between decentralized and centralized mechanisms,** $(\Pi_D^* - \Pi_C^*)/(\Pi_{opt}^* - \Pi_{no}^*)$**. Parameter values:** $\theta = 0.9$**,** $a = 0.5$**,** $\beta = 0.3$**,** $b = \underline{b}$**,** $c = 0$**,** $\pi(a) = 50a^2$**.**

where the platform's judgment accuracy is very low. Moreover, observe that the dominance of the decentralized mechanism is particularly pronounced when the judgment accuracy δ is high and the timeliness parameter γ is low. When γ is low, the centralized mechanism suffers as a result of the platform's inability to evaluate claims against malicious consumer behavior in a timely fashion; by contrast, the decentralized mechanism allows the firm to take action immediately, thus avoiding profit losses while the platform's investigation is conducted—the difference in performance between the two mechanisms in such cases is large.

While Figure 5 evaluates the performance of the decentralized mechanism relative to the centralized mechanism, Figure 6 evaluates how much of the firm's total profit loss can be recovered by the decentralized mechanism. It is instructive to compare this plot with that of Figure 4, which conducts the same experiment but for the centralized mechanism. Note the difference in the shape of the contours: while an increase in the performance of the centralized mechanism requires a simultaneous increase in both timeliness γ and accuracy δ, the decentralized mechanism requires only an increase in the accuracy δ, and is largely unaffected by changes in γ. The key to this observation is that, as the timeliness γ decreases, the platform can simply increase the penalty b appropriately, so as to ensure that the seller does not abuse the mechanism, while maintaining the mechanism's relevance for the seller-consumer interaction. Furthermore, observe that, provided the judgment accuracy is sufficiently high, the decentralized mechanism is able to recover most of the firm's profit loss. In contrast, for the centralized mechanism to achieve such performance, the

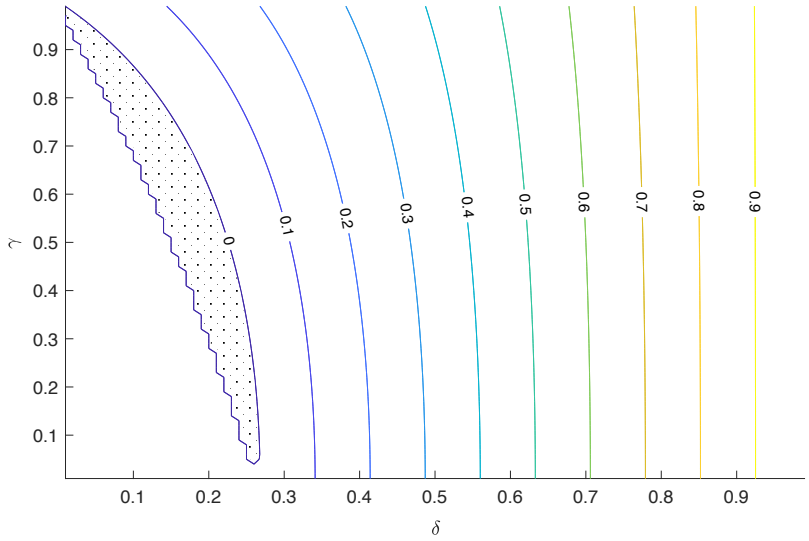

Figure 6 **Fraction of efficiency loss recovered by the decentralized dispute resolution mechanism,** $(\Pi_D^* - \Pi_{no}^*)/(\Pi_{opt}^* - \Pi_{no}^*)$**. Parameter values:** $\theta = 0.9$**,** $a = 0.5$**,** $\beta = 0.3$**,** $b = \bar{b}$**,** $\pi(a) = 50a^2$**.**

platform's judgments must be both very quick and highly accurate, and the hassle cost associated with using the mechanism must be very low.

6. Conclusion

In online marketplaces, customers rely on the reviews of their peers to help them distinguish between products of different quality levels. Recognizing this, malicious consumers may attempt to extort sellers, threatening with a negative review unless the seller agrees to pay a ransom.

In this paper, we develop a stylized model of the interactions between a seller, who takes into account the impact of consumer reviews on his future earnings, and a potentially-malicious consumer. We find that, apart from the direct impact of blackmail attempts, the presence of malicious consumers in the market may also cause upwards price distortions, leading to a significant loss in seller profit which is particularly pronounced when there is significant uncertainty about the product's quality. It is important to note that, while our analysis focuses on the implications of malicious consumer behavior for seller profit, the described price distortions are also detrimental with regards to consumer surplus. In particular, the surplus of nonmalicious consumers is hurt both directly (through the product's higher price), but also indirectly, through the decreased availability of review information, which occurs as a result of consumers' decreased probability of purchase (and therefore lower probability of producing a review).

In response to sellers' growing concerns about review blackmail, platforms such as TripAdvisor

and Taobao have implemented mechanisms for dispute resolution aimed at helping sellers mitigate the detrimental impact of malicious consumer behavior. The traditional versions of these mechanisms are "centralized": the seller reports the blackmail attempt to the platform, which then decides whether the malicious review should be removed. Our analysis of these mechanisms suggests that their effectiveness relies on the platform's ability to process seller claims in a timely and accurate manner, two objectives which are often at odds in practice.

More recently, a more "decentralized" approach to dispute resolution has emerged. Under decentralized dispute resolution, the platform grants sellers the autonomy to remove reviews without consulting with the platform, subject to ex post checks by the platform and penalties in the event that reviews are judged to have been removed unjustifiably. Our analysis suggests that such a mechanism, when implemented correctly, can significantly enhance outcomes while simultaneously reducing the need for platform resources. In particular, we observe that while accuracy remains important for the success of the decentralized mechanism, timeliness can be less of a concern, provided the penalty for wrongdoing is chosen appropriately.

Although the analysis of this paper focuses on disputes involving review blackmail, the qualitative nature of our results suggests that decentralization may be beneficial in a broader range of disputes that arise in online platforms. The key benefit of the decentralized approach lies in alleviating the need for the platform's investigation of disputes to be quick. Combined with appropriately chosen penalties for misuse of the mechanism, this benefit can be enjoyed by the platform while simultaneously ensuring desirable behavior from the platform's participants.

Appendix

Proof of Lemma 1

When using the centralized mechanism to remove a non-malicious negative review, the seller incurs cost $c \geq 0$, but the negative review is never removed, since the platform is assumed to never misjudge a genuine review as being malicious). □

Proof of Lemma 2

We calculate the posterior probability using Bayes' Rule. We start with a_P^i for $i = n, c, s$.

$$a_P^i = \frac{\Pr(q=h; R=P; i)}{\Pr(R=P, i)} = \frac{\Pr(R=P \mid q=h, i) \cdot \Pr(q=h)}{\Pr(R=P \mid q=h, i) \cdot \Pr(q=h) + \Pr(R=P \mid q=l, i) \cdot \Pr(q=l)}.$$

As $\Pr(R=P \mid q=l; i) = 0$, we have $a_P^i = 1$ for $i = n, c, s$.

Similarly, for a_0^i, we have:

$$a_0^i = \frac{\Pr(q=h; R=0; i)}{\Pr(R=0, i)} = \frac{\Pr(R=0 \mid q=h, i) \cdot \Pr(q=h)}{\Pr(R=0 \mid q=h, i) \cdot \Pr(q=h) + \Pr(R=0 \mid q=l, i) \cdot \Pr(q=l)}.$$

Note that $R=0$ could occur in two scenarios: when no customer (malicious or regular) purchases, or when a malicious review is removed. In each case, $\Pr(R=0 \mid q,i)=a$ for all q and i. Thus, $a_P^i=a$ for all i.

Finally, for a_N^i, we have:

$$a_N^i=\frac{\Pr(q=h;R=N;i)}{\Pr(R=N,i)}=\frac{\Pr(R=N \mid q=h,i)\cdot\Pr(q=h)}{\Pr(R=N \mid q=h,i)\cdot\Pr(q=h)+\Pr(R=N \mid q=l,i)\cdot\Pr(q=l)}.$$

We have: $\Pr(q=h)=a$ and $\Pr(q=l)=1-a$. For $\Pr(R=N \mid q,i)$ for $q=h,l$, we consider $i=n$ first:

$$\Pr(R=N \mid q=h,i=n)=\beta\cdot\Pr(R=N \mid q=h,j=M,i=n)+(1-\beta)\cdot\Pr(R=N \mid q=h,j=G,i=n),$$

where $j=M,G$ represents that the customer type malicious or genuine, respectively. If a malicious consumer purchases, leaves a negative review, and the firm does nothing ($i=n$), we have:

$$\Pr(R=N \mid q=h,j=M,i=n)=1.$$

As for the regular customer ($j=G$), $R=N$ if and only if a regular customer purchases (with probability $1-\frac{p}{a\theta}$) and has a negative experience (with probability $1-\theta$). Thus,

$$\Pr(R=N \mid q=h,j=G,i=n)=\frac{p}{a\theta}$$

Then

$$\Pr(R=N \mid q=h,i=n)=\beta(1)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta).$$

Similarly,

$$\Pr(R=N \mid q=l,i=n)=\beta(1)+(1-\beta)\left(1-\frac{p}{a\theta}\right).$$

Combining the two scenarios, we have

$$a_N^n=\frac{a\left[\beta+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]}{a\left[\beta+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]+(1-a)\left[\beta+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right]}=\frac{a}{a+(1-a)\frac{\left(\beta+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}{\left(\beta+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right)}}.$$

Similarly, for strategy $i=c$, we have

$$a_N^c=\frac{\Pr(R=N \mid q=h,i=c)\cdot\Pr(q=h)}{\Pr(R=N \mid q=h,i=c)\cdot\Pr(q=h)+\Pr(R=N \mid q=l,i=c)\cdot\Pr(q=l)}$$

Note that

$$\begin{aligned}\Pr(R=N \mid q=h,i=c)&=\beta\cdot\Pr(R=N \mid q=h,j=M,i=c)+(1-\beta)\cdot\Pr(R=N \mid q=h,j=G,i=c)\\&=\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta),\end{aligned}$$

where $(1-\gamma\delta)$ is the probability that the seller's report of a malicious review is not processed correctly and immediately by the platform. Moreover,

$$\Pr(R=N \mid q=l,i=c)=\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right).$$

Thus,

$$a_N^c=\frac{a\left[\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]}{a\left[\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]+(1-a)\left[\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right]},$$

$$= \frac{a}{a + (1-a)\frac{\left(\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}{\left(\beta(1-\gamma\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right)}}.$$

Finally, for strategy $i = s$,

$$a_N^s = \frac{\Pr(R = N \mid q = h, i = s) \cdot \Pr(q = h)}{\Pr(R = N \mid q = h, i = s) \cdot \Pr(q = h) + \Pr(R = N \mid q = l, i = c) \cdot \Pr(q = l)}.$$

Note that

$$\begin{aligned}\Pr(R = N \mid q = h, i = s) &= \beta \cdot \Pr(R = N \mid q = h, j = M, i = s) + (1-\beta) \cdot \Pr(R = N \mid q = h, j = G, i = s) \\ &= \beta(0) + (1-\beta)\left(1 - \frac{p}{a\theta}\right)(1-\theta),\end{aligned}$$

where the first term captures the scenario where the firm settles with the malicious customer, so that the negative review is removed by the malicious customer. Similarly,

$$\Pr(R = N \mid q = l, i = s) = \beta(0) + (1-\beta)\left(1 - \frac{p}{a\theta}\right).$$

Thus,

$$a_N^s = \frac{a\left[(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]}{a\left[(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right] + (1-a)\left[(1-\beta)\left(1-\frac{p}{a\theta}\right)\right]} = \frac{a}{a + (1-a)\frac{1}{1-\theta}}.$$

□

Proof of Proposition 1

The conditions under which $i = n$ is an equilibrium are detailed in the discussion before the proposition. In this proof, we focus on the conditions for $i = c$ and $i = s$.

First, $i = s$ is an equilibrium if and only if under the belief that $i = s$, the seller has no incentive to deviate to $i = c$ or $i = n$. We consider these two conditions in turns. First, when the seller faces a negative review and the belief is $i = s$, his net payoff by deviating from $i = s$ to $i = c$ is:

$$\Delta^{c|s} = \gamma\delta\pi(a_0^s) + (1-\gamma\delta)\pi(a_N^s) - c - [\pi(a_0^s) - r].$$

Thus, the seller does not deviate to $i = c$ if and only if $\Delta^{c|s} \leq 0$. As $a_0^s = a$, the condition becomes,

$$\pi(a_N^s) \leq \pi(a) + \frac{c - r}{1 - \gamma\delta}. \tag{7}$$

Similarly, the condition that the seller does not deviate to $i = n$ under the belief that $i = s$ is

$$\Delta^{n|s} = \pi(a_N^s) - [\pi(a_0^s) - r] \leq 0,$$

or equivalently,

$$\pi(a_N^s) \leq \pi(a) - r. \tag{8}$$

Combining the two conditions that preclude deviation, that is, (7) and (8), strategy $i = s$ is an equilibrium if and only if

$$\pi(a_N^s) \geq \max\left\{\pi(a) - r,\ \pi(a) - \frac{c}{\gamma\delta}\right\},$$

which corresponds to the second statement in the proposition.

Next, we consider to $i = c$, which is an equilibrium if and only if under this belief, the seller does not have incentive to deviate to $i = n$ and $i = s$. Using the same notation as in the paper, the seller will not deviate to $i = n$ if and only if

$$\Delta^{n|c} = \pi(a_N^c) - [\gamma\delta\pi(a_0^c) + (1-\gamma\delta)\pi(a_N^c) - c] \leq 0.$$

As $a_0^c = a$, the above condition is equivalent to

$$\gamma\delta[\pi(a) - \pi(a_N^c)] - c \geq 0. \tag{9}$$

Similarly, the seller will not deviating to $i = s$ when

$$\Delta^{s|c} = \pi(a_0^c) - r - [\gamma\delta\pi(a_0^c) + (1-\gamma\delta)\pi(a_N^c) - c] \leq 0,$$

that is,

$$(1-\gamma\delta)[\pi(a_N^c) - \pi(a)] - (c-r) \geq 0.$$

Combining this condition with (9), we have that $i = c$ is an equilibrium if and only if:

$$\pi(a_N^c) \in \left[\pi(a) + \frac{c-r}{1-\gamma\delta},\ \pi(a) - \frac{c}{\gamma\delta}\right],$$

which corresponds to the third statement in the proposition. □

Proof of Proposition 2

We prove the result by backward induction. First, assuming the malicious customer has purchased, he will request the equilibrium ransom r^* which is the maximum possible ransom such that $i = s$ is the seller's preferred equilibrium (that is, either $i = s$ is the only equilibrium, or the firm's payoff under $i = s$ is greater than that under $i = c$ or $i = n$ if both are equilibria). To identify the relevant conditions, we rearrange Proposition 1 to get the following scenarios:

1. When $c < \gamma\delta(\pi(a) - \pi(a_N^c))$, $i = n$ is not an equilibrium because $a_N^n > a_N^c$. On the other hand, $i = c$ is an equilibrium if and only if the ransom $r > (1-\gamma\delta)(\pi(a) - \pi(a_N^c)) + c$. When $i = c$ is the equilibrium, the seller's terminal payoff conditional on a malicious review is

   $$\pi^c = (1-\gamma\delta)\pi(a_N^c) + \gamma\delta\pi(a) - c.$$

   On the other hand, $i = s$ is an equilibrium if and only if

   $$r \leq \min(\pi(a) - \pi(a_N^s),\ (1-\gamma\theta)(\pi(a) - \pi(a_N^s)) + c).$$

   When this condition holds, the seller's terminal payoff conditional on a malicious review is

   $$\pi^s = \pi(a) - r.$$

   Thus, the sufficient and necessary condition for $i = s$ to be the preferred equilibrium for the seller is that $i = s$ is an equilibrium and $\pi^s \geq \pi^c$, or equivalently,

   $$r \leq \min\left(\pi(a) - \pi(a_N^s),\ (1-\gamma\theta)(\pi(a) - \pi(a_N^s)) + c\,(1-\gamma\delta)(\pi(a) - \pi(a_N^c)) + c\right).$$

   Since $c < \gamma\delta(\pi(a) - \pi(a_N^c))$ and $a_N^s > a_N^c$, the above condition can be simplified to

   $$r \leq (1-\gamma\delta)(\pi(a) - \pi(a_N^c)) + c.$$

2. When $c \in [\gamma\delta[\pi(a) - \pi(a_N^c)],\ \gamma\delta(\pi(a) - \pi(a_N^n)))$, both $i = n$ and $i = c$ are equilibria for sufficiently large r. By the first scenario, we know that $i = s$ is an equilibrium and it is preferred by the seller over $i = c$ if and only if

$$r \leq (1 - \gamma\delta)(\pi(a) - \pi(a_N^c)) + c. \tag{10}$$

Further, in this scenario, $i = n$ is an equilibrium if and only if $r \geq \pi(a) - \pi(a_N^n)$. In this case, the seller's terminal payoff conditional on a malicious review is $\pi^n = \pi(a_N^n)$, which is less than π^s if and only if $r \leq \pi(a) - \pi(a_N^n)$. This condition always holds when $r \leq (1 - \gamma\delta)(\pi(a) - \pi(a_N^c)) + c$, as $a_N^c < a_N^n$ and $c < \gamma\delta(\pi(a) - \pi(a_N^n))$. Combined, $i = s$ is the preferred equilibrium if and only if (10) holds. Combining this with the first scenario above leads to the equilibrium ransom r^* in the second statement in the proposition.

3. When $c \geq \gamma\delta(\pi(a) - \pi(a_N^n))$, $i = c$ is not an equilibrium because $a_N^n > a_N^c$. On the other hand, $i = c$ is an equilibrium if and only if $r > \pi(a) - \pi(a_N^n)$. From the analysis of the previous scenario, it follows that $i = s$ is the preferred equilibrium if and only if

$$r \leq \pi(a) - \pi(a_N^n),$$

leading to the the equilibrium ransom r^* in the first statement in the proposition.

Next, anticipating that if he purchases the equilibrium ransom will be r^* as described above, the malicious customer makes the purchase if and only if $p < r^*$. Substituting r^* from the first step into this condition leads to the purchase conditions in the proposition. □

Proof of Corollary 1

First, note that from the first statement in Proposition 2, when $\gamma\delta \leq \frac{c}{\pi(a) - \pi(a_N^n)}$ and $p < (1 - \gamma\delta)[\pi(a) - \pi(a_N^n)]$, the equilibrium ransom is $r_-^* = \pi(a) - \pi(a_N^n)$.

Next, by the second statement in Proposition 2, when $\gamma\delta > \frac{c}{\pi(a) - \pi(a_N^n)}$, and $p < (1 - \gamma\delta)[\pi(a) - \pi(a_N^n)]$, as $a_N^n < a_N^c$, the equilibrium ransom $r_+^* = (1 - \gamma\delta)[\pi(a) - \pi(a_N^c)] + c$. Let $\gamma\delta = \frac{c}{\pi(a) - \pi(a_N^n)} + \epsilon$ for $\epsilon > 0$. Thus,

$$r_+^* = \left(1 - \frac{c}{\pi(a) - \pi(a_N^n)} - \epsilon\right)[\pi(a) - \pi(a_N^c)] + c = (1 - \epsilon)[\pi(a) - \pi(a_N^c)] + c\left(1 - \frac{\pi(a) - \pi(a_N^c)}{\pi(a) - \pi(a_N^n)}\right).$$

Comparing r_+^* and r_-^*, we have:

$$\begin{aligned} r_+^* - r_-^* &= (1 - \epsilon)[\pi(a) - \pi(a_N^c)] + c\left(1 - \frac{\pi(a) - \pi(a_N^c)}{\pi(a) - \pi(a_N^n)}\right) - [\pi(a) - \pi(a_N^n)]. \\ &= [\pi(a_N^n) - \pi(a_N^c)]\left(1 - \frac{c}{\pi(a) - \pi(a_N^n)}\right) - \epsilon[\pi(a) - \pi(a_N^c)]. \end{aligned}$$

By the assumption that $\gamma\delta = \frac{c}{\pi(a) - \pi(a_N^n)} + \epsilon$, we have $\frac{c}{\pi(a) - \pi(a_N^n)} < 1$. In addition, by Lemma 2, $a_N^c > a_N^n$, and hence $\pi(a_N^n) > \pi(a_N^c)$. Therefore, for sufficiently small ϵ, we have $r_+^* - r_-^* > 0$. Put differently, when $\gamma\delta$ increases from $\frac{c}{\pi(a) - \pi(a_N^n)}$ to $\frac{c}{\pi(a) - \pi(a_N^n)} + \epsilon$, r^* increases. Therefore, we have that r^* is not monotonically decreasing in $\gamma\delta$. □

Proof of Proposition 3

Define function $\Pi_0(p)$ for $p \in [0,1]$ as

$$\Pi_0(p) = \beta\pi(a) + (1-\beta)\{p\pi(a) + (1-p)[p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)]\}.$$

By the definition of p_0, we have that $p_0 = \arg\max_{p\in[0,1]} \Pi_0(p)$.

To prove the first statement in the proposition, we note that by the definition of $\Pi_C(p)$, $\Pi_0(p) \geq \Pi_C(p)$ for all $p \in [0,1]$. Thus, if $p_0 \geq \bar{p}$, by the definition of $\bar{p}$, we have $p_0 \notin \mathcal{P}_{pur}$. Consequently, $\Pi_C(p^0) = \Pi_0(p_0) > \Pi_0(p) > \Pi_C(p)$ for all $p \in [0,1]$. Therefore, $p^* = p_0$.

For the second statement, note that as $\Pi_0(p)$ is concave oin p and $p_0 < \bar{p}$. Thus, $\Pi_0(p) < \Pi_0(\bar{p})$ for all $p > \bar{p}$. Further, by the definition of $\bar{p}$, we have that $p \notin \mathcal{P}_{pur}$ for $p > \bar{p}$. In other words, for $p > \bar{p}$, the malicious customer does not purchase, and hence $\Pi_0(p) = \Pi_C(p)$ for $p > \bar{p}$. Thus, $\Pi_C(p) < \Pi_C(\bar{p})$ for all $p > \bar{p}$, and hence $p^* \leq \bar{p}$. □

Proof of Proposition 4

As the seller would never use the decentralized mechanism to remove positive reviews, we focus on the case of negative reviews from regular customers. To prove the result, we use ij to represent the potential strategy the seller may follow when facing a malicious customer $i = s, n, d$ and when facing a regular customer $j = n, d$. Following the definition a_R^i in Lemma 2, we let a_R^{ij} be the posterior belief when future customers see review $R \in \{P, 0, N\}$ and believe that the seller adopts strategy ij. By this definition, we have that $a_R^{in} = a_R^i$ as in Lemma 2, and using the same technique as in the proof of Lemma 2, a_R^{id}, that is, the posterior belief when the seller uses the decentralized mechanism to remove non-malicious reviews, is: $a_P^{id} = 1$ for $i = s, n, d$, and

$$\begin{aligned}
a_0^{sd} &= \frac{a}{a + (1-a)\frac{\beta + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)\right]}{\beta + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)(1-\theta)\right]}} < a;\\
a_N^{sd} &= \frac{a}{a + (1-a)\frac{1}{1-\theta}} = a_N^s;\\
a_0^{nd} &= \frac{a}{a + (1-a)\frac{\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)}{\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)(1-\theta)}} < a;\\
a_N^{nd} &= \frac{a}{a + (1-a)\frac{\beta + (1-\beta)(1-\frac{p}{a\theta})\gamma}{\beta + (1-\beta)(1-\frac{p}{a\theta})\gamma(1-\theta)}} > a_N^s;\\
a_0^{dd} &= \frac{a}{a + (1-a)\frac{\beta(1-\gamma) + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)\right]}{\beta(1-\gamma) + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)(1-\theta)\right]}} < a;\\
a_N^{dd} &= \frac{a}{a + (1-a)\frac{\beta + (1-\beta)(1-\frac{p}{a\theta})}{\beta + (1-\beta)(1-\frac{p}{a\theta})(1-\theta)}} > a_N^s.
\end{aligned}$$

Under the above notation, to prove the proposition it suffices to prove that $ij \in \{sd, nd, dd\}$ cannot be an equilibrium for $b \geq \underline{b}$. We focus on showing $ij = sd$ is not an equilibrium, noting that the other two cases (i.e., $ij = nd$ and $ij = dd$) can be shown similarly.

Note that $ij = sd$ is an equilibrium only if deviating to $ij = sn$ is not profitable for the seller when facing a regular customer, that is,

$$(1-\gamma)\pi(a_0^{sd}) + \gamma\pi(a_N^{sd}) - b \geq \pi(a_N^{sd}),$$

As $a_0^{sd} = a_N^s$, the above condition is equivalent to

$$b \leq (1-\gamma)(\pi(a_0^{sd}) - \pi(a_N^s)).$$

Since $a_0^{sd} < a$, this condition cannot hold for $b \geq \underline{b}$. Thus, $ij = sd$ cannot be an equilibrium.

Similarly, it can be shown that $ij = dd$ and $ij = nd$ cannot be an equilibrium given the posterior belief sa_N^{id} and a_0^{id}. Combining these three cases, we have that when $b \geq \underline{b}$, the seller does not use the decentralized mechanism to remove nonmalicious reviews. □

Proof of Proposition 5

By Proposition 4, under the assumption that $b \geq \underline{b}$ it suffices to focus on the strategies $ij = sn, dn, nn$. To simplify the notation, in what follows we omit the component $j = n$ and write only $i = s, d, n$. The proof follows a similar structure of that of Propositions 1 and 2 with the centralized mechanism. Specifically, we follows three steps:

1. Establish conditions for $i \in \{s, d, n\}$ to be an equilibrium.
2. Determine the equilibrium ransom r^* given that a malicious customer has purchased.
3. Determine the malicious customer's purchase decision.

Step 1: Conditions for $i \in \{s, d, n\}$ as an equilibrium. First, $i = s$ is an equilibrium if and only if

$$\pi(a_0^s) - r \geq (1-\gamma(1-\delta))\pi(a_0^s) + \gamma(1-\delta)\pi(a_N^s) - (1-\delta)b;$$
$$\pi(a_0^s) - r \geq \pi(a_N^s).$$

where the first (second) condition guarantees that the seller has no incentive to deviate to $i = d$ ($i = n$). Since $a_0^s = a$, the above conditions are equivalent to:

$$r \leq \min(\pi(a) - \pi(a_N^s),\ (1-\delta)(\gamma(\pi(a) - \pi(a_N^s)) + b)).$$

Similarly, $i = n$ is an equilibrium if and only if

$$\pi(a_N^n) \geq \pi(a_0^n) - r;$$
$$\pi(a_N^n) \geq (1-\gamma(1-\delta))\pi(a_0^n) + \gamma(1-\delta)\pi(a_N^n) - (1-\delta)b,$$

where the first (second) condition guarantees that the seller has no incentive to deviate to $i = s$ ($i = d$). We note that $a_0^n = a$, so that the above conditions can be written as

$$r \geq \pi(a) - \pi(a_N^n);$$
$$b \geq \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a) - \pi(a_N^n)).$$

Finally, $i = d$ is an equilibrium if and only if

$$(1-\gamma(1-\delta))\pi(a_0^d) + \gamma(1-\delta)\pi(a_N^d) - (1-\delta)b \geq \pi(a_N^d),$$
$$(1-\gamma(1-\delta))\pi(a_0^d) + \gamma(1-\delta)\pi(a_N^d) - (1-\delta)b \geq \pi(a_0^d) - r,$$

which can be simplified to

$$b \le \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d));$$
$$b \le \gamma(\pi(a)-\pi(a_N^d))-r.$$

Step 2: Equilibrium ransom. To determine the equilibrium ransom, we first compare the magnitudes of the relevant posterior beliefs. In particular, we have $a_0^d = a$, and

$$a_N^d = \frac{a}{a+(1-a)\frac{\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)}}.$$

Thus, we have that $a = a_0^s = a_0^n = a_0^d > a_N^n > a_N^d > a_N^s$. Given this relationship, we determine the equilibrium ransom r^* according to the following three scenarios:

1. When $b \le \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n))$, $i=n$ is not an equilibrium. Thus, for $i=s$ to be the seller's preferred equilibrium, the seller's payoff under $i=s$ must not be less than under $i=d$, that is,

   $$\pi(a_0^s)-r \ge (1-\gamma(1-\delta))\pi(a_0^d)+\gamma(1-\delta)\pi(a_N^d)-(1-\delta)b,$$

   Since $a_0^s = a_0^d = a$, the above condition becomes

   $$r \le (1-\delta)[\gamma(\pi(a)-\pi(a_N^d))+b].$$

   Thus, the equilibrium ransom is $r^* = (1-\delta)[\gamma(\pi(a)-\pi(a_N^d))+b]$.

2. When $b \in \left[\frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n)),\ \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d))\right)$, for $i=s$ to be the preferred equilibrium, the seller's payoff under $i=s$ must not be less than that under $i=n$ and $i=d$, that is,

   $$\pi(a_0^s)-r \ge \max\left(\pi(a_0^n),\ (1-\gamma(1-\delta))\pi(a_0^d)+\gamma(1-\delta)\pi(a_N^d)-(1-\delta)b\right),$$

   or, equivalently,

   $$r \le \min\left(\pi(a)-\pi(a_N^n),\ (1-\delta)[\gamma(\pi(a)-\pi(a_N^d))+b]\right).$$

   Since $b < \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d))$ and $a_N^n < a_N^d$, we have $\pi(a)-\pi(a_N^n) < (1-\delta)[\gamma(\pi(a)-\pi(a_N^d))+b]$. Thus, the binding constraint is $r \le \pi(a)-\pi(a_N^n)$ and the equilibrium ransom is $r^* = \pi(a)-\pi(a_N^n)$.

3. When $b > \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d))$, $i=d$ cannot be an equilibrium. Thus, the equilibrium ransom is $r^* = \pi(a)-\pi(a_N^n)$, as in the previous scenario. Combining this case with the previous one, we arrive at the equilibrium ransom in the first statement of the proposition.

Step 3: Malicious customer's purchase decision. The malicious customer will purchase if and only if $r^* > p$, so that the purchase decision follows immediately from the equilibrium ransom. □

References

Alizamir, Saed, Francis De Véricourt, Peng Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Babich, Volodymyr, Simone Marinesi, Gerry Tsoukalas. 2020. Does crowdfunding benefit entrepreneurs and venture capital investors? *Manufacturing & Service Operations Management* .

Bakos, Y., C. Dellarocas. 2011. Cooperation without enforcement? a comparative analysis of litigation and online reputation as quality assurance mechanisms. *Management Science* **57**(11) 1944–1962.

Bimpikis, K., Y. Papanastasiou, W. Zhang. 2020. Information provision in two-sided platforms: Optimizing for supply. *Available at SSRN* .

Bolton, G., B. Greiner, A. Ockenfels. 2018. Dispute resolution or escalation? the strategic gaming of feedback withdrawal options in online markets. *Management Science* **64**(9) 4009–4031.

Chakraborty, Soudipta, Robert Swinney. 2020. Signaling to the crowd: Private quality information and rewards-based crowdfunding. *Manufacturing & Service Operations Management* .

Chen, Li, Yiangos Papanastasiou. 2019. Seeding the herd: Pricing and welfare effects of social learning manipulation. *Available at SSRN 3456139* .

Competition and Markets Authority. 2015. Report on the cma's call for information. *https://www.gov.uk/government/consultations/online-reviews-and-endorsements* .

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

Feldman, P., Y. Papanastasiou, E. Segev. 2019a. Social learning and the design of new experience goods. *Management Science* **65**(4) 1502–1519.

Feldman, Pnina, Andrew E Frazelle, Robert Swinney. 2019b. Can delivery platforms benefit restaurants? *Available at SSRN 3258739* .

Jin, Chen, Luyi Yang, Kartik Hosanagar. 2019. To brush or not to brush: Product rankings, customer search and fake orders. *Customer Search and Fake Orders (September 30, 2019) .NET Institute Working Paper* (19-02).

Kanoria, Yash, Daniela Saban. 2017. Facilitating the search for partners on matching platforms .

Kostami, Vasiliki, Sampath Rajagopalan. 2014. Speed–quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management* **16**(1) 104–118.

Lappas, Theodoros, Gaurav Sabnis, Georgios Valkanas. 2016. The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research* **27**(4) 940–961.

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

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

Mayzlin, Dina, Yaniv Dover, Judith Chevalier. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* **104**(8) 2421–55.

Mukherjee, Arjun, Bing Liu, Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. *Proceedings of the 21st international conference on World Wide Web*. ACM, 191–200.

Papanastasiou, Y., K. Bimpikis, N. Savva. 2018. Crowdsourcing exploration. *Management Science* **64**(4) 1727–1746.

Papanastasiou, Y., N. Savva. 2017. Dynamic pricing in the presence of social learning and strategic consumers. *Management Science* **63**(4) 919–939.

Papanastasiou, Yiangos. 2020. Fake news propagation and detection: A sequential model. *Management Science* **66**(5) 1826–1846.

Parker, G. G, M. W Van Alstyne, S. P. Choudary. 2016. *Platform revolution: How networked markets are transforming the economy and how to make them work for you*. WW Norton & Company.

Rule, Colin, Chittu Nagarajan. 2010. Leveraging the wisdom of the crowds: the ebay community court and the future of online dispute resolution. *ACResolution 2 (2)* **4–7**.

TripAdvisor. 2018. Reporting potential blackmail to tripadvisor: Report threats immediately. *www.tripadvisor.com/TripAdvisorInsights/w592* .

Yang, S. Alex, Angela Huyue Zhang. 2019. Crowd-judging. Working paper.

Yu, M., L. Debo, R. Kapuscinski. 2016. Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Science* **62**(2) 410–435.

Zhang, Jiding, Sergei Savin, Senthil K Veeraraghavan. 2017. Revenue management in crowdfunding. *Available at SSRN 3065267* .

Zhang, K., X. Chen, C. Wu. 2020. Review extortion in an on-line marketplace. University of British Columbia Working paper.