Improving 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 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 seller profit, especially in settings characterized by high uncertainty about product quality. The introduction of a standard centralized dispute resolution mechanism (whereby the seller can report allegedly 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 a more 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. Our results suggest that decentralization, when implemented correctly, may represent 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, they may 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 recently experimented with more decentralized forms of governance, aimed at alleviating the demand for platform resources as well as 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 also experimented with a more decentralized approach to dispute resolution, 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 (Zhang 2021).

The introduction of this "semi-decentralized" 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

[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).

and sales while their cases were pending resolution by the platform.[4] The rationale behind Taobao's introduction of the more decentralized mechanism is that, when faced with a blackmail attempt, the seller can now bypass the platform and remove the malicious review on his/her 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 dispute resolution outcomes between platform participants and, if so, how these outcomes can be achieved.

To keep the 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. Semi-Decentralized: The seller removes the review without consulting the platform. If the platform subsequently 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 seller'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.

[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).

Overall, we observe that the combined impact of these two effects is most pronounced in settings 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/her to lower the 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 platform 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. We further observe that even in those cases where the platform's investigation process is highly efficient, the fraction of the seller's profit loss which is recovered by the dispute resolution mechanism can be underwhelming.

ii. The semi-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 by the host platform: 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 some degree of decentralization may be beneficial in the resolution of a broader range of disputes in online platforms.[5] 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 meaningful and appropriately chosen penalties can be associated with the mechanism. 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 more decentralized dispute resolution processes may also have implications for longer-term operational decisions such as internal team organization and 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 the semi-decentralized mechanism may represent a more advantageous approach (in §6 we extend our main model to establish the robustness of our results). We conclude with a discussion in §7.

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, Kwan et al. (2020) 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. Lee and Cui (2022) analytically compare platform-led adjudication and crowd-sourced dispute resolution. 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.

[5] For example, similar mechanisms have been implemented to help vendors address spamming (i.e., irrelevant information) in the consumer comments section, damaging product reviews left by competing sellers, and reviews attempting to divert consumer traffic elsewhere. More generally, decentralization of the form considered in this paper can be particularly beneficial when the classic resolution approach requires a costly process of collection and verification of evidence/information held privately by one of the parties involved in the dispute.

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 (2021) 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 (2021) 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. (2021) consider whether food-delivery platforms benefit restaurants. Kanoria and Saban (2021) 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 who purchases 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 $\bar{F}(\cdot) = 1 - F(\cdot)$.[6] 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 that the product's quality is high is denoted by $a \in (0, 1)$.[7] 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 $R \in \{N, 0, P\}$ is then 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/she 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 the 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 ransom request and make use of the platform's dispute resolution mechanism (see next section for details).[8] If the consumer is non-malicious (i.e., a regular consumer), he/she purchases the product if and only if the expected utility from purchase $u_j = a\theta v_j - p$ is nonnegative, and subsequently posts a (truthful) review depending on his/her product experience.

Apart from the payoff associated with the seller's interaction with the consumer as 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, convex 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).[9]

[6] 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_j can be attributed to the consumers' idiosyncratic preferences. Note that we assume that each consumer knows their own v_j.

[7] 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. 2019, Papanastasiou and Savva 2017, Yu et al. 2016)

[8] 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.

[9] While not necessary for our main results, convexity of $\pi(\cdot)$ ensures that market learning is beneficial for the seller in expectation.

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:

(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) Semi-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 $b \geq 0$.[10]

Note that while the seller always knows whether a negative review was posted by a malicious consumer, the seller has no way of conveying this information to the platform efficiently, beyond presenting whatever evidence he/she can collect from the interaction with the consumer, and waiting for the platform to investigate.[11]

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 incurring significant delays, for example, owing to the platform's total case load at the time of a reported incident. In our main analysis, we assume that if the case is not investigated immediately, the investigation is delayed until the end of the selling season; in §6.1, we discuss an extension of our model that captures intermediate degrees of delay. Throughout our analysis, we refer to γ as the "timeliness" parameter.

Second, if the case involves a malicious review, we assume that the review is correctly identified by the platform as malicious with probability $\delta \in [0, 1]$. In particular, we note that the seller carries the responsibility of presenting convincing evidence that the consumer has engaged in malicious behavior (i.e., a ransom demand or equivalent threat). Thus, the assumption that errors may occur in the platform's final decision may relate to the seller's inability to collect evidence deemed to be satisfactory by the platform, the malicious consumers' skillfulness in conducting the ransom demand, and/or the subjective assessment of the available case information, among others.[12] We

[10] Assuming that the seller also incurs a hassle cost for using the decentralized mechanism has no qualitative bearing on our results (see also Figure 7 in §5).

[11] Note that a malicious review as defined in this paper is a review which is accompanied by a ransom demand, irrespective of the review content and/or the product's true underlying quality.

[12] We note that the platform only examines evidence pertaining to whether malicious behavior has occurred. It does not seek to verify whether or not the experience described in the consumer review is accurate, as such verification is typically impractical or impossible, not least due to the subjective nature of consumer reviews.

refer to δ throughout as the "accuracy" parameter, which we treat as exogenous in our main analysis; in §6.2, we investigate the impact of platform efforts to increase the judgment accuracy. For simplicity, we assume that the probability of a genuine negative review being misjudged by the platform as malicious is negligible.[13]

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.
   (1) If the consumer is malicious, he/she observes the price and decides whether to purchase. Following a purchase decision, he/she posts a fake negative review and chooses a ransom r to be demanded from the seller in exchange for removing the review.
   (2) If the consumer is non-malicious, he/she observes the price and decides whether to purchase. Following a purchase decision, he/she posts a truthful review according to his/her experience with the product.
3. The seller observes the posted review. If there is a ransom demand, the firm chooses whether to accept the demand, reject it, or utilize the available dispute resolution mechanism. Note that if there is no ransom demand, 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 formed via Bayes' rule.

In addition to the above events that apply to both mechanisms, in the case of the decentralized mechanism the platform also chooses the mechanism's penalty b at the beginning of the game (see Section 5). Throughout our analysis, we focus on perfect Bayesian equilibria (PBE) in pure

[13] That is, given that no misconduct has occurred in the first place and the review is genuine, we assume that the probability of being able to present evidence that convinces the platform that the review is malicious is small; we note that allowing for errors of this type does not affect the qualitative nature of our results.

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 entails 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, the seller does not dispute genuine reviews.*

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 seller 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

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

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*

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

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 θ (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 value a_N^n (which is greater than both a_N^c and a_N^s). 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 (i.e., either $a_0^c = a$ or a_N^s, depending on the outcome of the platform's investigation).

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

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

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

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

or, equivalently, when $\pi(a_N^n) \geq \pi(a) - \min\{r, \frac{c}{\gamma\delta}\}$. 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:*

(1) An equilibrium with $i = n$ exists if and only if $\pi(a_N^n) \geq \max\{\pi(a) - r, \pi(a) - \frac{c}{\gamma\delta}\}$.
(2) An equilibrium with $i = s$ exists if and only if $\pi(a_N^s) \leq \min\{\pi(a) - r, \pi(a) + \frac{c-r}{1-\gamma\delta}\}$.
(3) An equilibrium with $i = c$ exists if and only if $\pi(a) + \frac{c-r}{1-\gamma\delta} \leq \pi(a_N^c) \leq \pi(a) - \frac{c}{\gamma\delta}$.

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 negative review removed, and also (ii) satisfies $r > p$ yielding positive surplus for the malicious consumer.

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

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

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

(2) When $c < \gamma\delta\,(\pi(a) - \pi(a_N^n))$, the malicious consumer purchases if and only if $p < (1 - \gamma\delta)[\pi(a) - \pi(a_N^c)] + c$. He/she 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 ($\pi(a) - \pi(a_N^n)$)—this is the maximum ransom the malicious consumer can

extract from the seller. Having identified the maximum ransom he/she can extract, the consumer purchases if and only if the product's price is sufficiently low to allow positive surplus.

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 the ransom demand is too high, the mechanism provides a preferable course of action for the seller. To avoid this, the ransom demand cannot exceed $(1-\gamma\delta)[\pi(a)-\pi(a_N^c)]+c$, 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, the malicious consumer 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/her "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 manifests in 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 $\frac{c}{\pi(a)-\pi(a_N^n)}$.

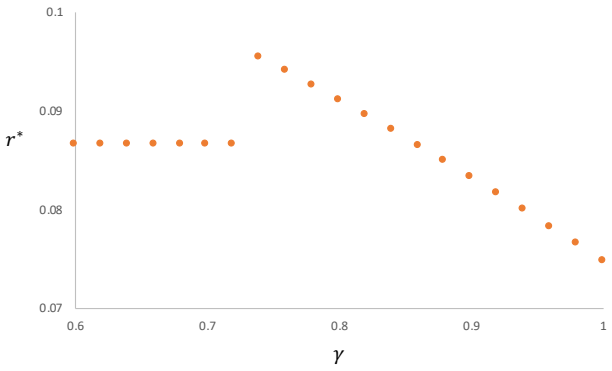

Figure 1 **Equilibrium ransom as a function of the centralized mechanism efficiency. Parameter values:** $a = \theta = 0.5$, $\beta = 0.1$, $\delta = 0.8$, $p = 0.08$, $c = 0.05$.

4.4. Seller's Pricing Decision

We consider next the seller's pricing problem. Building on the analysis of §4.3, let $\mathcal{P}_C^{pur}$ 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

$$\begin{aligned}\Pi_C(p) = \beta[\pi(a) - \mathbb{1}_{p\in\mathcal{P}_C^{pur}}(r^*(p) - p)] \\ + (1-\beta)\left[\left(\frac{p}{a\theta}\right)\pi(a) + \left(1 - \frac{p}{a\theta}\right)\left(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)\right)\right],\end{aligned} \tag{3}$$

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 $p \notin \mathcal{P}_C^{pur}$, 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 $p \in \mathcal{P}_C^{pur}$, then the malicious consumer purchases, and the seller agrees to pay the ransom $r^*(p)$ to have the malicious review removed. The second term is the seller's expected payoff in the event that the consumer is nonmalicious. In this case, the seller's payoff depends on whether the consumer chooses to purchases and, if so, the review he/she generates after experiencing the product.

Observe that by Lemma 2, the continuation payoff $\pi(a_N^s)$ is independent of the seller's chosen price. Moreover, the seller's payoff function $\Pi_C(p)$ is concave in p for all $p \notin \mathcal{P}_C^{pur}$. Therefore, the seller's problem may be viewed as choosing a price to maximize a concave function, less an (expected) penalty of $\beta[r^*(p) - p]$, which applies whenever a price $p \in \mathcal{P}_C^{pur}$ is chosen. Let p_0 be the unique maximizer of the seller's payoff function ignoring the penalty (or, equivalently, ignoring the presence of malicious consumers in the market),

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

which, assuming an interior solution, reduces to $p_0 = \frac{1}{2}\left[\pi(a) - \pi(a_N^s) + a\theta\left(1 - \pi(1) + \pi(a_N^s)\right)\right]$. It follows that if $p_0 \notin \mathcal{P}_C^{pur}$, 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 optimal. While the properties of the set $\mathcal{P}_C^{pur}$ depend on the functional form of the continuation payoff $\pi(\cdot)$, recall that in general the malicious consumer tends to purchase provided the price is sufficiently low (see Proposition 2). Accordingly, let us define

$$\bar{p}_C := \sup \mathcal{P}_C^{pur}, \tag{4}$$

which is the lowest price at which the malicious consumers choose not to purchase. Proposition 3 provides a characterization of the seller's pricing decision as a function of the prevalence of malicious consumer behavior in the market.

PROPOSITION 3. *There exists* $\underline{\beta}_C \in [0,1]$ *such that:*
(1) If $\beta < \underline{\beta}_C$, *then* $p^* \in [p_0, \bar{p}_C]$.
(2) If $\beta \geq \underline{\beta}_C$, *then* $p^* = p_0$.

Proposition 3 establishes that when the prevalence of malicious consumer behavior is below a threshold $\underline{\beta}_C$, the equilibrium price p^* exhibits an upwards distortion. This distortion helps to mitigate the damage incurred by the seller in the event that a malicious consumer is encountered, but it also causes a decrease in profit in the event that the seller encounters a nonmalicious consumer. In particular, we note that the higher price decreases the probability that a nonmalicious consumer will purchase and generate a review, which in turn causes a loss in (expected) future profit as a result of decreased market learning (see also §4.5, where we quantify the loss in profit incurred by the seller as a result of malicious consumer behavior).

To help explain in more detail the drivers underlying the behavior of the equilibrium price as described in Proposition 3, we enlist the example of Figure 2. In this example, we have $p_0 = 0.17$ and $\underline{\beta}_C = 0.78$; observe that, consistent with Proposition 3, the equilibrium price is higher than (equal to) p_0 for any $\beta < 0.78$ ($\beta \geq 0.78$). The figure also illustrates the malicious consumer's

ransom demand, in the event that such a demand occurs. In particular, observe that the malicious consumer purchases and successfully blackmails the seller only when $\beta < 0.3$; in all other cases, the seller's pricing decision is such that the malicious consumer chooses not to purchase. Based on these observations, we next discuss each qualitatively different region in more detail.

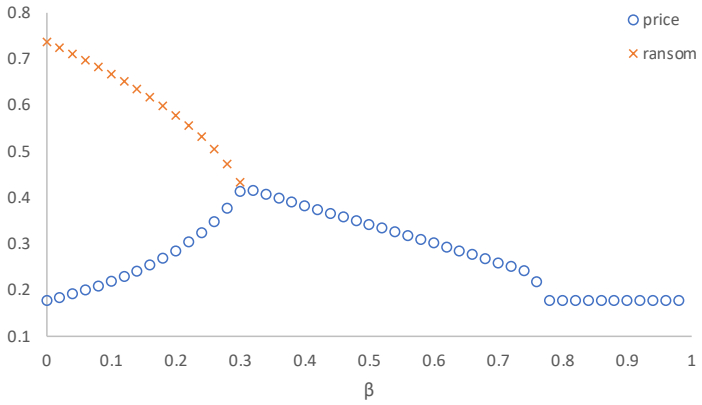

Figure 2 **Equilibrium price and ransom as a function of the probability that the customer is malicious. Parameter values:** $a = \theta = \gamma = \delta = 0.8$, $c = 0.1$, $\pi(a) = 4a^2$.

Consider first the region $\beta \in [0, 0.3]$, and note that in this region there is a positive ransom demand, 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 relatively close to p_0, in order to avoid a significant profit loss in the (much more probable) event that a non-malicious consumer is encountered. Observe, however, that while the probability of encountering a malicious consumer is small, whenever such an encounter does occur, the seller is forced to pay a heavy ransom; as the probability β increases, the seller thus 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 relatively low, probability.

Next, consider the region $\beta \in [0.3, 0.78]$. In this region, the probability of encountering a malicious consumer is sufficiently high so that the seller prefers to forgo some profit in the event that a non-malicious consumer is encountered, in order to completely avoid an encounter with the malicious consumer. To achieve this, 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 paying the ransom. Therefore, the optimal price in this region is the lowest price (i.e., the price which is closest to p_0) at which the malicious consumer is deterred from purchasing, that is, $\bar{p}_C$. We note that $\bar{p}_C$ is decreasing in β (see Lemma 2 and Proposition 2), which is why the equilibrium price in this region is also decreasing in β.

Finally, in the region $\beta \in [0.78, 1]$, the seller knows that there is a high chance of encountering a malicious consumer. However, this turns out to be irrelevant: in this region of β, the price threshold $\bar{p}_C$ falls below p_0, meaning that the price p_0 which maximizes profit in the event of a non-malicious consumer encounter is also high enough to deter the malicious consumer from attempting to blackmail the seller.

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.

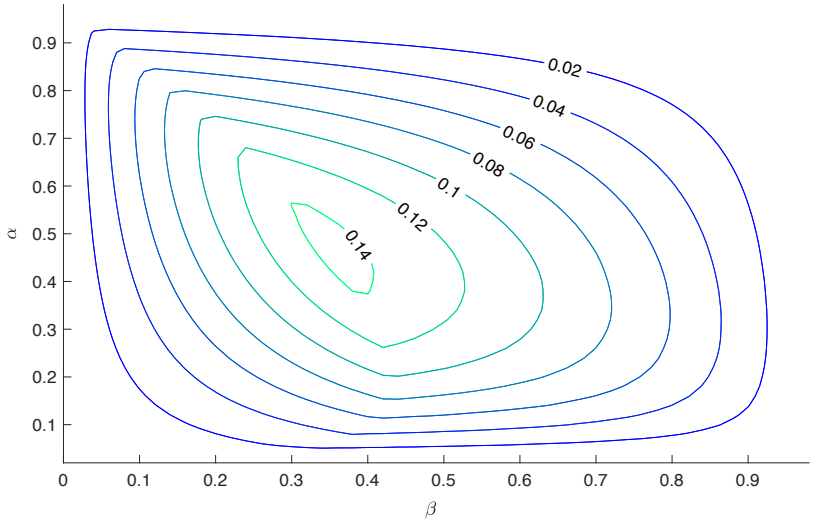

Figure 3 **Efficiency loss in the absence of a dispute resolution mechanism,** $1 - \Pi^*_{no}/\Pi^*_{opt}$**. Parameter values:** $\theta = 0.9$**,** $\pi(a) = 50a^2$**.**

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.

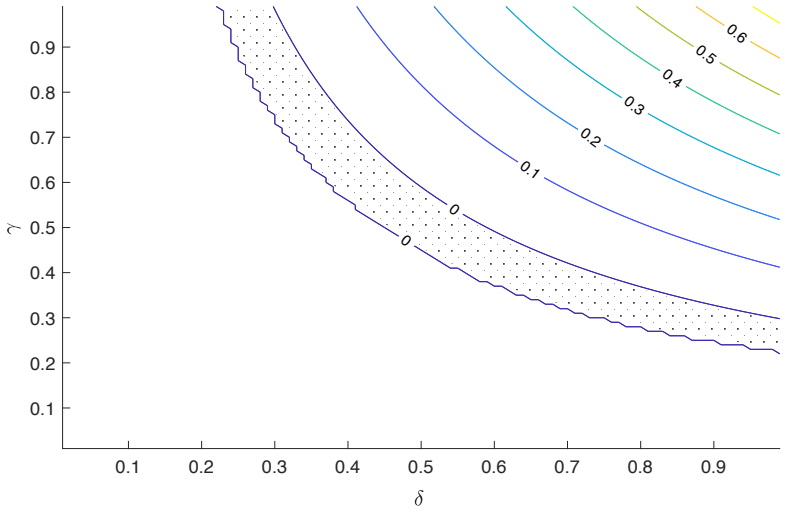

Figure 4 **Efficiency loss recovered by the centralized dispute resolution mechanism,** $(\Pi_C^* - \Pi_{no}^*)/(\Pi_{opt}^* - \Pi_{no}^*)$**. Parameter values:** $\theta = 0.9$**,** $a = 0.5$**,** $\beta = 0.3$**,** $c = 2$**,** $\pi(a) = 50a^2$**.**

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) for the mechanism to be 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 literature (see Alizamir et al. 2013, Kostami and Rajagopalan 2014, and references therein).

5. Semi-Decentralized Dispute Resolution

In this section, we analyze the properties of the "semi-decentralized" dispute resolution mechanism ("D"). Under this mechanism, the seller can remove a review immediately without consulting the platform; however, if the platform later 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. In implementing this mechanism for dispute resolution, the platform hosting the seller must also choose the penalty associated with the mechanism; accordingly, the analysis of this section includes an extra step at the beginning of the game, where the platform chooses the penalty b (see §5.3).

Under the semi-decentralized mechanism, the space of seller strategies to be considered is larger, because the seller may now choose to use the mechanism to remove genuine negative reviews (i.e., in addition to removing malicious negative reviews). It is useful to point out that such abuse of the mechanism by the seller would be undesirable for the platform, because a mechanism that facilitates the removal of genuine reviews by sellers implies a review system that lacks credibility—this can be particularly problematic in terms of the platform's ability to attract and retain customers. Accordingly, to focus the exposition on those cases which are more relevant from a practical standpoint, we will make use of the following result.

Proposition 4. *Under the semi-decentralized mechanism, the seller does not remove genuine reviews if and only if* $b \geq \underline{b} := (1-\gamma)\left[\pi(a) - \pi\left(a_N^s\right)\right]$.

The proof of the result establishes that for $b \geq \underline{b}$ an equilibrium exists, and that the equilibrium can only be such that the seller does not use the mechanism to remove genuine reviews; conversely, such an equilibrium can only exist when the penalty satisfies $b \geq \underline{b}$. Intuitively, abuse of the mechanism by the seller can be avoided if the platform chooses a sufficiently high penalty for wrongdoing.

Given that the penalty is chosen by the platform at the beginning of the game (and assuming the platform's penalty choice is such that abuse by the seller is deterred), in the analysis that follows we will focus on equilibria under mechanisms with $b \geq \underline{b}$ (the case of mechanisms with $b < \underline{b}$ is addressed in Proposition 13 of Appendix B).

5.1. Malicious Consumer's Strategy

We pick up the analysis of the semi-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 provided he/she 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

$$a_N^d = \frac{a}{a + (1-a)\frac{\left(\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}{\left(\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right)}}, \tag{5}$$

which denotes the market's posterior belief conditional on observing a negative review, when the seller uses the semi-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:*

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

$$r^* = \left(\pi(a) - \pi(a_N^n)\right). \tag{6}$$

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

$$r^* = (1-\delta)\left(\gamma\left(\pi(a) - \pi(a_N^d)\right) + b\right). \tag{7}$$

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/she is judged to have wrongfully removed a review, the decentralized mechanism does not constitute a credible course of action for the seller. Note that the threshold value of the penalty above which the 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 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 demand (should such a demand occur).
ii. Under the semi-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 demand.

The key difference between the two mechanisms with respect to the impact of the timeliness parameter γ lies is the fact that under the centralized mechanism, a more timely investigation of the incident by the platform can only benefit the seller, who suffers from the reputational damage inflicted by the negative review while the dispute remains unresolved; by contrast, under the semi-decentralized mechanism, the review is removed immediately by the seller and a more timely investigation by the platform can only harm the seller, in the event that the platform reinstates the review after judging it to have been wrongfully removed.

5.2. Seller's Pricing Decision

Having characterized the malicious consumer's strategy for a given price p, we now consider the seller's pricing decision. Building on §5.1 and following the same approach as in the analysis of the centralized mechanism, let $\mathcal{P}_D^{pur}$ be the set of prices at which the malicious consumer chooses to purchase and make a ransom demand. We may then express the seller's payoff as a function of his/her pricing decision,

$$\begin{aligned}\Pi_D(p) = \beta[\pi(a) - \mathbb{1}_{p\in\mathcal{P}_D^{pur}}(r^*(p)-p)] \\ + (1-\beta)\left[\left(\frac{p}{a\theta}\right)\pi(a) + \left(1-\frac{p}{a\theta}\right)\left(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)\right)\right],\end{aligned} \tag{8}$$

where $r^*(p)$ is given in Proposition 5. The first term captures the seller's expected profit in the event that a malicious consumer is encountered, while the second term is the seller's expected payoff in the event that the consumer is non-malicious.

Following the same approach as in §4.4, the seller's problem can be viewed as a problem of choosing a price to maximize a concave function, less an expected penalty equal to $\beta(r^*(p)-p)$ that applies whenever a price $p \in \mathcal{P}_D^{pur}$ is chosen (i.e., so that the malicious consumer enters the

market and extracts a ransom). Recall that the unique maximizer of the seller's problem ignoring the presence of malicious consumers is p_0 (see §4.4) and define

$$\bar{p}_D := \max \mathcal{P}_D^{pur}$$

as the lowest price at which the malicious consumer is deterred from entering the market. We then have the following result.

PROPOSITION 6. *There exists* $\underline{\beta}_D \in [0,1]$ *such that:*
(1) If $\beta < \underline{\beta}_D$, *then* $p^* \in [p_0, \bar{p}_D]$.
(2) If $\beta \geq \underline{\beta}_D$, *then* $p^* = p_0$.

Proposition 6 mirrors the corresponding result in the case of the centralized mechanism, while the numerical experiment of Figure 5 illustrates that the behavior of the equilibrium price is qualitatively similar. In particular, we observe that at relatively low values of β, the price chosen by the seller exhibits an upwards distortion (relative to p_0) and is such that the malicious consumer chooses to enter and is able to extract positive ransom from the seller. At intermediate values of β, the seller's price is higher than p_0 and the malicious consumer in equilibrium does not attempt to blackmail the seller. Finally, at high values of β, the equilibrium price returns to p_0 and the malicious consumer does not enter the market. The details underlying these observations parallel those discussed in §4.4 for the example of Figure 2; we thus refrain from repeating the more detailed discussion here.

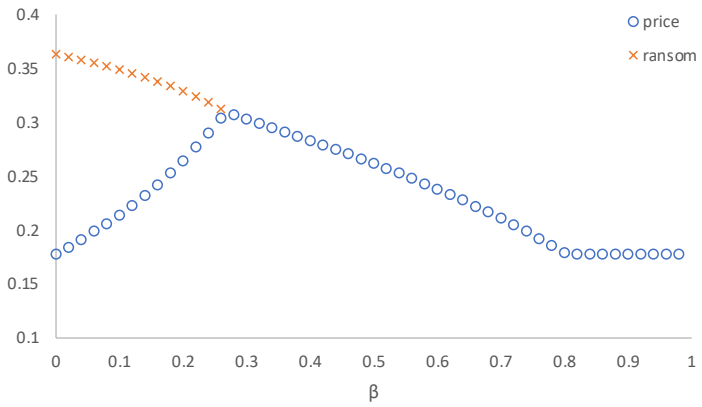

Figure 5 **Equilibrium price and ransom as a function of the probability that the customer is malicious. Parameter values:** $a = \theta = \gamma = \delta = 0.8$, $b = 0.4$, $\pi(a) = 4a^2$.

5.3. Platform's Penalty Decision

To complete the equilibrium analysis of the semi-decentralized dispute resolution mechanism, we now consider the platform's equilibrium choice of the penalty b. In choosing the penalty, we assume that the platform takes into consideration both sides of the market (i.e., the seller and the consumers). With regards to the consumer side, choosing a penalty that satisfies $b \geq \underline{b}$ ensures that the seller does not abuse the provided mechanism to remove genuine reviews, which in turn maintains the credibility of the platform's review system—this is crucial for the platform in terms of attracting and retaining consumers. Next, having ensured that sellers do not abuse the mechanism, we assume that the platform seeks to maximize the seller's equilibrium profit Π_D^* or, equivalently, seeks to maximize any increasing function of the seller's profit $g(\Pi_D^*)$ (for instance, the latter may relate to commission fees collected by the platform on transactions between the seller and the consumers).[14]

PROPOSITION 7. *There exists a threshold $\Delta \in (0,1]$ such that:*

(1) When $\delta \geq \Delta$, the platform implements the semi-decentralized mechanism with a penalty $b^* = \underline{b}$.

(2) When $\delta < \Delta$, the platform either (a) implements the semi-decentralized mechanism with a penalty $b^* = \underline{b}$, or (b) does not implement the mechanism (or, equivalently, sets $b^* \to \infty$).

The significance of Proposition 7 is twofold. First, it establishes that the best possible semi-decentralized mechanism is one with a penalty of $b^* = \underline{b}$. In particular, within the range of penalties where the mechanism is relevant to the interaction between the seller and the malicious consumer (this is the case as long as the penalty is not too high; see Proposition 5), the seller's equilibrium profit is at it's highest when the penalty is $\underline{b}$: higher values of the penalty would allow the malicious consumer to extract a higher ransom from the seller, while lower values would open up the mechanism to abuse from the seller, violating the credibility of the platforms review system. However, Proposition 7 also suggests that when the platform's judgment accuracy δ is low, it may be better for the platform not to offer the mechanism all together or, equivalently, to set the penalty high enough so that the mechanism does not affect the interaction between the seller and the malicious consumer. The latter may occur when the accuracy δ is very low, because under a mechanism with penalty $\underline{b}$ the platform's high rate of errors allows the malicious consumer to extract a higher ransom from the seller, as compared to the case where no mechanism is present (see Figure 6 for an example).

[14] We may also include a term that captures the expected penalties paid to the platform as a result of the seller's use of the decentralized mechanism; however, observe that in equilibrium this term would be equal to zero, because the seller never uses the mechanism on the equilibrium path.

5.4. Profit Implications

We now use the results of the preceding analysis to evaluate the effectiveness of the semi-decentralized mechanism in restoring the loss in profit incurred by the seller as a result of malicious consumer behavior.

Figure 6 evaluates how much of the firm's total profit loss can be recovered by the semi-decentralized mechanism (relative to the case where there is no dispute resolution mechanism in place). It is instructive to compare this plot with that of Figure 4, which conducts the same experiment for the centralized mechanism. Notice the difference in the shape of the contours: while an improvement 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 changes, the platform's equilibrium penalty b^* adjusts appropriately, so as to ensure that the seller does not abuse the mechanism, while at the same time maintaining the mechanism's relevance for the seller-consumer interaction. Furthermore, observe that, provided the judgment accuracy is sufficiently high, the semi-decentralized mechanism is able to recover most of the firm's profit loss; by contrast, for the centralized mechanism to achieve such performance, the platform's judgments must be both very quick and highly accurate, while the hassle cost associated with using the mechanism must be very low.

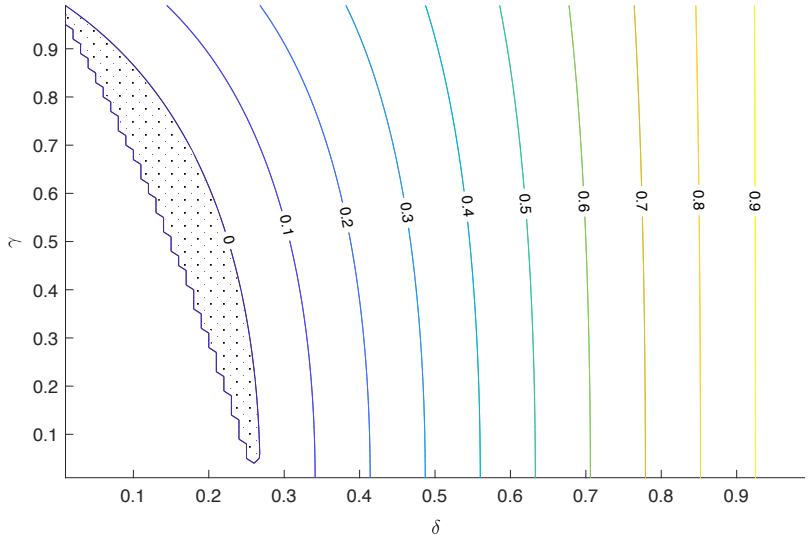

Figure 6 **Fraction of efficiency loss recovered by the decentralized dispute resolution mechanism,** $(\Pi_D^* - \Pi_{no}^*)/(\Pi_{opt}^* - \Pi_{no}^*)$**. Parameter values:** $\theta = 0.9$**,** $a = 0.5$**,** $\beta = 0.3$**,** $\pi(a) = 50a^2$**.**

To make the comparison between the two mechanisms clearer, we conduct the experiment of Figure 7 (see Appendix D for a more extensive numerical study). In particular, the experiment compares the seller's profit under the semi-decentralized mechanism against that of a centralized mechanism with a hassle cost $c = 0$ (i.e., the best possible version of the centralized version). Notice that even though the centralized mechanism in this example is costless, the semi-decentralized mechanism still dominates, with the exception of cases where the platform's judgment accuracy is very low. Furthermore, note 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 seller 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 at its largest.

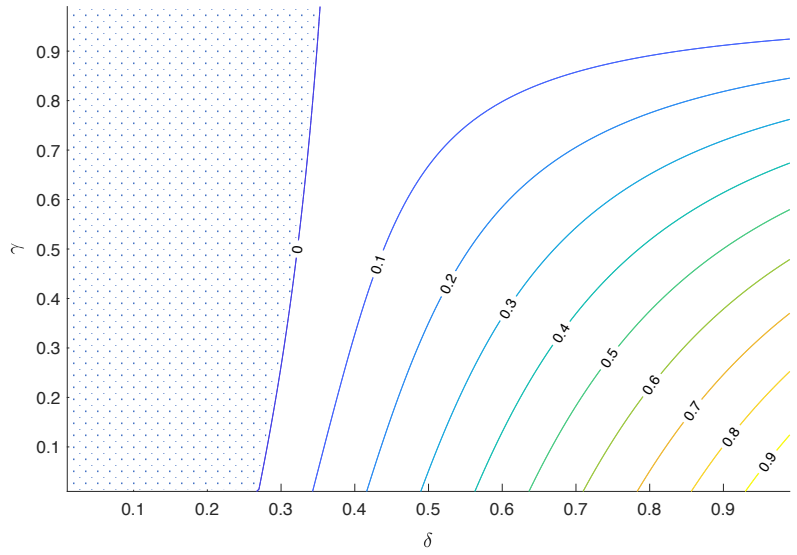

Figure 7 **Profit difference between semi-decentralized and centralized mechanisms,** $(\Pi_D^* - \Pi_C^*)/(\Pi_{opt}^* - \Pi_{no}^*)$. **Parameter values:** $\theta = 0.9$, $a = 0.5$, $\beta = 0.3$, $c = 0$, $\pi(a) = 50a^2$.

6. Model Extensions

6.1. Intermediate Degrees of Delay

In our main analysis, we have assumed that the platform's investigation occurs either immediately after the seller-consumer interaction, or is delayed until the end of the selling season. In this section, we consider cases of intermediate delay.

Let superscript $j \in \{c, d\}$ denote the type of mechanism under consideration (i.e., centralized or semi-decentralized, respectively). From the preceding analysis, if the seller makes use of the available mechanism and the platform's investigation occurs immediately, then the seller's subsequent payoff (excluding mechanism costs/penalties) under mechanism j is given by

$$\pi_{\text{no delay}} = \delta\pi(a) + (1-\delta)\pi(a_N^j). \tag{9}$$

Conversely, if the investigation is delayed until the end of the selling season, the seller's payoff is

$$\pi_{\text{full delay}} = \begin{cases} \pi(a_N^c) & \text{under the centralized mechanism, and} \\ \pi(a) & \text{under the semi-decentralized mechanism.} \end{cases}$$

The difference in the payoff expressions for the two mechanisms in the case of a delayed investigation reflects the difference in each mechanism's function: under the centralized mechanism, a delayed investigation allows the negative review to stay posted, which negatively influences the market belief to $a_N^c < a$ (see Lemma 2); under the semi-decentralized mechanism, the negative review is taken down pending the platform's investigation, and the market belief stays equal to the prior a.

To capture cases of intermediate delay, we assume that in the latter case where the investigation does not occur immediately, the seller's payoff lies between the two above extremes, that is,

$$\pi_{\text{delay}} = (1-\mu)\pi_{\text{no delay}} + \mu\pi_{\text{full delay}}, \tag{10}$$

where we have introduced the parameter $\mu \in (0, 1]$ as a measure of the delay associated with the platform's investigation. Note that the case of full delay considered in our main analysis is retrieved by setting $\mu = 1$, while smaller values of μ imply payoffs closer to immediate investigation, capturing lower degrees of delay. Under the extended model with intermediate delay, the seller now faces two possible scenarios conditional on making use of the available mechanism: (1) with probability γ, the platform's investigation occurs immediately and the seller's subsequent payoff is $\pi_{\text{no delay}}$; (2) with probability $1-\gamma$, the platform's investigation is delayed and the seller's subsequent payoff is π_{delay}, with the new parameter $\mu \in (0, 1]$ capturing the magnitude of the delay.

The analysis of the extended model follows the same approach and produces qualitatively similar results as the preceding analysis of the main model; in particular, we note that any differences in the analytical results of the two models are limited to differences in the respective algebraic expressions. Therefore, we relegate the analysis of the extended model to Appendix A and focus here on the impact of the new parameter μ. In particular, we note first that under the centralized mechanism, higher values of μ (i.e., larger delays) cause significant damage in the seller's future profit, because the delayed investigation causes a larger portion of the future market to be influenced by the negative review which remains posted pending the platform's investigation. By contrast, the

value of μ does not significantly impact the seller's future profit under the semi-decentralized mechanism, because the seller's ability to remove reviews immediately significantly limits the damage done by malicious reviews in terms of influencing subsequent consumers. Thus, from a qualitative standpoint, the impact of parameter μ on each mechanism is similar to the impact of $1-\gamma$ as seen in Figures 4 and 6 (note that μ and $1-\gamma$ are complementary measures of delay). The end result is that the difference between the two mechanisms is at its highest when μ is high and γ is low, as is demonstrated in Figure 8 below.

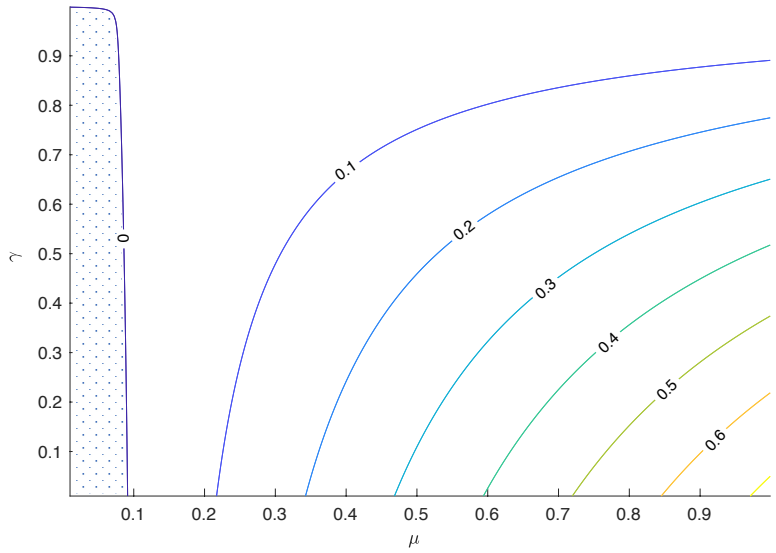

Figure 8 **Profit difference between semi-decentralized and centralized mechanisms,** $(\Pi_D^* - \Pi_C^*)/(\Pi_{opt}^* - \Pi_{no}^*)$**. Parameter values:** $\theta = 0.9$**,** $\delta = 0.8$**,** $a = 0.5$**,** $\beta = 0.3$**,** $c = 0$**,** $\pi(a) = 50a^2$**.**

6.2. Investment in Accuracy

Although we have so far assumed that the platform's judging accuracy δ is exogenous, it may be possible for the platform to improve its accuracy, especially in the long run. For instance, the platform may invest in technology that facilitates the collection of more evidence, or expend more resources in the investigation of seller claims.[15] In this section, we examine and compare the added

[15] Another way for the platform to increase the probability that malicious behavior is accurately judged as such is to lower the standard of conviction (i.e., to require a lower level of proof by the seller). However, we note that this could increase the likelihood of false convictions. Our model implicitly assumes that the platform requires the lowest level of proof for the rate of false convictions to be kept negligible, consistent with a long-standing guiding rule in legal jurisprudence known as the "Blackstone Principle" (e.g., Halvorsen 2004). The trade-off between false acquittal and false conviction rates in a platform dispute resolution setting may be a promising direction for future research.

value of improvements in judging accuracy under each type of mechanism. We begin with the structural result of Proposition 8.

PROPOSITION 8. *There exists a threshold $\bar{\Delta} \in (0,1)$ such that when $\delta \geq \bar{\Delta}$, the seller's profit under both the centralized and the semi-decentralized mechanisms is increasing in δ.*

Recall from Propositions 2 and 5 that a higher judgment accuracy δ weakens the malicious consumer's ability to extract ransom from the seller (either by reducing the equilibrium ransom demand or by deterring the malicious consumer from attempting blackmail). As a result, a higher judgment accuracy results in improved dispute resolution outcomes under both mechanisms. It is important to note, however, that while there are clear benefits to decreasing judgment errors, Proposition 8 does not account for the potential costs to the platform associated with increasing δ.

To better understand the tradeoff between the potential costs and benefits of improved adjudication outcomes, we conduct numerical experiments at different combinations of our model parameters. The example of Figure 9 highlights our main observations. The first observation is that the marginal benefit of increasing accuracy tends to be higher under the semi-decentralized, as compared to under the centralized mechanism. Importantly, this observation lends further support in favor of the cost efficiency associated with the semi-decentralized mechanism. Second, the marginal benefit of increasing accuracy is approximately constant under both mechanisms. While we have not explicitly modeled the costs associated with increasing accuracy, this observation suggests that if the platform faces increasing costs to accuracy (e.g., if the platform's cost function is convex increasing in accuracy), it may only be profitable to invest in accuracy up to a point. Third, as the platform's judgment accuracy increases, the dominance of the semi-decentralized mechanism over the centralized mechanism becomes stronger.

7. 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 non-malicious consumers is

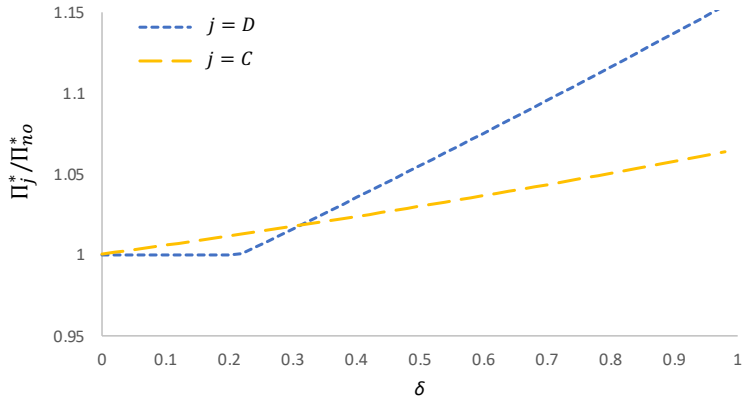

Figure 9 **Profit under the two mechanisms as a function of the judgment accuracy δ, relative to profit with no mechanism, Π_j^*/Π_{no}^*, where $j = C$ ($j = D$) denotes the centralized (semi-decentralized) mechanism. Parameter values:** $\theta = 0.9$, $a = 0.5$, $\beta = 0.3$, $\gamma = 0.5$, $c = 0$, $\pi(a) = 50a^2$.

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 this mechanism for 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 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.

In closing it is important to note that out model makes several simplifying assumptions which represent potentially fruitful avenues for future research. For example, our reduced-form model of the seller's future profit does not address repeated-interaction effects that would be present in a more detailed multi-period model. An intriguing question for future work is whether a multi-period model of the interactions between a seller and potentially malicious consumers admits steady-state equilibria and, if so, what the properties of such equilibria are. A more detailed model of sequential consumer learning may also lead to richer insights relating to the implications of malicious consumer behavior for long-term market outcomes.

Another simplifying assumption made in this work is that the seller either uses the dispute resolution mechanism made available by the platform, or deals directly with the malicious consumer to settle the dispute. In reality, sellers might employ a more complex approach, such as attempting first to use the mechanism, but later deciding to deal directly with the consumer if the platform's decision appears to be stalling. To consider such strategies, future work may incorporate more detailed models of delay and information disclosure relating to the platform's investigation process, coupled with richer models of the seller's decision problem when facing a blackmail attempts. Moreover, our analysis has focused on a single representative seller, and as such our model generates equilibrium outcomes where the platform's dispute resolution mechanism serves mainly to improve the seller's position against malicious consumers in a preemptive manner (either by deterring entry or by decreasing the malicious consumer's ability to extract ransom). In a model with multiple heterogeneous sellers, we would expect to see different sellers utilizing the mechanism differently, including both as preemptive support and as a defensive course of action following a blackmail attempt.

Future work might also consider more detailed modes of platform investigation and judgment processes. In this paper, we have assumed that the platform restricts attention to examining evidence that is directly related to potential blackmail attempts. However, platforms may have access to technology that allows for the collection of secondary information that can help improve the precision of adjudication outcomes, such as the content of reviews and the track record of individual buyers and sellers.

Acknowledgments

The authors thank the department editor, the associate editor, and three reviewers for their helpful comments. The authors also thank UC Berkeley, the London Business School, and the University of Hong Kong for their generous financial support.

References

Alizamir, S., F. De Véricourt, P. Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Babich, V., S. Marinesi, G. 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, S., R. Swinney. 2021. Signaling to the crowd: Private quality information and rewards-based crowdfunding. *Manufacturing & Service Operations Management* **23**(1) 155–169.

Chen, Li, Yiangos Papanastasiou. 2021. Seeding the herd: Pricing and welfare effects of social learning manipulation. *Management Science* **67**(11) 6734–6750.

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., A. E Frazelle, R. Swinney. 2021. Managing relationships between restaurants and food delivery platforms: Conflict, contracts, and coordination. *Management Science (Forthcoming)* .

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

Halvorsen, V. 2004. Is it better that ten guilty persons go free than that one innocent person be convicted? *Criminal Justice Ethics* **23**(2) 3–13.

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

Kanoria, Y., D. Saban. 2021. Facilitating the search for partners on matching platforms. *Management Science* **67**(10) 5990–6029.

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

Kwan, A. P., S A. Yang, A. H. Zhang. 2020. Decentralized governance on two-sided platforms: Crowdsourcing, learning, and debiasing. *Available at SSRN 3758359* .

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.

Lee, W. K., Y. Cui. 2022. Should gig platforms decentralize dispute resolution? *Working Paper, Cornell University, Available at SSRN 3719630* .

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, D., Y. Dover, J. 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. 2020. Fake news propagation and detection: A sequential model. *Management Science* **66**(5) 1826–1846.

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.

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* .

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, Angela. 2021. Decentralizing platform governance: Lawlessness, fraud and innovation. *Available at SSRN 3777697* .

Zhang, J., S. Savin, S. 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.

Appendix A: Analysis of the Extended Model with Intermediate Delay

A.1. Centralized Dispute Resolution

We begin from the seller's response to blackmail, which is described in the following proposition. Note that we use the same notation as that in §4.

PROPOSITION 9. *Suppose a malicious consumer enters the system, posts a negative review, and demands a ransom r. Then:*

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

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

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

Thus, the result mirrors Proposition 1 (i.e., the corresponding result in the main model), with the difference that γ is replaced by $1-\mu+\gamma\mu$. The same observation holds with regards to the malicious consumer's purchase and ransom strategy.

PROPOSITION 10. *The malicious consumer's equilibrium strategy is described as follows:*

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

$$r^* = \pi(a)-\pi(a_N^n). \tag{11}$$

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

$$r^* = [1-\delta(1-\mu+\gamma\mu)]\left[\pi(a)-\pi(a_N^c)\right]+c. \tag{12}$$

Finally, we note that at the pricing stage, the seller's objective function remains the same as that in the main model (i.e., as given in (3)), with the only difference being that the set of prices where a ransom demand occurs, $\mathcal{P}_C^{pur}$, is now defined by Proposition 10 (instead of Proposition 2). Therefore, Proposition 3 holds unchanged.

A.2. Semi-Decentralized Mechanism.

For the case of the semi-decentralized mechanism, we first present the following result which is analogous to Proposition 4 in the main model.

PROPOSITION 11. *Under the semi-decentralized mechanism, the seller does not remove genuine reviews if and only if $b \geq \underline{b}_I := (1-\gamma)\mu\left[\pi(a)-\pi\left(a_N^s\right)\right]$.*

The malicious consumer's equilibrium strategy (which corresponds to Proposition 5 in the main model) is then characterized as follows.

PROPOSITION 12. *The malicious consumer's equilibrium strategy is described as follows:*

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

$$r^* = \left(\pi(a)-\pi(a_N^n)\right). \tag{13}$$

(2) When $\underline{b} \le b < \frac{1-(1-\delta)(1-\mu+\gamma\mu)}{1-\delta}(\pi(a)-\pi(a_N^n))$*, the malicious consumer purchases if and only if* $p < (1-\delta)(\gamma(1-\mu)(\pi(a)-\pi(a_N^d))+b)$*. He/she then posts a negative review and demands a ransom*

$$r^* = (1-\delta)\left[(1-\mu+\gamma\mu)(\pi(a)-\pi(a_N^d))+b\right]. \tag{14}$$

Next, we note that as in the case of the centralized mechanism, the seller's problem is unchanged with respect to the main model, with the only difference being the definition of the set $\mathcal{P}_D^{pur}$, which is now determined according to Proposition 12. Finally, it is straightforward to show that the platform's optimal penalty is $\underline{b}$, for the same reasons as described in the analysis of the main model.

Appendix B: Supplemental Results

LEMMA 3. *Under the centralized mechanism, the following statements hold:*

(1) If $p_0 \ge \bar{p}_C$*, then* $p^* = p_0$*.*

(2) If $p_0 < \bar{p}_C$*, then* $p^* \in (0, \bar{p}_C]$*.*

PROPOSITION 13. *Under the semi-decentralized mechanism, in any equilibrium with* $b < \underline{b}$ *the seller uses the mechanism to remove genuine negative reviews.*

LEMMA 4. *Under the semi-decentralized mechanism, the following statements hold:*

(1) If $p_0 \ge \bar{p}_D$*, then* $p^* = p_0$*.*

(2) If $p_0 < \bar{p}_D$*, then* $p^* \in (0, \bar{p}_D]$*.*

Appendix C: Proofs

Proof of Lemma 1

When using the centralized mechanism to remove a non-malicious negative review, the seller incurs cost $c \ge 0$, 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, $a_0^i = a$ 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,

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

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,

$$a_N^c = \frac{a\left[\beta(1-\gamma\delta) + (1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right]}{a\left[\beta(1-\gamma\delta) + (1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right] + (1-a)\left[\beta(1-\gamma\delta) + (1-\beta)\left(1-\frac{p}{a\theta}\right)\right]}, \tag{15}$$

$$= \frac{a}{a + (1-a)\frac{\left(\beta(1-\gamma\delta) + (1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}{\left(\beta(1-\gamma\delta) + (1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)\right)}}. \tag{16}$$

Finally, for strategy $i = s$,

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

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,

$$\pi(a_N^s) \leq \pi(a) + \frac{c - r}{1 - \gamma\delta}. \tag{17}$$

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,

$$\pi(a_N^s) \leq \pi(a) - r. \tag{18}$$

Combining the two conditions that preclude deviation, that is, (17) and (18), strategy $i = s$ is an equilibrium if and only if

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

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

$$\gamma\delta[\pi(a) - \pi(a_N^c)] - c \geq 0. \tag{19}$$

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 (33), 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 either is an equilibrium). To identify the relevant conditions, we rearrange Proposition 1 to get the following scenarios:

1. When $c < \gamma\delta(\pi(a) - \pi(a_N^n))$, $i = n$ is not an equilibrium. 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

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

   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 $a_N^s < a_N^c < a_N^n$, $(1-\gamma\theta)(\pi(a)-\pi(a_N^s))+c > (1-\gamma\delta)(\pi(a)-\pi(a_N^c))+c$. Further, when $c < \gamma\delta(\pi(a)-\pi(a_N^n))$, we have $\pi(a)-\pi(a_N^s) > (1-\gamma\theta)(\pi(a)-\pi(a_N^s))+c$. Thus, the above condition can be simplified to

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

This corresponds to the second statement in the proposition.

2. When $c \in [\gamma\delta[\pi(a)-\pi(a_N^n)],\ \gamma\delta(\pi(a)-\pi(a_N^c)))$, 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

$$r \le (1-\gamma\delta)(\pi(a)-\pi(a_N^c))+c. \tag{20}$$

Further, in this scenario, $i=n$ is an equilibrium if and only if $r \ge \pi(a)-\pi(a_N^n)$. 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 $r \le \pi(a)-\pi(a_N^n)$. Thus, $i=s$ is the preferred equilibrium if and only if

$$r \le \min\{(1-\gamma\delta)(\pi(a)-\pi(a_N^c))+c,\ \pi(a)-\pi(a_N^n)\},$$

which is equivalent to $r \le \pi(a)-\pi(a_N^n)$ 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 (20) holds.

3. When $c \ge \gamma\delta(\pi(a)-\pi(a_N^c))$, $i=c$ is not an equilibrium because $a_N^n > a_N^c$. On the other hand, $i=n$ 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

$$r \le \pi(a)-\pi(a_N^n).$$

Combining this with the second scenario above leads to 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 $\gamma\delta \le \frac{c}{\pi(a)-\pi(a_N^n)}$ 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:

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

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

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

We prove the second point first. We show that there exists $\underline{\beta}_C$ such that if $\beta \geq \underline{\beta}_C$, then $p_0 \geq \bar{p}_C$ (and then we have $p^* = p_0$ by Lemma 3). Note first that p_0 does not depend on β. Next, note that a_N^n and a_N^c are both strictly increasing in β (see Lemma 2) and recall that $\pi'(\cdot) > 0$. It follows from Proposition 2 that $\bar{p}_C$ (i.e., the maximum price at which the malicious consumer purchases) is strictly decreasing in β. Therefore, there exists $\underline{\beta}_C \in [0,1]$ such that if $\beta \geq \underline{\beta}_C$ we have $p_0 \geq \bar{p}_C$ and $p^* = p_0$ by Lemma 3.

To prove the first point it suffices to show that if $\beta < \underline{\beta}_C$ the seller's profit function (3) is strictly increasing at any $p < p_0$ and is strictly decreasing at any $p > \bar{p}_C$. Note that according to the proof of Lemma 3, $\Pi_0(p)$ is a concave function which is maximized at p_0. Note first that in the case $\beta < \underline{\beta}_C$, we have $p_0 < \bar{p}_C$ and it follows by concavity of Π_0 that the seller's profit is strictly decreasing at any $p > \bar{p}_C$. Now consider any price $\hat{p} < p_0$. Note that both a_N^n and a_N^c are strictly increasing in p by Lemma 2. Since $\pi'(\cdot) > 0$, it follows by Proposition 2 that $\mathcal{P}_C^{Pur}$ is a connected set. Thus, for $\hat{p} < p_0$, $\hat{p} \in \mathcal{P}_C^{pur}$ the seller's profit at $\hat{p}$ can be written $\Pi_C(\hat{p}) = \Pi_0(\hat{p}) - \beta(r^*(\hat{p}) - \hat{p})$. The term $\Pi_0(\cdot)$ is strictly increasing at $\hat{p}$ by concavity. Therefore, to show that $\Pi_C(\cdot)$ is strictly increasing at $\hat{p}$ it suffices to show that $\beta(r^*(\hat{p}) - \hat{p})$ is strictly decreasing, and to show the latter it suffices to show that $r^*(\cdot)$ is strictly decreasing at $\hat{p}$. Recall that both a_N^n and a_N^c are strictly increasing in p by Lemma 2. Since $\pi'(\cdot) > 0$, it follows by Proposition 2 that $r^*(\cdot)$ is strictly decreasing at $\hat{p}$, completing the argument.

Proof of Proposition 4

Note that it is straightforward to show that the seller would never use the mechanism to remove genuine positive reviews. In what follows, we consider whether the seller uses the mechanism to remove genuine negative reviews.

For ease of reference, 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 $j = n, d$. Extending the definition a_R^i in Lemma 2, let a_R^{ij} be the market's posterior belief about the product when the review is $R \in \{P, 0, N\}$ and the market believes that the seller's strategy is ij. Further, with a slight abuse of notation, we represent cases where the malicious consumer does not purchase using $i = 0$.

There are two types of equilibria to consider: (1) the product price p is such that a malicious consumer does not purchase (this corresponds to $i = 0$) and (2) p is such that a malicious consumer does purchase and demands a ransom. In this case, we only need to consider $i = s$, that is, that the seller settles with the malicious customer by paying the ransom (that is, the other two possibilities, $i = n$ or $i = d$, cannot be part

of an equilibrium, because, anticipating such a strategy by the seller, the malicious consumer is better off by choosing not to purchase).

For each scenario, it is sufficient to prove the following two statements: (1) there exists an equilibrium such the seller does not remove genuine negative review ($ij = 0n$ for the first scenario, or $0n$ for the second scenario) if and only if $b \geq \underline{b}$; (2) For $b \geq \underline{b}$, no equilibrium exists such that the seller removes genuine negative reviews ($ij = 0d$ for the first scenario, or $0d$ for the second scenario). We note that the second statement follows directly from the proof of Proposition 13, which shows that pure strategy equilibria where the seller removes genuine negative reviews only exist for $b < \underline{b}_g < \underline{b}$. Thus, in what follows, we focus on proving the first statement.

Consider first the scenario where the malicious consumer purchases (and the seller settles, $i = s$). We establish the conditions under which strategy $ij = sn$ is an equilibrium (i.e., the seller settles with a malicious customer and does not remove a genuine negative review). For $ij = sn$ to be an equilibrium, we require that deviating to $ij = sd$ (when the market belief is that $ij = sn$) is not profitable for the seller. By considering the seller's payoff, such a deviation is not profitable provided

$$(1-\gamma)\pi(a_0^{sn}) + \gamma\pi(a_N^{sn}) - b \leq \pi(a_N^{sn}), \tag{21}$$

By the definition of a_R^{ij}, we have $a_R^{in} = a_R^i$ for $i = s, n$, as in Lemma 2. Thus, $a_0^{sn} = a$ and $a_N^{sn} = a_N^s$, and hence the above condition is equivalent to $b \geq \underline{b}$.

Next, consider an equilibrium where the malicious customer is deterred from purchasing ($i = 0$), we note that the market posterior beliefs are equivalent to the case where the malicious consumer purchases and the seller's strategy is $i = s$, that is, $a_R^{0j} = a_R^{sj}$ for $j = n, d$ and $R = P, 0, N$. It follows that the above analysis holds also for cases where the malicious consumer does not purchase, so that the condition with respect to the penalty b continues to hold.

To complete the proof, observe that $\underline{b}$ is independent of the product's price p, which implies that the condition $b \geq \underline{b}$ is necessary and sufficient for existence of an equilibrium with $j = n$. □

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:

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

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

$$\begin{aligned}&\pi(a_N^n)\ge \pi(a_0^n)-r;\\ &\pi(a_N^n)\ge (1-\gamma(1-\delta))\pi(a_0^n)+\gamma(1-\delta)\pi(a_N^n)-(1-\delta)b,\end{aligned}$$

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

$$\begin{aligned}&r\ge \pi(a)-\pi(a_N^n);\\ &b\ge \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n)).\end{aligned}$$

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

$$\begin{aligned}&(1-\gamma(1-\delta))\pi(a_0^d)+\gamma(1-\delta)\pi(a_N^d)-(1-\delta)b\ge \pi(a_N^d),\\ &(1-\gamma(1-\delta))\pi(a_0^d)+\gamma(1-\delta)\pi(a_N^d)-(1-\delta)b\ge \pi(a_0^d)-r,\end{aligned}$$

which can be simplified to

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

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

$$a_N^d=\frac{a}{a+(1-a)\frac{\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\beta\gamma(1-\delta)+(1-\beta)\left(1-\frac{p}{a\theta}\right)(1-\theta)}}. \tag{22}$$

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,

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

or, equivalently,

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

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 $r \leq \pi(a)-\pi(a_N^n)$ 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. □

Proof of Proposition 6

To prove the second point, we show that there exists $\underline{\beta}_D$ such that if $\beta \geq \underline{\beta}_D$ then $p_0 \geq \bar{p}_D$ (and then we have $p^* = p_0$ by Lemma 4). Note first that p_0 does not depend on β. Next, note that a_N^n and a_N^d are both strictly increasing in β (see Lemma 2 and (5)) and recall that $\pi'(\cdot) > 0$. It follows from Proposition 5 that $\bar{p}_D$ (i.e., the maximum price at which the malicious consumer purchases) is strictly decreasing in β. Therefore, there exists $\underline{\beta}_D \in [0,1]$ such that if $\beta \geq \underline{\beta}_D$ we have $p_0 \geq \bar{p}_C$ and $p^* = p_0$ by Lemma 4.

To prove the first point it suffices to show that if $\beta < \underline{\beta}_D$ the seller's profit function (8) is strictly increasing at any $p < p_0$ and is strictly decreasing at any $p > \bar{p}_D$. To show that, note that the function $\Pi_0(p)$ (as defined in Eq. (37) in Lemma 3) is maximized at p_0. In the case $\beta < \underline{\beta}_D$, we have $p_0 < \bar{p}_D$ and it follows by concavity of Π_0 that the seller's profit is strictly decreasing at any $p > \bar{p}_D$. Now consider any price $\hat{p} < p_0$. If $\hat{p} \notin \mathcal{P}_D^{pur}$, then the seller's profit function is strictly increasing at $\hat{p}$ by concavity of Π_0. On the other hand, if $\hat{p} \in \mathcal{P}_D^{pur}$ then the seller's profit at $\hat{p}$ can be written $\Pi_D(\hat{p}) = \Pi_0(\hat{p}) - \beta(r^*(\hat{p})-\hat{p})$. The term $\Pi_0(\cdot)$ is strictly increasing at $\hat{p}$ by concavity. Therefore, to show that $\Pi_D(\cdot)$ is strictly increasing at $\hat{p}$ it suffices to show that $\beta(r^*(\hat{p})-\hat{p})$ is strictly decreasing, and to show the latter it suffices to show that $r^*(\cdot)$ is strictly decreasing at $\hat{p}$. Observe that both a_N^n and a_N^d are strictly increasing in p by Lemma 2 and (5). Since $\pi'(\cdot) > 0$, it follows by Proposition 5 that $r^*(\cdot)$ is strictly decreasing at $\hat{p}$. □

Proof of Proposition 7

We prove the result in two steps. First, we show that among $b \leq \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n))$ (the second statement in Proposition 5), the seller's profit is the highest at $b = \underline{b}$. Second, we compare the seller's profit when the platform offers the semi-decentralized mechanism with $b = \underline{b}$ against that when the semi-decentralized mechanism is not offered (or equivalently, when it is offered with a high penalty satisfying $b > \frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n))$, such as $b \to \infty$).

For the first step, it suffices to show that under any penalty b' satisfying $\frac{1-\gamma(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n)) \geq b' > \underline{b}$, the seller's profit is no greater than that under $\underline{b}$. Let the equilibrium price under b' be p', and let the seller's equilibrium profit be $\Pi'_D := \Pi_D(p'|b')$. Furthermore, let $\Pi_D(p'|\underline{b})$ be the seller profit under the same price p', but under penalty $\underline{b}$. Note that since p' is not necessarily optimal under penalty $\underline{b}$, it follows that $\Pi_D(p'|\underline{b})$ is no greater than the seller's payoff under the optimal price at penalty $\underline{b}$. Therefore, to prove the proposition, it suffices to show that $\Pi_D(p'|b') \leq \Pi_D(p'|\underline{b})$. To show this, let $\bar{p}_D(b')$ and $\bar{p}_D(\underline{b})$ be the highest prices at which the malicious consumer chooses to enter the market under penalty b' and $\underline{b}$ respectively. Note that by Proposition 5, we have $\bar{p}_D(b') \geq \bar{p}_D(\underline{b})$. Next, in comparing $\Pi_D(p'|b')$ to $\Pi_D(p'|\underline{b})$, we have the following three scenarios depending on the relative magnitude between p', $\bar{p}_D(b')$ and $\bar{p}_D(\underline{b})$:

1. When $p' < \bar{p}_D(\underline{b})$, the malicious customer purchases under both b' and $\underline{b}$; thus, according to (8), the seller's profits under b' and $\underline{b}$ are:
$$\Pi_D(p'|b') = \beta[\pi(a) - (r^*(p'|b') - p')] + (1-\beta)\left[\frac{p}{a\theta}\pi(a) + \left(1-\frac{p}{a\theta}\right)(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s))\right];$$
$$\Pi_D(p'|\underline{b}) = \beta[\pi(a) - (r^*(p'|\underline{b}) - p')] + (1-\beta)\left[\frac{p}{a\theta}\pi(a) + \left(1-\frac{p}{a\theta}\right)(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s))\right].$$
By Proposition 5, we have $r^*(p'|b') > r^*(p'|\underline{b})$, so that $\Pi_D(p'|b') < \Pi_D(p'|\underline{b})$.
2. When $p' \in [\bar{p}_D(\underline{b}), \bar{p}_D(b'))$, a malicious customer purchases under b', but not under $\underline{b}$; thus, $\Pi_D(p'|b')$ is the same as in the above scenario, while
$$\Pi_D(p'|\underline{b}) = \beta[\pi(a)] + (1-\beta)\left[\frac{p}{a\theta}\pi(a) + \left(1-\frac{p}{a\theta}\right)(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s))\right].$$
Since $r^*(p'|b') \geq p'$, we have $\Pi_D(p'|b') < \Pi_D(p'|\underline{b})$.
3. When $p' > \bar{p}_D(b')$, a malicious customer does not purchase under either b' or $\underline{b}$; thus, $\Pi_D(p'|b') = \Pi_D(p'|\underline{b})$.

Combining the above three scenarios, we have $\Pi_D(p'|\underline{b}) \geq \Pi_D(p'|b')$. This completes the proof of the first step.

For the second step, observe that from Proposition 5 (first statement), when b is sufficiently large (which is equivalent to the platform not offering the semi-decentralized mechanism), the ransom decision, and the seller's profit function are both independent of δ. The optimal profit under this case is Π^*_{no}, as defined in §4.5. On the other hand, at $b = \underline{b}$, while b is independent of δ, under a given price p, the equilibrium ransom r monotonically decreases in δ. In particular, we note that at $\delta = 1$, we have $r^* = 0$. In this extreme, the malicious customer will not purchase (since he cannot extract any ransom), and the seller's profit will be $\beta\pi(a) + (1-\beta)\left[\frac{p}{a\theta}\pi(a) + \left(1-\frac{p}{a\theta}\right)\left(p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)\right)\right]$. The maximum profit in this case is Π^*_{opt}, which is achieved at $p^* = p_0$ (i.e., this is the "first best" profit where the malicious consumer does not purchase). Since $\Pi^*_{opt} \geq \Pi^*_{no}$, there are two relevant cases to consider. First, if $\Pi^*_{opt} > \Pi^*_{no}$, it follows by continuity of the seller's profit function in δ that there exists a threshold $\Delta < 1$ such that for any $\delta \geq \Delta$ the semi-decentralized mechanism with $b = \underline{b}$ strictly dominates not offering the mechanism (Statement 1 in the proposition), while for $\delta < \Delta$, the platform either chooses to offer the semi-decentralized mechanism with $b = \underline{b}$ (by the first step of the proof above), or not to offer the mechanism (Statement 2 in the proposition). Second, if $\Pi^*_{no} = \Pi^*_{opt}$, then it follows that not offering the mechanism weakly dominates for all values of δ (Statement 2 in the proposition). □

Proof of Proposition 8

We first consider the centralized mechanism. Note that for any $\delta > \bar{\delta}_C := \frac{c}{\gamma\left(\pi(a)-\pi(a_N^n)\right)}$, the second case of Proposition 2 holds and the equilibrium ransom is

$$r^*(\delta) = (1-\gamma\delta)[\pi(a) - \pi(a_N^c(\delta))] + c. \tag{23}$$

Here, we write r^* and a_N^c explicitly as functions of δ to highlight their dependence on δ. Next, we prove $r^*(\delta)$ decreases in δ for $\delta > \bar{\delta}_C$. To show this, we note:

$$\frac{dr^*(\delta)}{d\delta} = -\gamma[\pi(a) - \pi(a_N^c)] - (1-\gamma\delta)\pi'(a_N^c)\frac{\partial a_c^N}{\partial \delta}. \tag{24}$$

Since $\pi(a) > 0$ is convex and strictly increasing in a and $a > a_N^c$, we have $\pi(a) - \pi(a_N^c) > (a - a_N^c)\pi'(a_N^c)$. Thus, a sufficient condition for $\frac{dr^*(\delta)}{d\delta} < 0$ is

$$\gamma(a - a_N^c) > -(1-\gamma\delta)\frac{\partial a_c^N}{\partial \delta}. \tag{25}$$

By Eq. (16) in the proof of Lemma 2, we have a_c^N as:

$$a_c^N(\delta) = a - a\frac{(1-a)\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\beta(1-\gamma\delta) + \left(1 - a\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)}, \tag{26}$$

and,

$$\frac{\partial a_c^N(\delta)}{\partial \delta} = -\frac{\beta\gamma a(1-a)\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\left[\beta(1-\gamma\delta) + \left(1 - a\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)\right]^2} < 0. \tag{27}$$

Thus, Eq. (25) is equivalent to:

$$\gamma a\frac{(1-a)\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\beta(1-\gamma\delta) + \left(1 - a\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)} > (1-\gamma\delta)\frac{\beta\gamma a(1-a)\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)}{\left[\beta(1-\gamma\delta) + \left(1 - a\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)\right)\right]^2}, \tag{28}$$

which always holds because $\left(1 - a\theta(1-\beta)\left(1-\frac{p}{a\theta}\right)\right) > 0$. This establishes that r^* decreases in δ for $\delta > \bar{\delta}_c$.

Using this result, we next show that the seller's profit $\Pi_c(p)$ (Eq. 3) under any price p is non-decreasing in δ for $\delta \geq \bar{\delta}_c$. Consider any δ_1 and δ_2 such that $\delta_1 > \delta_2 > \bar{\delta}_c$. By the definition of the range of prices p that induce the malicious customer to purchase, $\mathcal{P}_C^{pur}(\delta) = \{p : p < r^*(\delta)\}$, we have, $\mathcal{P}_C^{pur}(\delta_1) \subset \mathcal{P}_C^{pur}(\delta_2)$. There are three possible scenarios depending on the magnitude of p.

1. For any $p \in \mathcal{P}_C^{pur}(\delta_1)$, the seller will pay ransom under both δ_1 and δ_2. Since $r^*(\delta_1) < r^*(\delta_2)$, we have $\Pi_c(p|\delta_1) > \Pi_c(p|\delta_2)$.
2. For any $p \in \mathcal{P}_C^{pur}(\delta_2) - \mathcal{P}_C^{pur}(\delta_1)$, the seller will only pay ransom under δ_2, but not under δ_1. Thus, $\Pi_c(p|\delta_1) > \Pi_c(p|\delta_2)$.
3. For any $p \notin \mathcal{P}_C^{pur}(\delta_2)$, the seller will not pay ransom under either δ_1 or δ_2. Thus, $\Pi_c(p|\delta_1) = \Pi_c(p|\delta_2)$.

Combining the three scenarios, we have $\Pi_c(p|\delta_1) \geq \Pi_c(p|\delta_2)$ for any price p. Therefore, we have that the seller's optimal profit increases in δ for $\delta > \bar{\delta}_c$.

The proof for the case of the semi-decentralized mechanism follows a similar structure. By Proposition 5, define $\bar{\delta}_d = 1 - \left(\gamma + \frac{b}{(\pi(a)-\pi(a_N^n))}\right)^{-1}$. For any $\delta > \bar{\delta}_d$, the second case of Proposition 5 holds and the equilibrium ransom is $r^* = (1-\delta)(\gamma(\pi(a) - \pi(a_N^d)) + b)$. By the definition of a_N^d (Eq. 22 in the proof of Proposition 5) and convexity of $\pi()$, we can show that r^* decreases in δ for $\delta > \bar{\delta}_d$ following the same procedure as above for

the centralized case. Finally, following the same argument as above, it can be shown that for any price p, the seller's profit increases in δ. Therefore, we have that the seller's optimal profit under the semi-decentralized mechanism also increases in δ for $\delta > \bar{\delta}_d$.

Combining the results for the two mechanisms, we have that for any $\delta > \bar{\Delta} := \max(\bar{\delta}_c, \bar{\delta}_d)$, the seller's profit increases in δ under both the centralized and decentralized mechanism. □

Proof of Proposition 9

We follow the same approach as in Proposition 1. First, $i = n$ is an equilibrium if and only if under the belief that $i = n$, the seller has no incentive to deviate to $i = s$ or $i = c$. We consider these two conditions in turns. First, when the seller faces a negative review and the belief is $i = n$, his net payoff by deviating from $i = n$ to $i = s$ is:

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

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

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

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

$$\Delta^{s|n} \leq 0 \iff \pi(a_N^n) \geq \pi(a) - r. \tag{29}$$

$$\Delta^{c|n} \leq 0 \iff \pi(a_N^n) \geq \pi(a) - \frac{c}{\delta(1-\mu+\gamma\mu)}. \tag{30}$$

Combining the two conditions that preclude deviation, that is, (29) and (30), strategy $i = n$ is an equilibrium if and only if

$$\pi(a_N^n) \geq \max\left\{\pi(a) - r,\, \pi(a) - \frac{c}{\delta(1-\mu+\gamma\mu)}\right\},$$

which corresponds to the first statement in the proposition.

Next, $i = s$ 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:

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

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,

$$\pi(a_N^s) \leq \pi(a) + \frac{c-r}{1-\delta(1-\mu+\gamma\mu)}. \tag{31}$$

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,

$$\pi(a_N^s) \leq \pi(a) - r. \tag{32}$$

Combining the two conditions that preclude deviation, that is, (31) and (32), strategy $i = s$ is an equilibrium if and only if

$$\pi(a_N^s) \leq \min\left\{\pi(a) - r,\ \pi(a) + \frac{c - r}{1 - \delta(1 - \mu + \gamma\mu)}\right\},$$

which corresponds to the second statement in the proposition.

Finally, 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

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

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

$$\pi(a_N^c) \leq \pi(a) - \frac{c}{\delta(1 - \mu + \gamma\mu)}. \tag{33}$$

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

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

that is,

$$\pi(a_N^c) \geq \pi(a) + \frac{c - r}{1 - \delta(1 - \mu + \gamma\mu)}.$$

Combining this condition with (33), we have that $i = c$ is an equilibrium if and only if:

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

which corresponds to the third statement in the proposition. □

Proof of Proposition 10

The proof is analogous to the 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 either is an equilibrium). To identify the relevant conditions, we rearrange Proposition 9 to get the following scenarios:

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

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

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

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

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,

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

Since $c < \delta(1 - \mu + \gamma\mu)(\pi(a) - \pi(a_N^n))$ and $a_N^s < a_N^c$, the above condition can be simplified to

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

This corresponds to the equilibrium ransom r^* in the second statement in the proposition.

2. When $c \in [\delta(1 - \mu + \gamma\mu)(\pi(a) - \pi(a_N^n)),\ \delta(1 - \mu + \gamma\mu)(\pi(a) - \pi(a_N^c)))$, 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

$$r \leq (1 - \delta(1 - \mu + \gamma\mu))(\pi(a) - \pi(a_N^c)) + c. \tag{34}$$

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 is tighter than Eq. 34, as $a_N^c < a_N^n$ and $c > \delta(1 - \mu + \gamma\mu)(\pi(a) - \pi(a_N^n))$.

3. When $c \geq \delta(1 - \mu + \gamma\mu)(\pi(a) - \pi(a_N^c))$, $i = c$ is not an equilibrium because $a_N^n > a_N^c$. On the other hand, $i = n$ 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

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

Combining this with the scenario above leads 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 Proposition 11

The proof is similar to that of Propositions 4 and 13, and it consists of two parts.

1. There exists an equilibrium such that the seller does not remove genuine review ($j = n$ using the notation in the proof of Proposition 4) if and only if $b \geq \underline{b}_I$.
2. There exists no equilibrium such that the seller removes genuine review ($j = d$) for $b \geq \underline{b}_I$.

For the first part, consider first the scenario where the malicious consumer purchases. We establish the conditions under which strategy $ij = sn$ is an equilibrium (i.e., the seller settles with a malicious customer and does not remove a genuine negative review). For $ij = sn$ to be an equilibrium, we require that deviating to $ij = sd$ (when the market belief is that $ij = sn$) is not profitable for the seller. By considering the seller's payoff, such a deviation is not profitable provided

$$(1-\gamma)\mu\pi(a_0^{sn}) + (1-\mu+\gamma\mu)\pi(a_N^{sn}) - b \leq \pi(a_N^{sn}), \tag{35}$$

Since $a_0^{sn} = a$ and $a_N^{sn} = a_N^s$, the above condition is equivalent to $b \geq \underline{b}_I$.

Next, consider the scenario where the malicious customer is deterred from purchasing, we note that the market posterior beliefs are equivalent to the case where the malicious consumer purchases and the seller's strategy is $i = s$. In particular, with a slight abuse of the notation, we may represent cases where the malicious consumer does not purchase using $i = 0$, and then we have the market posterior beliefs $a_R^{0j} = a_R^{sj}$ for $j = n, d$ and $R = P, 0, N$. It follows that the above analysis holds also for cases where the malicious consumer does not purchase, so that the condition with respect to the penalty b continues to hold.

To complete the proof of the first part, observe that $\underline{b}$ is independent of the product's price p, which implies that the condition $b \geq \underline{b}$ is necessary and sufficient for existence of an equilibrium with $j = n$.

To prove the second part, again, we first consider the scenario where the malicious customer purchases. In this case, the condition under which strategy $ij = sd$ is an equilibrium is:

$$b \leq (1-\gamma)\mu(\pi(a_0^{sd}) - \pi(a_N^s) < \underline{b}_I, \tag{36}$$

where the second inequality follows from $a_0^{sd} < a$. Thus, $ij = sd$ could not be an equilibrium for $b \geq \underline{b}_I$.

Similarly, we can also show that under the scenario where the malicious customer does not purchase, $ij = 0d$ also could not be an equilibrium for $b \geq \underline{b}_I$. This completes the proof of part 2. □

Proof of Proposition 12

By Proposition 11, under the assumption that $b \geq \underline{b}$ it suffices to focus on the strategies with $j = n$. 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 9 and 10 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

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

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:

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

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

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

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

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

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

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

which can be simplified to

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

Step 2: Equilibrium ransom. Next, we determine the equilibrium ransom r^* according to the following three scenarios:

1. When $b\le \frac{1-(1-\mu+\gamma\mu)(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n))$, $i=n$ 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,

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

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

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

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

2. When $b\in\left[\frac{1-(1-\mu+\gamma\mu)(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n)),\ \frac{1-(1-\mu+\gamma\mu)(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d))\right)$, 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,

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

   or, equivalently,

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

   Since $b>\frac{1-(1-\mu+\gamma\mu)(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^n))$ and $a_N^n>a_N^d$, we have $\pi(a)-\pi(a_N^n)<(1-\delta)[(1-\mu+\gamma\mu)(\pi(a)-\pi(a_N^d))+b]$. 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 $b > \frac{1-(1-\mu+\gamma\mu)(1-\delta)}{1-\delta}(\pi(a)-\pi(a_N^d))$, $i=d$ 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. □

Proof of Lemma 3

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

$$\Pi_0(p) = \beta\pi(a) + (1-\beta)\left(\frac{p}{a\theta}\pi(a) + \left(1-\frac{p}{a\theta}\right)[p + a\theta\pi(1) + (1-a\theta)\pi(a_N^s)]\right). \tag{37}$$

which is the seller's profit ignoring the ransom term and recall that this is a concave function which is maximized at p_0. 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 result, we note that by the definition of $\Pi_C(p)$ and $\Pi_0(p)$, $\Pi_0(p) \geq \Pi_C(p)$ for all $p \in [0,1]$. Thus, if $p_0 \geq \bar{p}_C$, by the definition of $\bar{p}_C$, we have $p_0 \notin \mathcal{P}_C^{pur}$. Consequently, $\Pi_C(p^0) = \Pi_0(p_0) \geq \Pi_0(p) \geq \Pi_C(p)$ for all $p \in [0,1]$. Therefore, $p^* = p_0$.

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

Proof of Proposition 13

Note that it is straightforward to show that the seller would never use the mechanism to remove genuine positive reviews. Thus, in what follows, we consider whether the seller uses the mechanism to remove genuine negative reviews.

For ease of reference, 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 $j = n,d$. Extending the definition a_R^i in Lemma 2, let a_R^{ij} be the market's posterior belief about the product when the review is $R \in \{P,0,N\}$ and the market believes that the seller's strategy is ij. Accordingly, we have $a_R^{in} = a_R^i$ as in Lemma 2, and using the same technique as in the proof of Lemma 2, a_R^{id} (i.e., the posterior belief assuming the seller uses the decentralized mechanism to remove non-malicious reviews) is given by: $a_P^{id} = 1$ for $i = s,n,d$, and

$$\begin{aligned} a_0^{sd} &= \frac{a}{a + (1-a)\frac{\beta + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)\right]}{\beta + (1-\beta)\left[\frac{p}{a\theta} + (1-\frac{p}{a\theta})(1-\gamma)(1-\theta)\right]}} < a; \\ a_N^{sd} &= \frac{a}{a + (1-a)\frac{1}{1-\theta}} = a_N^s; \end{aligned} \tag{38}$$

There are two possible types of equilibria to consider: (1) the product price p is such that a malicious consumer does not purchase (this corresponds to $i = 0$), and (2) p is such that a malicious consumer does purchase and demands a ransom. In the latter case, we only need to consider the strategy $i = s$, that is, that the seller settles with the malicious customer by paying the ransom (that is, the other two possibilities, $i = n$

or $i = d$, cannot be part of an equilibrium, because, anticipating such a strategy by the seller, the malicious consumer is better off by choosing not to purchase).

Now, consider first the scenario in which the malicious consumer purchases (and the seller settles, $i = s$). In this case, we show the following two results: (i) there exists $\underline{b}_g \in (0, \underline{b})$ such that the equilibrium sd exists (that is, the seller uses the mechanism to remove genuine negative reviews) if and only if $b \leq \underline{b}_g$; (ii) if $\underline{b}_g < b < \underline{b}$, no pure strategy equilibrium does not exist. To show this, we establish the conditions under which strategy $ij = sd$ is an equilibrium. For $ij = sd$ to be an equilibrium, we require that deviating to $ij = sn$ (when the market belief is $ij = sd$) is not profitable for the seller. By considering the seller's payoff, such a deviation is not profitable provided

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

where a_0^{sd}, a_N^{sd} are given in (38). Since $a_N^{sd} = a_N^s$, the last inequality is equivalent to

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

It is straightforward to verify that since $a_0^{sd} < a$, we have $\underline{b}_g < \underline{b}$, where $\underline{b}$ is defined in Proposition 4. It follows that, if $b \leq \underline{b}_g$, then there exists an equilibrium in which the malicious consumer purchases (and the seller settles, i.e., $i = s$), while if the non-malicious consumer purchases and leaves a negative review, the seller uses the semi-decentralized mechanism to remove the review (i.e., $j = d$). By contrast, if $\underline{b}_g < b < \underline{b}$, then there is no such equilibrium because the seller has an incentive to deviate from $j = d$ to $j = n$. Furthermore, note that by Proposition 4 we also have that if $\underline{b}_g < b < \underline{b}$ there is also no equilibrium such that $ij = sn$, since in this case the seller has an incentive to deviate from $j = n$ to $j = d$.

Next, consider the scenario where the malicious consumer does not purchase ($i = 0$) . In this case, we note that the market posterior beliefs are equivalent to the case where the malicious consumer purchases and the seller's strategy is $i = s$, that is, $a_R^{0j} = a_R^{sj}$ for $j = n, d$ and $R = P, 0, N$. It then follows that the above analysis holds also for cases where the malicious consumer does not purchase, so that the same conditions with respect to the penalty b hold. □

Proof of Lemma 4

The proof is analogous to that of Lemma 3. To prove the first statement in the result, we note that by the definition of $\Pi_0(p)$ and $\Pi_D(p)$, $\Pi_0(p) \geq \Pi_D(p)$ for all $p \in [0,1]$. Thus, if $p_0 \geq \bar{p}_D$, by the definition of $\bar{p}_D$, we have $p_0 \notin \mathcal{P}_D^{pur}$. Consequently, $\Pi_D(p^0) = \Pi_0(p_0) \geq \Pi_0(p) \geq \Pi_D(p)$ for all $p \in [0,1]$. Therefore, $p^* = p_0$.

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

Appendix D: Numerical Experiments

In this section, we conduct numerical experiments to evaluate the difference in the seller's payoff under the semi-decentralized versus under the centralized mechanism for dispute resolution. In each experiment, we fix the model parameters θ, a, β, c, $\pi(\cdot)$ and vary parameters δ and γ in the range $[0,1]$ in steps of 0.025. For each of the 1600 δ-γ combinations, we calculate the normalized difference $Diff = (\Pi_D^* - \Pi_C^*)/(\Pi_{opt}^* - \Pi_{no}^*)$ and report summary statistics on this difference. In the experiments presented below, we use $\theta = 0.9$, $a = 0.5$, $\beta = 0.3$, $c = 0$, and $\pi(a) = 50a^2$ as base values (i.e., unless otherwise stated in the table, these are the values at which the parameters are fixed in the experiment).

Diff	min	max	median	average
$\theta = 0.3$	-0.1518	0.9470	0.0729	0.1641
$\theta = 0.6$	-0.1536	0.9469	0.0731	0.1642
$\theta = 0.9$	-0.1358	0.9489	0.0961	0.1795
$a = 0.2$	-0.0883	0.9511	0.1357	0.2072
$a = 0.5$	-0.1358	0.9489	0.0961	0.1795
$a = 0.8$	-0.051	0.9402	0.0001	0.1514
$\beta = 0.1$	-0.0331	0.9507	0.1680	0.2318
$\beta = 0.3$	-0.1358	0.9489	0.0961	0.1795
$\beta = 0.5$	-0.2298	0.9399	-0.002	0.1169
$c = 0$	-0.1358	0.9489	0.0961	0.1795
$c = 1$	-0.0788	0.9911	0.1980	0.2656
$c = 2$	-0.0788	1.0086	0.2833	0.3172
$\pi(a) = 50a^2$	-0.1358	0.9489	0.0961	0.1795
$\pi(a) = 50a^3$	-0.0828	0.9513	0.1418	0.2122
$\pi(a) = 50a^4$	-0.0471	0.9520	0.1650	0.2305
$\pi(a) = 50a^2$	-0.1358	0.9489	0.0961	0.1795
$\pi(a) = 5000a^2$	-0.1469	1.0264	0.1039	0.1941
$\pi(a) = 500000a^2$	-0.0766	1.1677	0.1182	0.2322

Table 1 **Summary statistics for the normalized difference** $(\Pi_D^* - \Pi_C^*)/(\Pi_{opt}^* - \Pi_{no}^*)$.

Across our experiments, we observe the same qualitative pattern as that observed in Figure 7: the centralized mechanism tends to perform (modestly) better only in cases where δ is very low, while the semi-decentralized mechanism tends to perform better at most combinations of δ and γ, with a dominance that becomes more pronounced at higher values of δ and lower values of γ.[16] In addition, as Table 1 suggests, the dominance of the semi-decentralized mechanism over the centralized mechanism is greater when the performance of a high-quality product θ is higher; the prior belief about the product's quality a is lower; the probability that the seller encounters a malicious consumer β is lower; the hassle cost associated with the centralized mechanism c is higher; and the seller's future profit potential $\pi(\cdot)$ is greater and more convex in the market belief a.

[16] Note that maximum values greater than one can be observed in Table 1 in instances where the centralized mechanism results in a lower payoff for the seller as compared to no mechanism being present (see Corollary 1 and Figure 4).