Implications of Revenue Models and Technology for Content Moderation Strategies

Yi Liu, T. Pinar Yildirim, Z. John Zhang *

Abstract

This paper develops a theoretical model to study the economic incentives for a social media platform to moderate user-generated content. We show that a self-interested platform can use content moderation as an effective marketing tool to expand its installed user base, to increase the utility of its users, and to achieve its positioning as a moderate or extreme content platform. For the purpose of maximizing its own profit, a platform will balance pruning some extreme content, thus losing some users, with gaining new users because of a more moderate content on the platform. This balancing act will play out differently depending on whether users will have to pay to join (subscription vs advertising revenue models) and on whether the technology for content moderation is perfect.

We show that when conducting content moderation optimally, a platform under advertising is more likely to moderate its content than one under subscription, but does it less aggressively compared to the latter when it does. This is because a platform under advertising is more concerned about expanding its user base, while a platform under subscription is also concerned with users' willingness-to-pay. We also show a platform's optimal content moderation strategy depends on its technical sophistication. Because of imperfect technology, a platform may optimally throw away the moderate content more than the extreme content. Therefore, one cannot judge how extreme a platform is by just looking at its content moderation strategy. Furthermore, we show that a platform under advertising does not necessarily benefit from a better technology for content moderation, but one under subscription does, as the latter can always internalize the benefits of a better technology. This means that platforms under different revenue models can have different incentives to improve their content moderation technology. Finally, we draw managerial and policy implications from our insights.

keywords: social media platforms, content moderation, revenue models, technology

*Liu is a doctoral student at the University of Pennsylvania's Wharton School, Marketing Department. Yildirim is an Assistant Professor and Zhang is the Tsai Wan-Tsai Professor at the same department.

1 Introduction

A significant challenge that online social media platforms such as Facebook and Twitter face today is acting as the custodians of the Internet while at the same time being the center of self-expression and user-generated content (Gillespie, 2018). Social media platforms allow millions of users with diverse views to post their opinions on issues day-to-day, some of which are deemed offensive, harmful, or "extreme[1]," by majority of users. Users demand, on the one hand, to freely express their views on ongoing political, social, and economic issues on social media platforms without intervention and without being told their views are "inappropriate." On the other hand, they abhor the content that they themselves view as inappropriate, sensitive, harmful, or extreme. So platforms, in one form or another, moderate content to protect individual users and their interests, by removing posts they deem extreme. In fact, some executives at Facebook view content moderation as "the most important thing they do" (Lomas, 2017). In 2019, Facebook CEO Mark Zuckerberg declared that they would be allocating 5% of the firm revenues, $3.7 billion, on content moderation, an amount greater than Twitter's entire annual revenue (Roettgers, 2019). In this paper, we take a theoretical look at how a self-interested platform can do "the most important thing."

Content moderation is no simple feat. Zuckerberg (2020) states that "platforms like Facebook have to make trade-offs ... between free expression and safety" and that there is rarely a clear "right" answer. According to a *Morningconsult.com* survey[2], consumers vary on their tolerance to potentially harmful content. Of those surveyed, 80% wish to see hate speech such as posts using slurs against a racial, religious or gender group removed, 73% wish to see videos depicting violent crimes removed, and only 66% wish to see depictions of sexual acts removed. This user heterogeneity adds complexity to content moderation, but it also gives a social media platform the cover and leeway to conduct content moderation to achieve its own profit objective. This is because the balance between self-expression and safety in the context of user heterogeneity can justify any strict or lax content moderation strategy motivated by a platform's profit. In this study, we incorporate the user heterogeneity and derive a platform's optimal content moderation strategy.

The practice in the real world has shown that platforms have a wide latitude in pruning different kinds of contents that have frequently incited partisan bickering. In this paper, we shall avoid any partisan contents and focus on the contents that mostly transcend partisan politics such as hate speech, sexual or adult content, graphic violence, illegal activities, harassment, bullying, threats, etc. These are the contents that users by and large agree to be harmful and extreme, albeit they may have a

[1] Through the rest of the paper, we will refer to such content as "extreme content."

[2] Source: `https://morningconsult.com/wp-content/uploads/2019/10/190859_crosstabs_CONTENT_MODERATION_Adults_v4_JB-1.pdf`. The survey also states that 56% of adults think that edited or distorted images of public officials and celebrities should be removed by social media sites. The same number is 69% for "Misleading health information," 32% for "Fad diets, such as detoxes."

varying degree of sensitivity to them, as we have noted earlier. These are also the contents on which a vast majority of social media platforms have sworn to moderate, although their policy coverage may differ and enforcement may vary. Since the basic product offered by social media platforms is user-generated content, removal of some of these content simultaneously changes the design of the product offered to other users, and endogenously determines which users may enjoy the content enough to stay on the social media as well. Put differently, content moderation for a social media platform is a decision that simultaneously determines the content offered and the platform's positioning as a moderate or extreme content platform. It is a decision that combines "product" and "promotion" in one. In this paper, we will investigate how a platform makes that decision.

Content moderation is also a decision that attracts extensive public scrutiny and regulatory attention. How exactly content moderation should be implemented is an issue that is of high priority to policymakers, academics, and industry pundits (Jhaver et al., 2018; Schomer, 2019; Feiner, 2020), and there is a heated debate ongoing on the topic. Given the importance of content moderation to the public, it is important to understand that there can be many different worthy objectives related to the well-being of the society in conducting content moderation, such as racial harmony, national security, crime prevention, gender equality, etc. In this paper, we approach this big and complex topic from the ground up, focusing on how a self-interested platform may conduct content moderation, and take a first step toward understanding the economic foundations of a platform's desire to moderate content and to invest in content moderation technology. We study three key questions that are at the heart of marketing and management for social media platforms. First, how do self-interested social media platforms moderate content given their revenue model? More specifically, does the revenue source (advertising vs. subscription) matter in their motivation and strategy to conduct content moderation? Second, does a platform with a given revenue model always prefer a better technology for content moderation, so that they have sufficient incentives to pursue the best technology on their own? Finally, how does a free-market content moderation compare to the social optimum? The answers to these questions are important not only to understand a platform's behaviors with regard to content moderation, but also to identify the rationale for any regulatory intervention or non-intervention.

In this paper, we develop a theoretical model to address all these questions. In our model, a platform allows users to post and share content with others, and earns revenue either through advertising or subscription fees. Users enjoy the ability to express their opinions, and read others' content, from which they may or may not obtain positive utility depending on their own preferences and also on how extreme the content on the platform is. A platform moderates online content to maximize its revenue. This model setup allows us to study content moderation as a marketing tool and explore its strategy and public policy implications.

The analysis of our model shows that content moderation by a platform is primarily motivated

by users' preferences for posting vs. reading content on the platform. A platform conducts content moderation only if users care more about reading others' content than posting their own. As a marketing tool, content moderation can perform two functions for a platform: expand its user base and increase users' willingness-to-pay. The user base is expanded through pruning extreme users to get more users of moderate opinions. Content moderation can also increase users' willingness-to-pay by reducing their reading disutility from extreme content. This result establishes content moderation as an effective product design and positioning tool.

When applying this tool, a platform's optimal content moderation strategy will depend on its revenue model and also its technological sophistication. When a platform chooses optimally not to conduct content moderation, the platform under advertising will field less extreme content than a platform under subscription, all else being equal. This is because users on the advertising-based platform need not pay to join and hence the platform attracts more moderate users, while a subscription-based platform screens them out with a fee. When a platform optimally conducts content moderation, its content would be less extreme under subscription than under advertising. This is because under subscription, the platform controls the fee that marginal users pay to join and content moderation combined with a lower fee becomes a more effective tool to expand its user base. Furthermore, because of imperfect technology, the optimal content moderation strategy may call for a platform to prune the moderate content more than the extreme content on the platform and vice versa depending on the revenue model. This analysis thus suggests that the content moderation strategy for a platform, given its standard on what is extreme or not, may not be as straightforward as removing the extreme content while keeping all the moderate content.

Technology plays an important role in content moderation also in a different way. When criticized for insufficient effort at content moderation, social media executives frequently blame imperfect technology and promise to remedy the inadequacy through technology improvement (Dave, 2020; Gershgorn, 2020; Gagliordi, 2020). Interestingly, our analysis shows that for a self-interested social media platform, technological improvement does not always lead to more content moderation or to less extreme content on the platform. In addition, a platform under advertising may not even benefit from a better technology because a better technology may reduce its user base. In other words, a social media platform under advertising may not have the incentive to perfect its technology for content moderation. This result demonstrates that content moderation on online platforms is not merely an outcome of their technological capabilities, but economic incentives. This result thus casts some doubts on whether social media platforms will always remedy the technological deficiencies on their own.

The regulatory concerns are even deeper when one compares the content moderation strategy for a self-interested platform with that for a social planner. We show that a social planner will use

content moderation to prune the users whose net utility contribution to the society is negative. In addition, the social planner always pursues perfect technology if the cost of developing technology is not an issue. In contrast, a self-interested platform under either advertising or subscription is always more likely to conduct content moderation than a social planner, and when conducting content moderation, a platform under advertising (subscription) will be less (more) strict than a social planner. Only a platform under subscription will have an interest aligned with a social planner in perfecting the technology for content moderation. These conclusions thus demonstrate that there is room for government regulations and when they are warranted, they need to be differentiated with regard to the revenue model a platform adopts.

Studies in the past on user-generated content (UGC), social media, and firm strategy have taken a number of different directions. Many studies have looked into the dynamics of and the motivations for user-generated content (e.g., Toubia and Stephen, 2013; Daugherty et al., 2008; Sun et al., 2017; Iyer and Katona, 2016; Ahn et al., 2016; Bazarova and Choi, 2014; Buechel and Berger, 2015). A number of empirical and theoretical studies have also investigated how firms can glean information from UGC and use it strategically to perform their marketing functions (e.g., Ghose et al., 2012; Timoshenko and Hauser, 2019; Iyengar et al., 2011; Tirunillai and Tellis, 2012; Goh et al., 2013; Godes and Mayzlin, 2004). As the UGC provides a different dimension for firms' offerings, a number of theoretical papers have derived the optimal differentiation strategy for competing firms (e.g., Yildirim et al., 2013; Zhang and Sarvary, 2015). However, these studies do not address hate content or the issue of content moderation.

To the extent that content moderation is carried out with artificial intelligence (AI) algorithms, our research is also related to the growing stream of literature on the implications of applying AI algorithms. A number of papers have studied the application in fintech (Wei et al., 2015), hiring (Cowgill and Tucker, 2020; Lee, 2018), online dating (Abeliuk et al., 2019), and advertising (Lambrecht and Tucker, 2019). Our study differs from these papers in that we theoretically explore the strategic implications of using AI algorithms in content moderation and also the incentives for platforms to perfect their technology.

Content moderation is a hotly debated issue in political science, communications, and economics, and many of the discussions involve free speech, censorship, and the merits or demerits of content moderation (e.g., Gillespie, 2018; Myers West, 2018; Gorwa et al., 2020). In a complementary theoretical paper, Madio and Quinn (2020) study content moderation as a tool to attract content-sensitive advertisers and as a way to manage its advertising price. The content moderation we study focuses on the interactions between a platform and users and it differs from theirs in three ways. First, the user content subject to content moderation in our model is the one that all users do not like, to a varying degree. In their case, it is the content that some users like and some do not, and overall users'

demand for the platform actually goes up with more such content. Second, content moderation in our paper is motivated by a platform's effort to please and attract users, while their content moderation is motivated solely by attracting advertisers. Third, we examine how a platform's content moderation strategy interacts with technology and whether a platform under different revenue models has sufficient incentive to perfect its technology, and they do not.

The rest of the paper is structured as follows. In Section 2, we develop our theoretical model and discuss a platform's content moderation strategies with perfect technology. In Section 3, we discuss how imperfect technology can affect content moderation and what incentives a platform faces in developing a better technology. Section 4 explores the policy implications of our model. Finally, in Section 5, we conclude.

2 Model

Consider a social media platform, with users of mass of 1, where they post their opinions and read those from others. We assume that all those posts can be evaluated on a vertical scale between 0 and 1 in terms of how extreme or offensive they are. We let a user located at x be the one who expresses opinions with extremeness index $x \in [0, 1]$. Here, the vertical scale captures the fact that users agree, to a varying degree, whether a particular content is more or less extreme as measured by the index. In other words, we are modeling "vertical differentiation" in user preferences with regard to content, rather than "horizontal differentiation," where partisan users do not agree on the extremeness of a particular issue. We believe that such a vertical differentiation model is better suited for issues such as graphic violence, hate speech, bullying and threats, sexual harassment, etc., on which content moderation mostly takes place. In our conclusion section, we will discuss how a horizontal differentiation model can be relevant for future research on content moderation.

Users are heterogeneous with respect to how extreme their expressed opinions typically are. To capture this heterogeneity, we assume that users are distributed uniformly over the index range, i.e., $x \sim U[0, 1]$. We make this assumption for analytical simplicity and clarity. However, the effect of an alternative distribution will become quite clear once we understand this model with a uniform distribution. When a user posts content, as the literature has shown (e.g., Bazarova and Choi, 2014; Buechel and Berger, 2015), she gains a utility of $u(x)$ from sharing her opinions on the platform. This utility differs amongst users. The literature in consumer psychology shows that individuals with more extreme opinions are also more vocal in expressing their opinions (Miller and Morrison, 2009; Yildirim et al., 2013; Mathew et al., 2019). Based on this finding, we model the utility from posting content on social media as $u(x) = \alpha x$, where $\alpha \geq 0$, such that a user with a higher extremeness index gains more utility from posting. Here a larger α implies a greater difference in posting utilities between any

two users.[3]

A user on the platform also derives utility from reading content posted by others. Past research has suggested that a user always appreciates her own content or like-minded contents (e.g., Garimella et al., 2018; Cinelli et al., 2021). In the context of our model, this means that a user located at x reading something also located at x derives the highest reading utility, which we denote as v. However, past research has offered little direct guidance in terms of how a user may react to contents more vs less extreme than her own in a vertical context. The studies on extremeness aversion both for product choice in marketing (Simonson and Tversky, 1992; Neumann et al., 2016) and also for candidate choice in politics (Hall and Thompson, 2018; Mebane Jr and Waismel-Manor, 2005) all suggest that people have the tendency to favor more moderate choices and avoid more extreme alternatives. Absent of any study directly on social media content, extremeness aversion is a reasonable and good assumption we adopt for this paper. Specifically, in our context, extremeness aversion means that a user at x tends to feel uncomfortable about contents more extreme than x, but she may tolerate contents less extreme than x. The simplest possible way to model this asymmetry is to assume that a user at any x is troubled by the contents more extreme than x, but not at all affected by more moderate contents.[4] Algebraically, a user at x will find a post with extremeness index $\tilde{x} > x$ objectionable and her utility will be reduced by $\tilde{x}$ per post with the same index. In other words, all posts more extreme than her own will reduce her reading utility. Then, the utility for a user at x from reading the posts in the extremeness index range of $[0, \overline{x}]$ where $\overline{x} > x$ is given by $v - \int_x^{\overline{x}} \tilde{x} d\tilde{x}$. We assume a user is exposed to all content on the platform. We further assume $v < \frac{1}{2}$ to ensure that the least extreme users ($x = 0$) have a negative utility if she is exposed to all the content on the platform without moderation.[5]

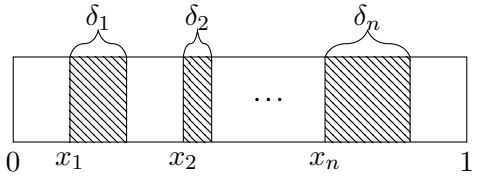
The platform can moderate the user-generated content. Due to the large volume of UGC, it typically relies on artificial intelligence (AI) and natural language processing algorithms to identify

[3]In our model, the posting utility depends only on a user's extremeness index or αx. Alternatively, the posting utility can also depend on the size of the user base on the platform X, or $\alpha x + \beta X$. Our conclusions are not affected with this extended model if β is not too large. When β is too large, multiple equilibria occur, depending on users' expectation of the platform's user base. However, in all equilibria, content moderation is less likely to happen when β is sufficiently large because the positive network effect will dominate any user expansion effect from content moderation. Detailed analysis is available upon request and we thank the Associate Editor for urging us to do this robustness check.

[4]It is conceivable that a user may be bothered by both more extreme and less extreme contents than her own, or she may be more interested in content even more extreme than her own. Future research can explore those alternative models. Ultimately, which model is more reasonable should be judged by the insights they generate and also by future empirical research findings.

[5]This formulation assumes that a user puts 100% weight on the absolute extremeness index in evaluating the disutility from a more extreme content. Alternatively, we can model the user as putting a γ% weight on the difference between the extremeness index of a more extreme content and that of her own and $(1-\gamma)$% weight on the absolute extremeness index as in our model, or $v - \gamma \int_x^{\overline{x}} (\tilde{x} - x) d\tilde{x} - (1-\gamma) \int_x^{\overline{x}} \tilde{x} d\tilde{x} = v - \int_x^{\overline{x}} (\tilde{x} - \gamma x) d\tilde{x}$. We can show that even though the complexity in analysis has increased significantly, as long as $\gamma < 1$, our analysis based on the simpler model with $\gamma = 0$ will not qualitatively change, and when $\gamma = 1$, no content moderation will take place. This is because at $\gamma = 1$, the effect of content moderation in expanding the user base disappears as content moderation does not increase enough reading utility to draw more users to the platform. We thank an anonymous reviewer for suggesting this robustness check and the complete analysis is available upon request.

Figure 1: Platform's content moderation strategy



and remove intended content. We start in our benchmark model with the assumption of a perfect content moderation technology such that a platform can get rid of any content with perfect accuracy. This means that a platform can eliminate any subset of content in the regions denoted by $\{x_i, \delta_i\}$ where $i = 1, 2, ..., n$. Here, x_i denotes the location and δ_i denotes the width of the region associated with x_i to be eliminated such that we have $0 \leq x_i \leq 1$, $x_i + \delta_i < x_{i+1}$ and $x_n + \delta_n \leq 1$, as illustrated in Figure 1. We show in Appendix on page A1, the platform will optimally choose a threshold strategy $y \in [0, 1]$ such that any content with extremeness index $x > y$ is eliminated while any content with $x \leq y$ is kept. Therefore, from this point on, our analysis focuses only on the optimal threshold strategy. Later in our analysis (Section 3), we will also look into imperfect technologies where the platform cannot perfectly moderate the intended content but can only remove any content $x > y$ and preserve any content $x \leq y$ with a higher than random probability. This imperfect technology nests our benchmark model as a special case.

If the platform engages in content moderation and deletes a user's content because it is deemed offensive, then the user experiences a psychological cost c when she could not post or her post cannot be seen by others. This cost captures how people treasure freedom of expression. Then, the total cost borne by a user at location x who is pruned by the platform is actually $c + \alpha x$ (the psychological cost plus the opportunity cost), indicating that if a user cares more about posting the content than others, then she would be more upset if her post is pruned.[6] Throughout the model, we shall maintain the assumption $c > v$ to ensure that users subject to content moderation with certainty will not participate in the platform. Without loss of generality, we set $c \leq \alpha + 2v$. This assumption is sufficient to guarantee that a user can still participate in the platform facing uncertain prospects of content moderation, as we will see in Section 3.

The anticipated utility of a user from participating in a social media platform $U(x)$ is the sum of utilities from both reading and posting content. A user participates in the platform if $U(x) \geq 0$.

[6] We thank an anonymous reviewer for suggesting this clarification.

Mathematically, $U(x)$ is given by

$$U(x) = \begin{cases} \underbrace{\alpha x}_{\text{posting utility}} + \underbrace{v - \int_{\tilde{x}\in\hat{\mathcal{X}}, x<\tilde{x}\leq y} \tilde{x} d\tilde{x}}_{\text{reading utility}} & \text{if } x \leq y, \\ \underbrace{-c}_{\text{posting utility}} + \underbrace{v}_{\text{reading utility}} & \text{if } x > y. \end{cases} \tag{1}$$

where $\hat{\mathcal{X}}$ is the expected set of participants on the platform.

Depending on whether a platform uses advertising or subscription as its revenue model, it may earn revenues from advertisers or from users through subscription fees. We focus on these two revenue models to investigate a platform's content moderation strategy because they provide distinctly different economic incentives for a platform. In the case of advertising, the platform's incentive is to maximize its user base. Under subscription, maximizing its user base will not maximize its profitability. The platform in this case needs to focus on the high willingness-to-pay users and strike a balance between the subscription fee and its user base. Then, how does the difference in economic incentives shape a platform's content moderation strategy? We shall address this question now.

Specifically, with an advertising model, the platform charges the advertisers for each user. This means that the total advertising revenue is proportional to the number of users on the platform, i.e.,

$$\pi^A = \zeta X^A,$$

where X^A is the user base of the platform, and ζ is the advertising value of each user, or ARPU (average revenue per user).[7] In some cases, the advertising rate may increase with the number of users on a platform so that we have $\zeta = \zeta_0 + \zeta_1 X^A$. We can easily show that this modification in advertising rate will not change our results.[8] We can also show that our results are robust to an advertising rate that depends on the average extremeness index on a platform ($\hat{x}$): $\zeta = \zeta(\hat{x}) = \zeta' - \omega\hat{x}$, as long as ω is sufficiently small.[9]

If the platform earns revenue from subscription instead, it sets a subscription fee p to its users. Its entire revenues will come from paying users instead of advertisers. Denote the number of users who

[7]Here we have abstracted away from the nuisance factor associated with advertising. If we were to introduce this cost for users, we can simply add a negative reading utility for all users due to advertising, which effectively is to add a negative constant to v in our model. This addition does not change our results in a substantive way. If more moderate users suffer more from advertising, our results also are not affected because that nuisance factor can be absorbed by the parameter α in our model.

[8]With the constant rate, the platform chooses its content moderation threshold y to optimize $\pi^A = \zeta X^A$. With $\zeta = \zeta_0 + \zeta_1 X^A$, the platform now maximizes $(\zeta_0 + \zeta_1 X^A)X^A$. The solution to both optimization problems is identical. As our paper focuses on content moderation, we choose the simpler model.

[9]The analysis is available upon request.

choose to use the platform when it charges p as $X^S(p)$. Then the platform's revenue is

$$\pi^S = pX^S(p).$$

In our model, the subscription fee is endogenously determined by the platform whereas the per-user advertising fee (ζ) is determined by a competitive market and is exogenous to our model.[10]

By juxtaposing these two revenue models, we can examine the incentives a platform faces in content moderation in each of the revenue models. Moreover, we can also examine how the ability of a platform to conduct content moderation may influence the choice of its revenue model. Admittedly, a platform's choice of revenue model may not depend primarily on its ability to conduct content moderation. However, given the importance of content moderation and the public nature of this activity, it is important to investigate how the ability to conduct content moderation can affect how best a platform can take advantage of a revenue model.[11] The timeline of the game is as follows:

1. If the platform uses advertising, it takes the advertising fee (ζ) as given; if it uses subscription, it sets the subscription fee (p) for all users.

2. The platform determines its content moderation strategy (y).

3. Users decide whether to stay on the platform, and those who stay on post content. Users read the content remaining on the platform after moderation, and obtain utility from posting and reading.

In the following section, we analyze the platform's content moderation strategy separately for both advertising-based and subscription-based revenue models. Then, we will discuss how content moderation can in turn affect a platform's preference for revenue models.

2.1 Advertising-Supported Social Media Platforms

We start the analysis by considering the case for an ad-supported platform. Let the users who actually participate in the platform be $\mathcal{X}$. $\hat{\mathcal{X}}$, as introduced above in Equation (1), is a user's expected set of participants on the platform. Each user will decide whether to participate in the social media platform based on her utility $U(x)$ and we derive the equilibrium where $\mathcal{X} = \hat{\mathcal{X}}$ (Katz and Shapiro, 1985; Easley et al., 2010).

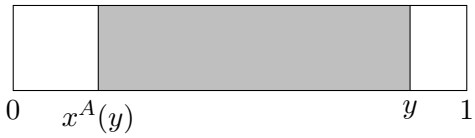
We start by characterizing the equilibrium configuration for users for any given content moderation policy y. The following lemma summarizes our analysis.

[10] As a standard practice, advertising fees on major social media platforms like Facebook and Twitter are not set by the platforms, but determined through auctions.

[11] In choosing a revenue model, business executives are clearly mindful of their ability to conduct content moderation. See `https://techcrunch.com/2020/12/22/substack-explains-its-hands-off-approach-to-content-moderation/`.

Lemma 1. *For any $y \in [0,1]$, there exists $x^A(y) \in [0,y]$ such that in equilibrium, the set of people who participate in the platform $\mathcal{X}$ is a continuum on $[x^A(y), y]$. Furthermore, $U(x)$ is increasing in x on $[x^A(y), y]$.*

Figure 2: Illustration of the platform's user base (advertising)



The proof of Lemma 1 is on p. A11 in Appendix A.2. This lemma is illustrated in Figure 2 and the platform's user base is given by the shaded area. Based on Lemma 1, we know that in equilibrium, for any $x \in [x^A(y), y]$, the utility of a user with extremeness index x from participating on the platform can be expressed as:

$$U(x) = \alpha x + v - \int_x^y \tilde{x} d\tilde{x}. \tag{2}$$

Based on the utility function, we can solve for $x^A(y)$ by setting $U(x^A(y))$ to zero, for any given level of content moderation y. The solution is given below:

$$x^A(y) = \begin{cases} -\alpha + \sqrt{\alpha^2 + y^2 - 2v} & \text{if } y \geq \sqrt{2v}, \\ 0 & \text{if } y < \sqrt{2v}. \end{cases} \tag{3}$$

Here, when $y \geq \sqrt{2v}$, we have a regime of a more lax content moderation. In this case, more content moderation will decrease $x^A(y)$, or draw more users of less extreme views to the platform. This is the case where moderating extreme views may help the platform to expand its market. When $y < \sqrt{2v}$ we have a regime of a more strict content moderation. In this case, $x^A(y)$ is bounded at zero and more content moderation will simply reduce the platform's customer base.

Thus, the revenue of a platform under advertising is given by

$$\pi^A = \zeta(y - x^A(y)), \tag{4}$$

and the platform chooses the optimal level of content moderation, y^{A*}, to maximize its revenue π^A. The following proposition summarizes the optimal content moderation strategy of a platform under advertising.

Proposition 1. ***(Advertising model and content moderation)*** *A platform with advertising as its revenue model does not always have incentives to conduct content moderation. It will conduct content moderation $y^{A*} = \sqrt{2v}$ if and only if the posting utility in the market is sufficiently small relative*

to the maximum reading utility, or $\alpha < \alpha^A \equiv \sqrt{2v}$. *The optimal revenue is given by* $\pi^{A*} = \zeta\sqrt{2v}$. *Otherwise, the platform does not moderate content* ($y^{A*} = 1$) *and its optimal revenue is given by* $\pi^{A*} = \zeta(1 + \alpha - \sqrt{\alpha^2 + 1 - 2v})$.

The proof of Proposition 1 is on page A11 in Appendix A.2. Proposition 1 suggests that content moderation is a tool for the platform to achieve its revenue objectives. When a platform uses advertising as its revenue model, it needs to maximize its user base to maximize its advertising revenues. Content moderation can help the platform to maximize its user base if cutting extreme content and pruning extreme users expand the user base amongst the less extreme users. We can see this more clearly by comparing the marginal gain and marginal loss in market size associated with content moderation.

Note that at any given y, the platform will always want to do more content moderation dy if doing so can draw more users dx^A to the platform, or algebraically $dx^A > dy$. This implies that if $\frac{dx^A}{dy} > 1$, the platform will do content moderation to expand its customer base. It is simple to show that $\frac{dx^A}{dy} = \frac{y}{\sqrt{y^2+\alpha^2-2v}}$ and it is larger than 1 if and only if $\alpha < \sqrt{2v} \equiv \alpha^A$, which is the condition given in Proposition 1.

To probe deeper, content moderation can expand the platform's customer base fundamentally because moderating extreme views and hence pruning extreme users on a platform can attract more users with less extreme opinions by increasing their reading utility. To see this, if a platform wants to expand its customer base beyond x^A to the left by Δx, the new users have a lower posting utility by $\alpha\Delta x$. These new users will only participate in the platform if their reading utility is increased by $\alpha\Delta x$. The platform can increase the reading utility for the marginal user only by further content moderation by the amount of Δy. The amount of reading utility increase by Δy is given by the expression $\Delta y \frac{\alpha y}{\sqrt{y^2+\alpha^2-2v}}$. Thus, the amount of Δy needed to increase the marginal user's reading utility by $\alpha\Delta x$, denoted as $\Delta\tilde{y}$, is given by $\Delta\tilde{y} = \frac{\Delta x\sqrt{y^2+\alpha^2-2v}}{y}$. This expression decreases with a higher v, or at a higher v a smaller change in content moderation is required to deliver the same amount of reading utility to the marginal users. This explains why at the optimal content moderation $y^{A*} = \sqrt{2v}$, the platform does less content moderation when v is large. This also explains why a platform would not do any content moderation if v is too small (the threshold $\alpha^A \equiv \sqrt{2v}$ is too small): too much content moderation to deliver too little reading utility. As a platform's objective is to maximize its customer base, it will do more content moderation as long as $\Delta\tilde{y} < \Delta x$, which also gives us the condition in Proposition 1.

Through this analysis, we can see that Proposition 1 reveals an interesting insight about content moderation across different platforms. Under the advertising revenue model, whenever users care sufficiently more about reading utility than posting utility, the platform is motivated to moderate

content. Otherwise, it is not. This proposition suggests a testable hypothesis: if users on a social media platform care more about posting than reading, then we will see little content moderation, and if they care more about reading than posting, then we will see more content moderation. However, our proposition does not rule out the possibility that if α is very large so that extreme users are vocal, the platform actually wants to moderate the content. This is because as α increases, the reading utility v may also increase with it.[12]

2.2 Subscription-Supported Social Media Platforms

When the revenue source is subscription fees, the platform determines a content moderation strategy (y) as in the case of advertising model, and also sets a subscription fee (p) for all users. Then, a user at x will participate in the platform if her net utility $U(x) - p > 0$. Similar to the advertising case, a platform's customer base is illustrated in Figure 3.

Figure 3: Illustration of the platform's user base (subscription)

The marginal user $x^S(y,p)$, which is dependent also on p now, is given by

$$x^S(y,p) = \begin{cases} -\alpha + \sqrt{\alpha^2 + y^2 - 2(v-p)} & \text{if } y \geq \sqrt{2(v-p)}, \\ 0 & \text{if } y < \sqrt{2(v-p)}. \end{cases} \tag{5}$$

The platform maximizes its subscription revenue π^S by setting its content moderation strategy y, and subscription fee p, and the revenue is given by

$$\pi^S = p(y - x^S(y,p)). \tag{6}$$

The following proposition summarizes the optimal strategy for the platform under a subscription revenue model.

Proposition 2. ***(Subscription model and content moderation)*** *Under the subscription model, there exists $\alpha^S \in (0, \alpha^A)$, such that the platform will conduct content moderation if $\alpha < \alpha^S$. The optimal content moderation strategy, equilibrium subscription fee, and the resulting revenue are given respectively by $y^{S*} = \sqrt{\frac{2v}{3}}$, $p^* = \frac{2v}{3}$, and $\pi^{S*} = (\frac{2v}{3})^{\frac{3}{2}}$. Otherwise, if $\alpha \geq \alpha^S$, the platform does*

[12] We thank an anonymous reviewer for raising this possibility.

not moderate content ($y^{S*} = 1$). The platform's optimal price and profit are given by $p^* = p_1^* \equiv \frac{1}{9}\Big[(1+\alpha)\sqrt{2(2-3v+2\alpha^2+\alpha)} - 2(1-3v+\alpha^2-\alpha)\Big]$, *and* $\pi^{S*} = p_1^*(1+\alpha-\sqrt{\alpha^2+1-2(v-p_1^*)})$.

The proof of Proposition 2 is on pp. A12-A14 in Appendix A.2. By analyzing Proposition 2, we can develop insights about what motivates a platform under subscription to do more or less content moderation, how content moderation affects its pricing, and finally how content differs under advertising vs subscription revenue models because of a platform's effort in content moderation.

By comparing Propositions 1 and 2, we see that a platform under advertising is more likely to do content moderation ($\alpha^A > \alpha^S$). This seems to be consistent with casual observations. Facebook and Twitter are two prominent examples of advertising-based platforms, and they both actively moderate content. Even though a platform under subscription is less likely to moderate its content, when it does, it will moderate content more aggressively than what it would under advertising ($y^{S*} = \sqrt{\frac{2v}{3}} < y^{A*} = \sqrt{2v}$). The platform is less likely to engage in content moderation because the subscription fee screens out less extreme users on the platform, and the remaining users are more extreme, getting less disutility from other more extreme users. For that reason, moderating extreme content adds less utility to the marginal users under subscription than to those under advertising. In other words, content moderation is less effective in attracting marginal users when the subscription model is used or $\frac{\partial x^S}{\partial y} < \frac{\partial x^A}{\partial y}$, all else being equal. This explains why content moderation is more sparingly used under subscription.

The reason why a platform under subscription may behave more aggressively once it decides to do content moderation is related to the role of pricing. Under subscription, a platform can use pricing to internalize its decision on the extent of content moderation, which is not possible under advertising. To see this clearly, we can derive how the optimal price for the platform may change with its content moderation decision, and the expression in a general form is given by

$$\frac{\partial p^*}{\partial y} = \frac{1 - \frac{\partial x^S}{\partial y} - p^* \frac{\partial^2 x^S}{\partial p \partial y}}{2\frac{\partial x^S}{\partial p} + p^* \frac{\partial^2 x^S}{\partial y^2}}.$$

The denominator of this expression is positive guaranteed by the second-order condition. Therefore, content moderation will lead to a lower price by the platform if the numerator is positive, which is the case if content moderation adds little utility to the marginal users (a small $\frac{\partial x^S}{\partial y}$), or content moderation increases price sensitivity on the part of marginal users (a large $|\frac{\partial^2 x^S}{\partial p \partial y}|$). Thus, price can help the platform to expand its user base more effectively in conjunction with content moderation so that it wants to do it more aggressively. If the numerator is negative, which is the case if marginal users are very responsive to content moderation but their price sensitivity does not change much with content moderation, the platform will increase its price with content moderation. This is the case where the platform internalizes content moderation efforts by charging a higher subscription

fee. Given our modeling assumptions, price is used to enhance the user expansion effect of content moderation.[13]

The conclusions that a platform under advertising is more likely to conduct content moderation and that when conducting content moderation, a platform under subscription does more aggressively can both be tested with suitable data. To provide some prima facie evidence, we have collected data on 103 social media platforms based on the "101+ Social Media Sites You Need to Know in 2021" composed by *Influencer Marketing Hub*.[14] As shown in Appendix A.5, we collect the texts of their content moderation policy and also information about their revenue models. In addition, we hire independent graders from Mechanical Turk to read and code the texts of content moderation policy for each platform. Our analysis shows that out of all the social media platforms in our analysis, only two platforms do not conduct content moderation and they both adopt subscription as their major revenue model. Our regression analysis further shows that the platforms with advertising as their revenue model tend to have a less restrictive content moderation policy than those with subscription (see Appendix A.5 for details). While not being conclusive, these findings are consistent with the conclusions coming out of our theoretical analysis, providing some preliminary external validity for our modeling efforts.

The first two propositions also allow us to shed light on whether content tends to be more or less extreme on a platform with subscription vs advertising model as a result of conducting content moderation. Our analysis shows that whenever a platform under subscription does not moderate content, it has more extreme content and appeals to more extreme users than a platform under advertising. However, when a platform under subscription does conduct content moderation, it fields less extreme content and caters to less extreme users than a platform under advertising. This is because subscription fee serves to screen out less extreme users when a platform does not moderate content, and when it does, as discussed previously, it uses content moderation more aggressively and charges a lower subscription fee to draw moderate users to the platform. This analysis offers a testable hypothesis that platforms under subscription tend to have the most extreme or the least extreme content.

2.3 Content Moderation and Revenue Models

The previous two sections show that a platform's revenue model will influence its content moderation strategy. In this section, we push that line of inquiry one step further to see how the ability to conduct content moderation can affect how best a platform can take advantage of a revenue model. We will do

[13] In our model, users' disutility from reading extreme content comes from all users with higher extremeness indices. This assumption, although more realistic, reduces the response of marginal users to a platform's content moderation. If we were to let a user's disutility only come from the most extreme content on the platform, we would enhance this response greatly so that the platform will want to raise its price to internalize any content moderation.

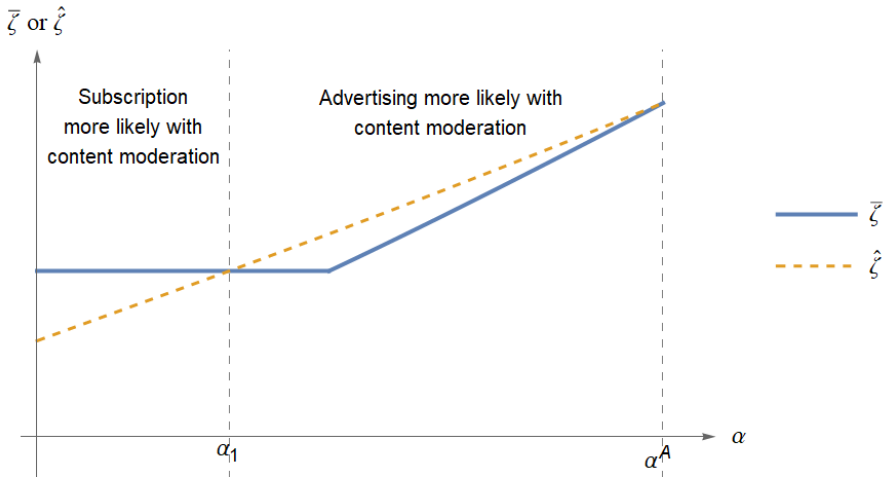
[14] `https://influencermarketinghub.com/social-media-sites/`.

so by examining when a platform may choose subscription over advertising with and without content moderation.

When content moderation is allowed, a platform will choose advertising over subscription if and only if π^{A*} is larger (smaller) than π^{S*}. This comparison will define a $\overline{\zeta}$ such that a platform will choose advertising if and only if $\zeta > \overline{\zeta}$. Here $\overline{\zeta}$ is the minimum advertising value per user needed for a platform to embrace advertising model. Thus, a larger $\overline{\zeta}$ will make it less likely for a platform to choose advertising. Similarly, when content moderation is not allowed, we can define a $\hat{\zeta}$ such that the platform will choose advertising if and only if $\zeta > \hat{\zeta}$. By comparing $\overline{\zeta}$ and $\hat{\zeta}$, we can isolate how optimal content moderation can alter a platform's preference for advertising vs subscription model. We will make the comparison for all $\alpha < \alpha^A$. For any $\alpha \geq \alpha^A$, we have the trivial case where content moderation makes no difference in the choice of revenue model because no content moderation will be conducted regardless of this choice, even if content moderation is allowed. The following proposition summarizes the findings.

Proposition 3. ***(Content moderation and revenue model choice)*** *Relative to the case of no content moderation, a platform conducting optimal content moderation is more likely to choose subscription over advertising ($\overline{\zeta} > \hat{\zeta}$) if the maximum posting utility is sufficiently small, i.e., $\alpha < \alpha_1$. Otherwise, i.e., $\alpha_1 < \alpha < \alpha^A$, optimal content moderation increases the likelihood of a platform choosing advertising ($\overline{\zeta} < \hat{\zeta}$).*

Figure 4: Revenue model choice and content moderation



The proof of Proposition 3 is on pp. A14-A16 in Appendix A.2. Proposition 3 is illustrated in Figure 4. For $\alpha < \alpha_1$, we see $\overline{\zeta} > \hat{\zeta}$ in Figure 4, implying that it takes a higher advertising rate per user for a platform to choose advertising over subscription model when content moderation is introduced. For $\alpha > \alpha_1$, we have $\overline{\zeta} < \hat{\zeta}$, which implies that a platform is willing to embrace advertising model at

a lower advertising rate when content moderation is allowed. This proposition suggests an intriguing insight that at a low α, it requires advertisers to pay a higher advertising rate to switch a platform from subscription to advertising model when content moderation is allowed. Equivalently, at a low α, content moderation makes it more likely for a platform to adopt the subscription model at a given advertising rate.

Intuitively, content moderation helps a platform under subscription more than that under advertising because without content moderation, marginal users are more sensitive to any change in maximum posting utility (α) under subscription than advertising, and hence a platform under subscription suffers more in profitability with any reduction in α. Content moderation neutralizes that effect to deliver more profit gain to a platform under subscription.

Proposition 3 suggests a testable hypothesis that in the environment where social media platforms are free to conduct content moderation vs one where platforms are constrained for one reason or another, we shall see more platforms choosing to adopt a subscription model over an advertising model if users care more about reading than posting. Otherwise, we would expect social media to use advertising model more. The variation in the extent of content moderation across Europe, US, and China may provide a good testing ground for this hypothesis.

It is important to note that our analysis on the content moderation strategy and revenue models is based on the assumption that users know the content moderation policy of the platform. This assumption captures the fact that most platforms do indeed try to publicize their content moderation policies. However, if users were to have imperfect knowledge about a platform's content moderation strategy and form their expectations about it, we can show that there will be a set of equilibria where users form their expectations independently, the platform sets its own strategies given those expectations, and users' expectations are confirmed in the respective equilibria.[15] The analysis suggests that the equilibrium we have derived is in the set of rational expectations equilibria, and it is the equilibrium that yields the maximum revenue for the platform.

3 Content Moderation and Technology

The analysis in the previous section delivers the key insights into the incentives a platform faces under advertising or subscription in using content moderation and also the impact of content moderation on how best a platform can take advantage of a revenue model. These insights are delivered under the assumption of perfect technology for content moderation. In this section, we shall expand our analysis and explore a platform's content moderation strategy under imperfect technology.

[15] We thank an anonymous reviewer and the associate editor for suggesting this robustness check. Detailed analysis is available from authors upon request.

In reality, an accurate technology for content moderation is still many years into the future (Singh, 2020). As Mark Zuckerberg commented, "over a five-to-ten-year period we will have AI tools that can get into some of the linguistic nuances of different types of content to be more accurate, to be flagging things to our systems, but today we're just not there on that... There's a higher error rate than I'm happy with" (Gershgorn, 2020). In a well-publicized example, in the days leading up to 4th of July, 2018, Facebook's algorithm for "hate speech detection" flagged down and removed a post of the Declaration of Independence because of paragraphs 27-31, which include the phrase "merciless Indian savages" (Sandler, 2018). The existence of imperfect technology raises a number of questions about the practice and management of content moderation.

First, if technology has a "higher error rate" than a platform is "happy with," how should a platform employ the technology given the choice of its revenue model? In this regard, a related question is whether a platform has the incentive to embrace an inaccurate technology to do content moderation? Second, how can a platform best manage its content moderation to achieve its profit objectives? Given that a platform's primary objective is to maximize its profit, could content moderation with imperfect technology lead to a higher extremeness index for the platform? Finally, if today's technology is "just not there," and there is "a higher error rate," what kind of a platform has the most incentives to improve it or not to improve it? In this section, we address all those questions by extending our model to incorporate imperfect technology for content moderation.

When a platform uses imperfect technology, it can err in two ways. On the one hand, it may not be able to prune the extreme content a platform wants to eliminate completely so that part of the extreme content remains on the platform. On the other hand, it may accidentally prune the content it wants to preserve. To capture both types of errors and also to nest our main model as a special case, we specify the content moderation technology $q_k(x|y)$ as the probability that a content generated by a user with extremeness index x is removed by the platform when it intends to prune all $x > y$ given its technology accuracy k. Specifically, we have:

$$q_k(x|y) = \begin{cases} \frac{1}{2} - k & \text{if } x \leq y; \\ \frac{1}{2} + k & \text{if } x > y. \end{cases} \tag{7}$$

More generally, we can specify the same probabilities for the content intended to be pruned and the content not intended to be pruned for any arbitrary content moderation strategies in the same way as in the perfect technology case. As we have shown in Appendix A.1 on page A4, the threshold strategy specified in Equation (7) dominates any other arbitrary content moderation strategies. Therefore, we focus on this threshold strategy hereafter.

The imperfect technology in Equation (7) prunes any content $x > y$ with probability $\frac{1}{2} + k$, where

$k \in [0, \frac{1}{2}]$. It also accidentally deletes any content $x < y$ with probability $\frac{1}{2} - k$. In other words, the technology allows a platform to prune extreme content with a higher probability than it deletes moderate content accidentally. When $k = \frac{1}{2}$, we go back to our main model where extreme content is cut with perfect accuracy. When $k = 0$, all content on the platform is cut with equal probability and we have a random technology at work. Thus, a higher k indicates a more accurate technology.[16] We also focus our analysis on $\alpha < \alpha^S$ such that at $k = \frac{1}{2}$ a platform always chooses to do content moderation regardless of whether it is under advertising or subscription models. Then, with this assumption, whenever a platform does not want to do content moderation, it will be due to imperfect technology. We maintain all other assumptions in the previous section. Our analysis will unfold by first looking at content moderation in advertising, then in subscription, and finally the incentives a platform faces in advancing its content moderation technology.

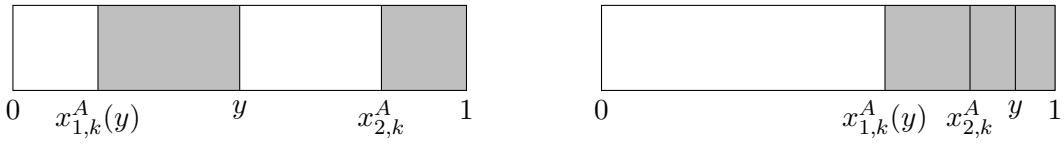
3.1 Content Moderation with Imperfect Technology

Due to imperfect technology, when a platform tries to prune all content $x > y$, the content by users with extremeness index x will be eliminated with probability $q_k(x|y)$ and it remains on the platform with probability $1 - q_k(x|y)$, as defined in equation (7). Therefore, a user's expected utility from posting is given by $\alpha x(1 - q_k(x|y)) - cq_k(x|y)$ and her utility from reading is correspondingly adjusted by the probability. We can write the total utility for a user at x as

$$U(x) = \underbrace{\alpha x(1 - q_k(x|y)) - cq_k(x|y)}_{\text{posting utility}} + \underbrace{v - \int_{\tilde{x} \in \hat{\mathcal{X}}, \tilde{x} > x} \tilde{x}(1 - q_k(\tilde{x}|y))d\tilde{x}}_{\text{reading utility}}, \tag{8}$$

which is a generalization of equation (1).

Figure 5: User base of an ad-supported platform with imperfect content moderation technology



(a) $y < x^A_{2,k}$ (b) $y \geq x^A_{2,k}$

As we show in Appendix A.3.1, whenever a platform conducts content moderation, the users on the platform fall into one of the two configurations illustrated in Figure 5. In Figure 5a, content

[16] If we were to introduce a secondary feature that the technology works more precisely as the content is far more extreme or far less extreme, the technology can be specified as $q_k(x|y) = \min\left\{\max\left\{\frac{1}{2} + k(x - y), 0\right\}, 1\right\}$, where $k \in [0, \infty)$. This model of technology is analytically intractable. However, we can numerically show that our conclusions about a platform's incentives to choose technology under our simpler model does not qualitatively change. This analysis is available upon request. We thank an anonymous reviewer for suggesting this robustness check.

moderation creates two disjoint segments of users. The appearance of these two disjoint segments is due to imperfect technology. This is because for all users subject to content moderation, it is the more moderate users that suffer the most disutility both from reading extreme content and also from their content being possibly removed. Their utility can be low enough so that they may leave the platform. In Figure 5b, we have a contiguous user segment, all dependent on the extent of content moderation y. In this figure, the variables $x^A_{2,k}$ and $x^A_{1,k}(y)$ are respectively given as:

$$x^A_{2,k} = \begin{cases} \sqrt{\alpha^2 + 1 + \frac{2c(1+2k)-4v}{1-2k}} - \alpha < 1 & \text{if } k \leq \overline{k}; \\ 1 & \text{if } k > \overline{k}, \end{cases} \tag{9}$$

where $\overline{k} = \frac{\alpha+2v-c}{2(\alpha+c)}$, and

$$x^A_{1,k}(y) = \begin{cases} \sqrt{\alpha^2 + \max\left\{0, y^2 + \min\{2\alpha y, \frac{(1-2k)(2c+1-(x^A_{2,k})^2)-4v}{1+2k}\}\right\}} - \alpha & \text{if } y < x^A_{2,k}; \\ \sqrt{\alpha^2 + \max\left\{0, y^2 + \min\{2\alpha y, \frac{(1-2k)(2c+1-y^2)-4v}{1+2k}\}\right\}} - \alpha & \text{if } y \geq x^A_{2,k}, \end{cases} \tag{10}$$

where $x^A_{1,k}(y)$ is the location for the marginal users on the platform whose content is not intended to be removed and $x^A_{2,k}$ is for those whose content is intended. Furthermore, $x^A_{2,k}$ is increasing in k and $x^A_{1,k}(y)$ is decreasing in k.

Interestingly, if the technology is not sufficiently good ($k < \overline{k}$), then we have $x^A_{2,k} < 1$, which means the most extreme users in $[x^A_{2,k}, 1]$ will stay on the platform regardless of how the platform conducts its content moderation, as extreme users always derive the highest utility from the platform. In other words, inaccurate technology can no longer screen out the most extreme users, a fact about imperfect technology that a platform under advertising can benefit from, as we will see soon.

Similar to the advertising case, when a platform adopts subscription, we can show in Appendix A.3.2 that we have similarly well-defined user configurations as in Figure 5. As in Section 2.2, the platform once again chooses its moderation strategy and subscription price. We refer readers to Appendix A.3.2 for detailed analysis of this case.

A social media platform at any given point in time typically sets a clear standard about what is extreme and what is not extreme, or what is allowed and what is not allowed on the platform. In addition, users on the platform also know and are frequently reminded of the standard the platform uses. However, even though the standard is clear, with imperfect technology, it is no longer the case that a platform can remove what it deems extreme with perfect accuracy. This means that the content removed from the platform includes those intended as well as unintended. We will refer to the content that the platform intends to remove based on its own standard as the "extreme" content, and the content that the platform does not intend to remove as the "moderate" content. Depending on the

technology it uses (for a given k), the platform may prune the extreme content more than the moderate content and vice versa. Thus, the first question to ask is: to maximize its profit, should a platform always prune the extreme content more than the moderate content? From a user's perspective, we can ask the second question: when a platform prunes the moderate content more than the extreme content, does the platform always have a high average extremeness index?[17] In other words, could users conclude based on what is thrown out from a platform whether the platform is on average more or less extreme? This question should also be of interest to regulators, policymakers, and consumer advocacy groups.[18] Our analysis provides answers to both questions. The following proposition summarizes a platform's optimal content moderation strategy with imperfect technology.

Proposition 4. ***(Content moderation with imperfect technology)*** *For both advertising and subscription, a platform will conduct content moderation only if technology is sufficiently accurate. When conducting content moderation, the platform may prune the moderate content more than the extreme content, but the average extremeness index on the platform may be lower than when it prunes the extreme content more or does not prune any content.*

The proof of Proposition 4 can be found on pp. A16-A18 in Appendix A.2 (part of the proof is based on exhaustive numerical analyses). Proposition 4 first suggests that a platform needs a sufficiently accurate technology to start content moderation. Secondly, when the technology is sufficiently good, the optimal content moderation strategy may call for a platform to prune the moderate content more than the extreme content. This is because a sufficiently accurate technology already deters extreme users from participating in the platform but encourages more moderate users based on the platform's standard to participate, so there is less extreme content on the platform in the first place. Finally, whether a platform prunes extreme content more than the moderate content is not a good yardstick to judge whether a platform is extreme or moderate. This means that a user looking to join a platform may not find a moderate outlet even if the outlet is pruning a lot of extreme content. This may be because there are many extreme users on the platform in the first place.

However, the case where pruning the moderate content may lead to a low extremeness index deserves a closer look. The optimal content moderation strategy calls for pruning the moderate content more than the extreme content when there is little extreme content on the platform in the first place. Then the question is, why prune moderate content if there is little extreme content on

[17] The average extremeness index on the platform can be calculated as $\hat{x} = \frac{\int_{\mathcal{X}} x(1-q(x))dx}{\int_{\mathcal{X}} (1-q(x))dx}$, where the numerator is the weighted sum of the location index (weighted by the probability of not being removed), which represents the "total" extremeness of all remaining content, and the denominator is the expected number of posts remaining after content moderation.

[18] For instance, the EU does pay attention to the content pruned from a platform, and regularly publishes Evaluation of the Code of Conduct on Countering Illegal Hate Speech Online (see `https://ec.europa.eu/info/sites/default/files/codeofconduct_2020_factsheet_12.pdf`).

the platform? The reason is strategic. With a blunt instrument or imperfect technology, pruning the moderate content is the collateral damage to pruning the extreme content, or the price a platform pays to reduce extreme content. Thus, it may be necessary for a platform to prune only the moderate content in order to deter extreme users from ever getting onto the platform.

Proposition 4 suggests two managerial as well as policy insights about content moderation. First, no one should be alarmed about a platform pruning moderate content or not eliminating extreme content, and it is part of a platform's optimal strategy when technology is imperfect. For this reason, we may see more social media executives blaming technology. Second, the content moderation strategy by a platform and the diligence with which it is pruning the extreme content may not tell the full story about how extreme the content may be on the platform. To tell the full story, one will have to also consider the technology used and the preferences of the user base.

3.2 Content Moderation and Incentive for Technology Improvement

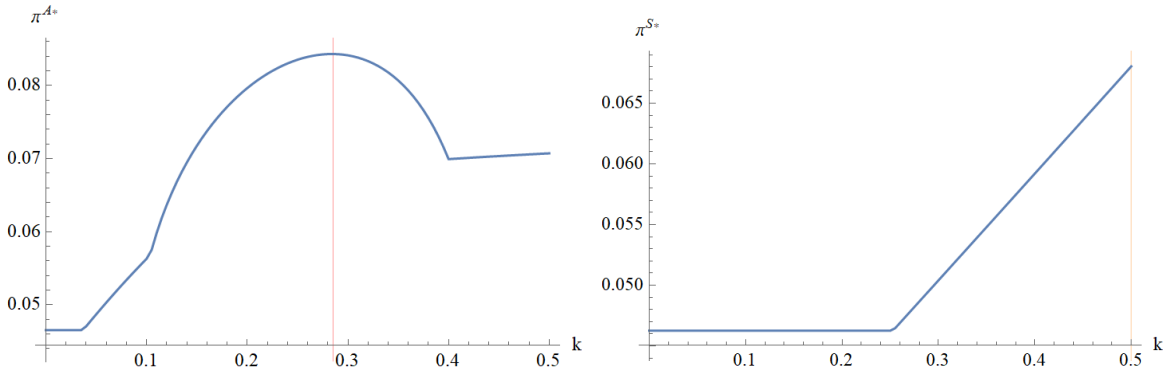
As the technology improves, a platform's strategy in content moderation will also change. In this regard, our model of imperfect technology allows us to shed light on two related questions. First, will a platform impose a more strict standard for content moderation when technology improves? Second, as the content moderation strategy of a platform also affects its profitability, does a platform actually have an incentive to improve the technology? The following two propositions suggest some nuanced answers to these two questions.

Proposition 5. ***(Better technology and less content moderation)*** *When technology is sufficiently accurate (a sufficiently large k), a platform under either advertising or subscription will adopt a more relaxed standard for content moderation as the technology further improves. As a result, the average extremeness index increases.*

The proof of Proposition 5 is on p. A18 in Appendix A.2 (part of the proof is based on exhaustive numerical analyses). Intuitively, as the technology improves, the platform can prune extreme content more accurately to keep moderate marginal users happy so that it does not need to prune as much. In addition, by pruning the extreme content less, the platform can increase its customer base to increase its profit when it is under advertising, and keep more of its high willingness-to-pay users on the platform when it is under subscription.

Proposition 6. ***(Incentive for imperfect content moderation technology)*** *Under advertising, the platform may choose imperfect technology even if there is no cost involved in improving the technology when the cost to users subject to pruning (c) is small. Under subscription, the platform always chooses a perfect technology.*

Figure 6: Platform profit and technology accuracy ($v = 0.25, \alpha = 0.2, c = 0.3$)



(a) Advertising ($\zeta = 0.1$) (b) Subscription

The proof of Proposition 6 is on p. A19 in Appendix A.2 (part of the proof is based on exhaustive numerical analyses), and the proposition is illustrated in Figure 6. A platform under advertising may not want to develop a perfect technology because its primary objective is to maximize its customer base. When c is small, a less accurate technology will increase the number of extreme users more than it reduces the number of moderate users, thus increasing the installed customer base for the platform. This is because a less accurate technology offers more benefits to extreme users than the loss it imposes on moderate users when technology is accurate in the first place. However, too much extreme content on the platform will alienate moderate users. This effect becomes increasingly dominant when technology is at lower accuracy (smaller k). This explains why in Figure 6a we have an inverted U-shaped relationship between accuracy and platform profit under advertising. When c is sufficiently large, the segment of moderate users is also sufficiently large relative to the segment of extreme users because the cost carries more weight for the extreme users as they have a higher probability of being pruned. In this case, a less accurate technology can still increase the segment of extreme users but it will impose unintended damage on the relatively large segment of moderate users. Therefore, the most effective way to increase the installed customer base is not to increase extreme users but to retain and expand moderate users by reducing the likelihood of unintended pruning, which is to increase the accuracy (k). This is why when c is sufficiently large, even a platform under advertising will be motivated to pursue the perfect technology.

In the case of subscription, however, maximizing a platform's customer base can no longer maximize the platform's profit as the platform has the subscription fee as the second instrument. With this second instrument, the platform can fully internalize the benefit of technology improvement. This is evident from the fact that both the platform's customer base and the optimal fee increase with technology improvement. Therefore, costs aside, the platform always has the incentive to pursue the perfect technology.

Propositions 5 and 6 offer two rather surprising perspectives on content moderation and technology. First, as content moderation technology becomes more accurate, one should not expect that a profit-maximizing platform will always do more content moderation and curate more moderate content. Second, a number of executives of large social media platforms, including those of Facebook and Twitter (Dave, 2020; Gershgorn, 2020), often complain about the limits of technology in content detection. However, our analysis also suggests an intriguing possibility that a platform under advertising may not have the incentive to pursue a more accurate technology in the first place.

These two perspectives suggest to a manager that conducting content moderation is not as simple as just reducing the extremeness index. In the pursuit of a better technology, the optimal strategy calls for a manager to relax the criteria for pruning and to increase the average index. Moreover, the cost to users subject to pruning is an important parameter to watch. When that cost is low, imperfect technology is conducive to attracting a large installed customer base to the platform. When it is large, technology improvement is always a winning strategy.

One testable hypothesis from Propositions 5 and 6 is that when conducting content moderation, all else being equal, we will expect to observe that platforms under subscription have a better technology for content moderation than those under advertising.

4 Content Moderation and Policy Implications

Content moderation is a hotly debated issue that has many policy implications. Many questions are raised in this context. For instance, do platforms have sufficient incentives to conduct content moderation on their own relative to what is optimal for users? When they do conduct moderation, are they doing too much or too little? Is the technology that is optimal for platforms also optimal for users? We can address all these questions with our model by investigating how a social planner with user interest at heart will conduct content moderation to maximize user welfare. Our answers can then inform the ongoing debate on whether and how much the government should get involved in regulating online content if it wants to advance users' interests and how the regulatory effort may need to be nuanced with regard to platforms of different revenue models.

To conduct our analysis, we note that the objective of a social planner in content moderation is to maximize user welfare, which is the sum of the utilities for all users for the platform, and the expression of the user welfare, denoted as $W(y)$, when technology is perfect is given by

$$W(y) = \int_{x^P(y)}^{y} \left(\alpha x + v - \int_{x}^{y} \tilde{x} d\tilde{x} \right) dx, \tag{11}$$

where $x^P(y)$ is the marginal user who is indifferent between participating in the platform and not.[19] As we show in the proof of Proposition 7, the social planner will conduct content moderation if and only if $\alpha < \alpha^P$, where at α^P the social planner is indifferent between conducting content moderation and not. When the social planner does conduct content moderation ($\alpha < \alpha^P$), the optimal content moderation strategy is given by $y^{P*} = \frac{1}{2}(\alpha + \sqrt{\alpha^2 + 4v})$. By comparing what the social planner does with what a platform under advertising or subscription does, we have the following proposition.

Proposition 7. ***(Social planner's content moderation strategy)*** *All else being equal, a social planner is less likely to conduct content moderation than a platform under either advertising or subscription ($\alpha^P < \alpha^S < \alpha^A$). When it does, it adopts a more relaxed standard for content moderation than a platform under subscription, but a more strict one than a platform under advertising ($y^{S*} < y^{P*} < y^{A*}$).*

The proof of Proposition 7 is on p. A20 in Appendix A.2. This proposition suggests three insights about the content moderation strategy in a decentralized market. First, left to market forces, a platform with the profit motivation has even more incentives to engage in content moderation than a social planner with user welfare as its objective. This is because content moderation is less effective at increasing user welfare than at increasing a platform's profitability. Extreme users contribute positively to the user welfare so that the social planner will be more inclusive. Second, more incentives for a platform do not mean right incentives. To maximize the user welfare, the social planner only prunes users with a negative utility contribution to the society.[20] A user's utility contribution to the society includes her posting utility, reading utility, and the total negative utility her post imposes on other more moderate users. A platform under advertising will keep users with negative utility contribution, all for the purpose of maximizing its installed customer base. A platform under subscription will prune users even with positive utility contribution to increase the willingness-to-pay of other users. Third, because of the different incentives that platforms with different revenue models have in content moderation and also because of the severity with which different platforms are motivated to prune content for their own profitability, any regulatory measures may need to account for the difference in platforms' revenue models. In other words, sweeping regulations for all social media platforms for the purpose of advancing users' interests regardless of their revenue models could be ill-advised.

Indeed, revenue models also provide different incentives for a platform to perfect its technology. A natural question arises: what technology a social planner would prefer for content moderation, perfect or imperfect? The following proposition addresses this question.

[19] It is important to note that this user welfare function is the same as the social welfare function in the case of subscription because of the payment a platform receives is a transfer payment from users.

[20] This is checked in Appendix on page A23.

Proposition 8. ***(Social planner's technology preference)*** *When a social planner conducts content moderation, it always prefers a better technology (higher accuracy k) such that, cost aside, it always pursues the perfect technology ($k = \frac{1}{2}$).*

The proof of Proposition 8 is on p. A23 in Appendix A.2 based on exhaustive numerical analyses. Comparing Proposition 8 with Proposition 6, we see that a platform under advertising does not always have the right incentives to perfect its technology, unless c is sufficiently large. However, a platform under subscription does have the right incentive to develop the technology for content moderation, although the technology is not applied in a way that is maximizing user welfare, as discussed in Proposition 7.

5 Conclusion

Content moderation on social media platforms is an important issue that has attracted increasing attention in the past few years from practitioners, scholars, social activists, policy makers, and regulators alike. At a high level, the issue concerns the freedom of expression, political discourse, personal liberty, civil society, and government regulations. At a more basic level, it is a platform's marketing decisions, like any other product or service company would do, on what revenue models to use, what content to allow or what "product" to design, and what kind of users to attract or to discourage, all for the purpose of achieving its highest revenues. In addressing this complex issue, it is quite understandable that experts with different objectives offer different perspectives as to whether platforms should do self regulation themselves, or a government intervention is needed to regulate social media content. In this paper, we take a first step to unpack this complex issue and investigate how a self-interested social media platform may conduct content moderation, how its content moderation strategy may hinge on its revenue model and technology, and what incentives a platform with advertising or subscription as its primary revenue model may have in perfecting content moderation technology. This investigation not only offers normative insights about how a self-interested platform will or will not do content moderation, but also sheds light on whether government interventions are needed and if they are, what those interventions may entail.

Our analysis shows that a self-interested platform does not need to care about any social cause to actively engage in content moderation. It can use content moderation as a tool to perform two marketing functions: to expand its user base and to increase the willingness-to-pay of the users on its platform. These dual functions are rooted in the nature of social media where users gain utilities from posting and reading user-generated content on a platform, but they are also sensitive to content more extreme than what they prefer. For a social planner who cares about user welfare, content moderation is a tool to eliminate users who make negative utility contributions to society. In this

regard, we show that self-interested platforms are more likely to use the dual functions and conduct content moderation than a social planner. In other words, platforms are more eager than a social planner to conduct content moderation motivated by their own self-interest.

Because a self-interested platform conducts content moderation for profit, the economics dictate that its strategy will depend on its revenue model and hence the resulting content on the platform, as measured by the extremeness index, will also depend on the same. We show that in the absence of any content moderation, all else being equal, a platform under subscription revenues will field more extreme content than a platform under advertising. However, when content moderation is conducted, a platform under subscription revenues will curate a more moderate content than one under advertising. Interestingly, the social planner will conduct content moderation to achieve a body of content that is more extreme than under subscription, but more moderate than under advertising.

For most social platforms, technology for content moderation is imperfect as many executives have readily admitted (Dave, 2020; Gershgorn, 2020). Our analysis shows that a platform's strategy in content moderation critically depends on the technology it uses. A platform may choose not to do any content moderation at all if its technology is not sufficiently accurate. When it is, a platform may conduct content moderation in an unexpected way. Under imperfect technology, a platform may throw away the moderate content more than the extreme content as part of its optimal strategy. We show that when this happens, it does not necessarily result in a more extreme platform. Conversely, when a platform prunes the extreme content more than the moderate content, we do not necessarily have a more moderate platform. In other words, one cannot judge how extreme a platform is by looking at its content moderation strategy. This insight is especially germane to policy makers when they try to reduce hate content on a platform by focusing on the removal of hate content upon user complaints, such as what is currently practiced in EU (Reynders, 2020).

It is common for social media executives to blame imperfect technology for some lapses in content moderation, and those blames are well-placed, as our analysis shows. However, our analysis also sheds some light on whether a self-interested platform actually has incentives to perfect its content moderation technology. A platform under advertising may not pursue the perfect technology, even if doing so is costless. We further show that a platform under subscription will pursue the perfect technology, as does a social planner. Overall, our analysis shows that self-interested platforms are motivated to do content moderation, but their strategy diverges from a social planner's. In this sense, there can be grounds for government interventions. We show that such interventions can only be effective if they are differentiated and nuanced according to revenue models and technology levels that different platforms are adopting.

As managerial insights, our analysis has articulated the marketing roles that content moderation plays in achieving a platform's profit objectives. It also prescribes the normative strategies that

a platform can use in content moderation: what content to moderate for what purpose and what strategic adjustments to make regarding revenue models and technology. Finally, platforms under subscription are well advised to invest in their technology for content moderation.

Content moderation as a research topic is a target-rich area. We hope our research kindles some interest in this important and timely subject. Future research can take a number of directions. First, in our model, we assume a uniform distribution for users over the extremeness index. This enables us to conduct our analysis with clarity and gains a good intuition about what content moderation strategy helps a platform to do. In reality, it is conceivable that users who hold extreme views are probably in the minority. We venture to suggest, based on our analysis, that the platform should be more willing to prune more extreme content since there were fewer users to prune. Future research can extend our analysis to different distributions, such as a normal distribution.[21] Second, our model is based on a vertical differentiation model, which applies to many different kinds of content that are currently subject to moderation. Future research can extend this analysis to perhaps a combination of vertical and horizontal models. Such a model can be suitable for political issues where partisans agree within the group but disagree between groups. These are the types of issues that our model does not address. Third, as a first paper on content moderation strategies, we have abstracted away from the possibility of strategic users. These are the users who may engage in self-censorship and who may change the content they post because of a platform's content moderation strategy. We believe that such a strategic behavior can reduce the cost of content moderation for the platform and may encourage more content moderation. Fourth, in our model, we identify the difference between what a social planner will do with content moderation and what a self-interested platform will do, thus probing into the rationale for and approach toward any regulatory interventions. Future research can develop concrete regulatory measures that can induce platforms under advertising or subscription to conduct content moderation in alignment with a social planner. Lastly, many of our theoretical insights are empirically testable. Future research can put them to a test with suitable data.

References

Abeliuk, A., Elbassioni, K., Rahwan, T., Cebrian, M., and Rahwan, I. (2019). Price of anarchy in algorithmic matching of romantic partners. *arXiv preprint arXiv:1901.03192*.

Ahn, D.-Y., Duan, J. A., and Mela, C. F. (2016). Managing user-generated content: A dynamic rational expectations equilibrium approach. *Marketing Science*, 35(2):284–303.

Bazarova, N. N. and Choi, Y. H. (2014). Self-disclosure in social media: Extending the functional

[21] We thank the Associate Editor and an anonymous reviewer for suggesting these alternative distributions.

approach to disclosure motivations and characteristics on social network sites. *Journal of Communication*, 64(4):635–657.

Buechel, E. C. and Berger, J. (2015). Motivations for consumer engagement with social media. In *Consumer Psychology in a Social Media World*, pages 31–50. Routledge.

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9).

Cowgill, B. and Tucker, C. E. (2020). Algorithmic fairness and economics (February 14, 2020). *Available at SSRN*. https://ssrn.com/abstract=3361280 or http://dx.doi.org/10.2139/ssrn.3361280.

Daugherty, T., Eastin, M. S., and Bright, L. (2008). Exploring consumer motivations for creating user-generated content. *Journal of Interactive Advertising*, 8(2):16–25.

Dave, P. (2020). Social media giants warn of AI content moderation errors, as employees sent home. World Economic Forum. https://www.weforum.org/agenda/2020/03/social-media-giants-ai-moderation-errors-coronavirus/. Accessed December 27, 2020.

Easley, D., Kleinberg, J., et al. (2010). *Networks, Crowds, and Markets*. (Vol. 8). Cambridge: Cambridge University Press.

Feiner, L. (2020). Biden tech advisor: hold social media companies accountable for what their users post. CNBC. https://www.cnbc.com/2020/12/02/biden-advisor-bruce-reed-hints-that-section-230-needs-reform.html. Accessed December 27, 2020.

Gagliordi, N. (2020). Facebook says AI enhancements have bolstered its content moderation efforts. ZDNet. https://www.zdnet.com/article/facebook-says-ai-enhancements-have-bolstered-its-content-moderation-efforts/. Accessed December 27, 2020.

Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922.

Gershgorn, D. (2020). Mark Zuckerberg just gave a timeline for AI to take over detecting internet hate speech. Quartz. https://qz.com/1249273/facebook-ceo-mark-zuckerberg-says-ai-will-detect-hate-speech-in-5-10-years/. Accessed December 27, 2020.

Ghose, A., Ipeirotis, P. G., and Li, B. (2012). Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Science*, 31(3):493–520.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. New Haven: Yale University Press.

Godes, D. and Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, 23(4):545–560.

Goh, K.-Y., Heng, C.-S., and Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information Systems Research*, 24(1):88–107.

Gorwa, R., Binns, R., and Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Hall, A. B. and Thompson, D. M. (2018). Who punishes extremist nominees? candidate ideology and turning out the base in us elections. *American Political Science Review*, 112(3):509–524.

Iyengar, R., Van den Bulte, C., and Valente, T. W. (2011). Opinion leadership and social contagion in new product diffusion. *Marketing Science*, 30(2):195–212.

Iyer, G. and Katona, Z. (2016). Competing for attention in social communication markets. *Management Science*, 62(8):2304–2320.

Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. (2018). Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25(2):1–33.

Katz, M. L. and Shapiro, C. (1985). Network externalities, competition, and compatibility. *The American Economic Review*, 75(3):424–440.

Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684.

Lomas, N. (2017). Facebook's content moderation rules dubbed 'alarming' by child safety charity. TechCrunch. `https://techcrunch.com/2017/05/22/facebooks-content-moderation-rules-dubbed-alarming-by-child-safety-charity/`. Accessed December 27, 2020.

Madio, L. and Quinn, M. (2020). User-generated content, strategic moderation, and advertising (November 6, 2020). *Available at SSRN*. `https://ssrn.com/abstract=3551103` or `http://dx.doi.org/10.2139/ssrn.3551103`.

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. (2019). Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Mebane Jr, W. R. and Waismel-Manor, I. S. (2005). Does it help or hurt kerry if nader is on the ballot? In *Annual Meeting of the Midwest Political Science Association, Chicago, IL, April*. Citeseer.

Miller, D. T. and Morrison, K. R. (2009). Expressing deviant opinions: Believing you are in the majority helps. *Journal of Experimental Social Psychology*, 45(4):740–747.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Neumann, N., Böckenholt, U., and Sinha, A. (2016). A meta-analysis of extremeness aversion. *Journal of Consumer Psychology*, 26(2):193–212.

Reynders, D. (2020). Countering illegal hate speech online: 5th evaluation of the code of conduct. `https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf`. Accessed December 27, 2020.

Roettgers, J. (2019). Mark Zuckerberg says Facebook will spend more than $3.7 billion on safety, security in 2019. Variety. `https://variety.com/2019/digital/news/facebook-2019-safety-speding-1203128797/`. Accessed December 27, 2020.

Sandler, R. (2018). Facebook has apologized for flagging parts of the Declaration of Independence as hate speech. Business Insider. `https://www.businessinsider.com/facebook-declaration-of-independence-hate-speech-2018-7`. Accessed December 27, 2020.

Schomer, A. (2019). The content moderation report: Social platforms are facing a massive content crisis – here's why we think regulation is coming and what it will look like. Business Insider. `https://www.businessinsider.com/content-moderation-report-2019-11`. Accessed December 27, 2020.

Simonson, I. and Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research*, 29(3):281–295.

Singh, S. (2020). AI proves it's a poor substitute for human content checkers during lockdown. Venture Beat. `https://venturebeat.com/2020/05/23/ai-proves-its-a-poor-substitute-for-human-content-checkers-during-lockdown/`. Accessed December 27, 2020.

Sun, Y., Dong, X., and McIntyre, S. (2017). Motivation of user-generated content: Social connectedness moderates the effects of monetary rewards. *Marketing Science*, 36(3):329–337.

Timoshenko, A. and Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science*, 38(1):1–20.

Tirunillai, S. and Tellis, G. J. (2012). Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Science*, 31(2):198–215.

Toubia, O. and Stephen, A. T. (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3):368–392.

Wei, Y., Yildirim, P., Van den Bulte, C., and Dellarocas, C. (2015). Credit scoring with social network data. *Marketing Science*, 35(2):234–258.

Yildirim, P., Gal-Or, E., and Geylani, T. (2013). User-generated content and bias in news media. *Management Science*, 59(12):2655–2666.

Zhang, K. and Sarvary, M. (2015). Differentiation with user-generated content. *Management Science*, 61(4):898–914.

Zuckerberg, M. (2020). Big tech needs more regulation. Facebook. `https://about.fb.com/news/2020/02/big-tech-needs-more-regulation/`. Accessed December 27, 2020.

ONLINE APPENDIX

A.1 Proof for the Optimality of Threshold Strategy

Proof for the optimality of threshold strategy under perfect technology

Lemma A.1. *When the content moderation technology is perfect, under both advertising and subscription revenues, any content moderation strategy which removes content with extremeness indices in $X \subset [0,1]$, is (weakly) dominated by a "threshold" strategy which removes content with extremeness indices greater than y.*

Proof of Lemma A.1. To prove Lemma A.1, we start with an arbitrary content moderation strategy, denoted as $\mathcal{C}_X$, which removes content with extremeness indices in $X \subset [0,1]$. We will show that there exists a threshold strategy that dominates $\mathcal{C}_X$. We first consider the case when the platform is earning revenue from advertising.

First, notice that for the individuals who participate in the platform, the utility from participation is increasing in their extremeness index, x. To see this, consider two users with extremeness indices x_1, x_2 whose content is not removed, where without loss of generality, $x_1 < x_2$. Let $\mathcal{X}$ be the users who participate in the platform. We can express the difference in the utility of these consumers as

$$\begin{aligned} U(x_1) - U(x_2) =& \alpha x_1 - \int_{\tilde{x}\in\mathcal{X}, \tilde{x}>x_1} \tilde{x} d\tilde{x} - (\alpha x_2 - \int_{\tilde{x}\in\mathcal{X}, \tilde{x}>x_2} \tilde{x} d\tilde{x}) \\ =& \alpha(x_1 - x_2) - \int_{\tilde{x}\in\mathcal{X}, x_1<\tilde{x}\leq x_2} \tilde{x} d\tilde{x} \\ <& 0, \end{aligned}$$

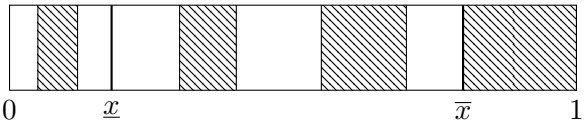
which implies that $U(x)$ is increasing in x. Then, if a user at x participates in the platform (has a non-negative utility from participating), then all other users with extremeness index greater than x should also participate, as long as the platform does not remove their content.

Next, take any content moderation strategy $\mathcal{C}_X$ which removes all the content with extremeness index $x \in X \subset [0,1]$. We will prove that $\mathcal{C}_X$ is dominated by a threshold strategy. As an illustration, the shaded areas in Figure A1 indicate the content that is removed (X) in this moderation strategy. We intentionally use an example where the set X contains four disjoint "blocks" to illustrate the steps in the proof. This approach can be generalized to rule out strategies with fewer or more disjoint blocks.

Let the user with the highest and the lowest extremeness indices among all users participating in the platform under $\mathcal{C}_X$ be denoted by $\overline{x}$ and $\underline{x}$, respectively, as illustrated in Figure A1. First, notice that by the monotonicity of $U(x)$, $(\overline{x}, 1] \in X$ must hold, since otherwise these users would have participated as well. Put differently, content in $(\overline{x}, 1]$ must be in the removed set. Second, again by

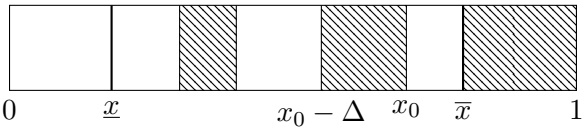
the monotonicity of $U(x)$, all users located between $\underline{x}$ and $\overline{x}$ must participate unless their content is removed.

Figure A1: Content Moderation Strategy $\mathcal{C}_X$



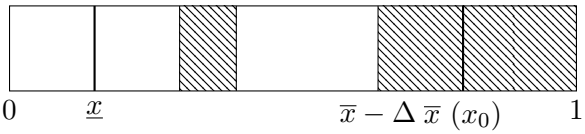
Third, $\mathcal{C}_X$ must be equivalent to (and is weakly dominated by) $\mathcal{C}_{X'}$ where $X' = X\backslash[0, \underline{x})$ and $\mathcal{C}_{X'}$ is illustrated in Figure A2. This is because, since by definition, users in $[0, \underline{x})$ do not participate, and a strategy removing their content cannot do better.

Figure A2: Content Moderation Strategy $\mathcal{C}_{X'}$



We have completed the proof if $\mathcal{C}_{X'}$ is a threshold strategy. If $\mathcal{C}_{X'}$ is not a threshold strategy, then we can still show that a threshold strategy dominates it. To see this, consider the "block" for removal that is to the left of and the closest to $\overline{x}$. Let the right border of this block be denoted with x_0 and its width be Δ, so that it covers the region $[x_0 - \Delta, x_0]$, as illustrated in Figure A2. If the platform moves this block to the right, its user base X^A will not decrease if we can show that $\frac{\partial X^A}{\partial x_0} \geq 0$. So the platform can move the block $[x_0 - \Delta, x_0]$ all the way to the right until $x_0 = \overline{x}$. That is, strategy $\mathcal{C}_{X'}$ is dominated by another strategy $\mathcal{C}_{X''}$ where $X'' = X'\backslash[x_0 - \Delta, x_0] \cup [\overline{x} - \Delta, \overline{x}]$, as illustrated in Figure A3.

Figure A3: Content Moderation Strategy $\mathcal{C}_{X''}$



In the following, we will show that $\underline{x}$ is non-increasing in x_0. For ease of expression, we denote the other blocks for removal in X' as $\hat{X}$, which does not change when the block $[x_0 - \Delta, x_0]$ is moved to the right. Mathematically, $\hat{X} = X'\backslash[x_0 - \Delta, x_0]$. When $\underline{x} > 0$ (strictly), by definition, the marginal user at $\underline{x}$ gets zero utility, i.e.,

$$0 = U(\underline{x}) = \alpha\underline{x} + v - \int_{\tilde{x} \in [\underline{x}, \overline{x}]\backslash X'} \tilde{x} d\tilde{x} \tag{A1}$$

$$= \alpha\underline{x} + v - \int_{\tilde{x} \in [\underline{x}, \overline{x}]\backslash (\hat{X} \cup [x_0 - \Delta, x_0])} \tilde{x} d\tilde{x} \tag{A2}$$

$$= \alpha \underline{x} + v - \int_{\tilde{x} \in [\underline{x}, \overline{x}]} \tilde{x} d\tilde{x} + \int_{\tilde{x} \in (\hat{X} \cup [x_0 - \Delta, x_0])} \tilde{x} d\tilde{x} \tag{A3}$$

$$= \alpha \underline{x} + v - \int_{\underline{x}}^{\overline{x}} \tilde{x} d\tilde{x} + \int_{\tilde{x} \in \hat{X}} \tilde{x} d\tilde{x} + \int_{x_0 - \Delta}^{x_0} \tilde{x} d\tilde{x}. \tag{A4}$$

Taking the derivative of $\underline{x}$ w.r.t. x_0 on both sides of Equation (A4) yields

$$0 = \alpha \frac{\partial \underline{x}}{\partial x_0} - (-\underline{x} \frac{\partial \underline{x}}{\partial x_0}) + x_0 - (x_0 - \Delta),$$

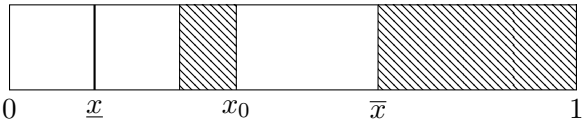
which implies

$$\frac{\partial \underline{x}}{\partial x_0} = -\frac{\Delta}{\alpha + \underline{x}} < 0$$

whenever $\underline{x} > 0$ since $\alpha \geq 0$ and $\Delta > 0$. That is, as long as $\underline{x}$ does not "bump" into zero, moving the block $[x_0 - \Delta, x_0]$ to the right will increase the user base X^A. If at some point, $\underline{x}$ bumps into zero and cannot further decrease, it does not hurt to keep moving it to the right, until $x_0 = \overline{x}$. Therefore, $\underline{x}$ is non-increasing in x_0 and thus $\mathcal{C}_{X'}$ is dominated by $\mathcal{C}_{X''}$. Note that going from strategy $\mathcal{C}_{X'}$ to strategy $\mathcal{C}_{X''}$ implies connecting the blocks $[x_0 - \Delta, x_0]$ and $[\overline{x}, 1]$ together. This can be seen more clearly by comparing Figures A2 and A3.

Let's redefine $\overline{x}$ as the highest extremeness index member among the participating users under strategy $\mathcal{C}_{X''}$, and redefine x_0 as the right border of the "block" for removal that is to the left of and closest to the "new" $\overline{x}$, as shown in Figure A4. Note that Figure A4 is identical to A3 except that we have redefined $\overline{x}$ and x_0. One can repeat the procedure of moving x_0 to $\overline{x}$ many times until it reaches a single contiguous threshold strategy. This threshold strategy dominates $\mathcal{C}_X$. In our example, we do this step once to get to the threshold strategy. Thus, any arbitrary content moderation strategy $\mathcal{C}_X$ is weakly dominated by a threshold strategy, and this concludes the proof for the advertising case.

Figure A4: Redefine $\overline{x}$ and x_0 for induction



When the platform earns revenue from subscription fees, all above derivations showing that the threshold strategy maximizes the user base still hold. We need to show that for any subscription fee p, an arbitrary content moderation strategy $\mathcal{C}_X$ is dominated by a threshold strategy.

Again, let the user with the highest extremeness index among all participating users be denoted by $\overline{x} \leq 1$ under content moderation strategy $\mathcal{C}_X$. When $p > \alpha \overline{x} + v$, which is the highest utility that a user can get, then $U(x) < 0$ for all users and no user will participate in the platform, so moderation strategy $\mathcal{C}_X$ is equivalent to the threshold strategy since there is no revenue anyway.

When $p \leq \alpha\overline{x}+v$, there exist some users whose utility from participating in the platform is positive. Note that for any given p, a larger user base means a larger revenue. We can simply replace "v" in the proof for the advertising case as "$v-p$" and everything still holds in that proof. Therefore, $\mathcal{C}_X$ can also be (weakly) dominated by a threshold strategy due to the same logic as in the advertising case.

□

Proof for the optimality of threshold strategy under imperfect technology

Recall that if the content moderation strategy is imperfect, then the platform can choose any $X \subset [0,1]$ as a "target zone" such that all content with extremeness index $x \in X$ is intended to be removed. Since the technology is imperfect, any content with $x \in X$ is removed with probability $\frac{1}{2}+k$ and any content with $x \notin X$ is removed with probability $\frac{1}{2}-k$, where $k \in [0, \frac{1}{2}]$. In this case, we can describe the optimal content moderation strategy in Lemma A.2.

Lemma A.2. *If the content moderation technology is imperfect, for both advertising and subscription revenues, any content moderation strategy targeting the content with extremeness indices in $X \subset [0,1]$ for removal is (weakly) dominated by a threshold strategy targeting content with extremeness indices greater than y for removal.*

Proof of Lemma A.2. To prove the lemma, we start with some arbitrary content moderation strategy $\mathcal{C}_X$ which targets content with extremeness indices in $X \subset [0,1]$ for removal. We will show that a threshold strategy dominates this strategy in five steps.

First, let's assume that the platform earns revenue from advertising. For any given content moderation strategy $\mathcal{C}_X$ and any user at x, let her utility from participating in the platform be $U^T(x)$ if she is in the "target zone" (i.e., $x \in X$) and $U^{NT}(x)$ if she is not (i.e., $x \notin X$). Then we have

$$U^T(x) = \alpha x(\frac{1}{2}-k) - c(\frac{1}{2}+k) + v - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\in X,\tilde{x}>x} \tilde{x}(\frac{1}{2}-k)d\tilde{x} - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\notin X,\tilde{x}>x} \tilde{x}(\frac{1}{2}+k)d\tilde{x}, \quad \text{(A5)}$$

$$U^{NT}(x) = \alpha x(\frac{1}{2}+k) - c(\frac{1}{2}-k) + v - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\in X,\tilde{x}>x} \tilde{x}(\frac{1}{2}-k)d\tilde{x} - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\notin X,\tilde{x}>x} \tilde{x}(\frac{1}{2}+k)d\tilde{x}. \quad \text{(A6)}$$

Since $U^T(x) - U^{NT}(x) = -2\alpha kx - 2ck \leq 0$,

$$U^T(x) \leq U^{NT}(x) \quad \text{(A7)}$$

holds for any x. Moreover, $U^T(x)$ and $U^{NT}(x)$ are both increasing in x since for any $x_1 < x_2$

$$\begin{aligned} U^T(x_1) - U^T(x_2) =& \alpha(x_1-x_2)(\frac{1}{2}-k) - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\in X,x_1\leq\tilde{x}\leq x_2} \tilde{x}(\frac{1}{2}-k)d\tilde{x} - \int_{\tilde{x}\in\mathcal{X},\tilde{x}\notin X,x_1\leq\tilde{x}\leq x_2} \tilde{x}(\frac{1}{2}+k)d\tilde{x} \\ <& 0, \end{aligned}$$

and using a similar derivation

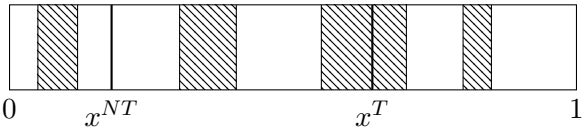
$$U^{NT}(x_1) - U^{NT}(x_2) < 0.$$

Thus, the utility from participating in the platform is increasing in x within and outside the target zone. By monotonicity, we claim that there exists a marginal user $x^T \in X$ such that any user at $x \in X$ will participate if $x \geq x^T$, and will not do so if $x < x^T$. Similarly, there exists $x^{NT} \notin X$ such that any user at $x \notin X$ will participate if $x \geq x^{NT}$ and will not do so if $x < x^{NT}$.

We first claim that $x^{NT} \leq x^T$. This is because otherwise there exists $x_0 \in (x^T, x^{NT})$ and thus a user at x_0 will participate if she is within the target zone but will not do so if she is outside the target zone. That is, $U^T(x_0) \geq 0$ and $U^{NT}(x_0) < 0$, and thus $U^T(x_0) > U^{NT}(x_0)$, which is a contradiction to the inequality given in (A7).

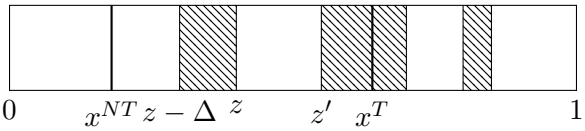
Next, we show that the content moderation strategy $\mathcal{C}_X$ is dominated by a threshold strategy. As an illustration, the shaded areas in Figure A5 indicate the target zone (X) of this moderation strategy. We intentionally use an example with four disjoint "blocks" as the target zone to illustrate the steps of finding the dominant threshold strategy. Note that this figure is an illustration to help readers understand the procedure for the proof described later, but the proof holds for any general X. We prove the statement by constructing a threshold strategy that induces a higher revenue for the platform in five steps.

Figure A5: Content moderation strategy $\mathcal{C}_X$



Step 1. For any content with $x < x^{NT}$, there is no need to include them in the target zone. This is because by definition, no users with $x < x^{NT}$ will participate under strategy $\mathcal{C}_X$ and these users will not affect the utility of the participating users since they are more moderate than them. Therefore, we claim that $\mathcal{C}_X$ is weakly dominated by $\mathcal{C}_{X_1}$ where $X_1 \equiv X \backslash [0, x^{NT})$. We illustrate the moderation strategy after Step 1, i.e., $\mathcal{C}_{X_1}$ in Figure A6.

Figure A6: Content moderation strategy $\mathcal{C}_{X_1}$



Step 2. After Step 1, there is no content pruned to the left of x^{NT}. Now consider the content in region $[x^{NT}, x^T]$. We claim that if there is any content intended to be pruned within region $[x^{NT}, x^T]$, it should be next to x^T. For example, in our illustration above in Figure A6, there is one block $[z-\Delta, z]$ within region $[x^{NT}, x^T]$ and $z < z'$, where z' is the left border of the block that contains x^T. We will show that it is better for the platform to move the block of target zone $[z-\Delta, z]$ to the right such that $z = z'$. Without loss of generality, we will only prove the claim for this "one block" example for ease of articulation. The same logic can be applied to the case of multiple blocks within the region $[x^{NT}, x^T]$.

In short, pushing the block of target zone $[z-\Delta, z]$ to the right next to z' will not affect anyone to the right of z', but will move x^{NT} to the left and thus increase the user base. To see this, recall that the marginal user outside the target zone x^{NT} is given by

$$\alpha x^{NT}(\frac{1}{2}+k) - c(\frac{1}{2}-k) + v - \int_{x^{NT}}^{z-\Delta} \tilde{x}(\frac{1}{2}+k)d\tilde{x} - \int_{z}^{z'} \tilde{x}(\frac{1}{2}+k)d\tilde{x} - A = 0, \tag{A8}$$

where A is all the negative utility that a user at x^{NT} suffers from reading content with extremeness index $> x^T$, and thus A is independent of z. Taking the first-order derivative w.r.t. z on both sides of Equation (A8) yields

$$\alpha(\frac{1}{2}+k)\frac{\partial x^{NT}}{\partial z} - \left((z-\Delta)(\frac{1}{2}+k) - x^{NT}(\frac{1}{2}+k)\frac{\partial x^{NT}}{\partial z}\right) - \left(-z(\frac{1}{2}+k)\right) = 0,$$

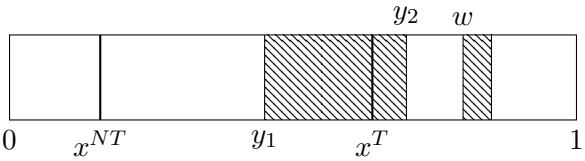
which implies

$$\frac{\partial x^{NT}}{\partial z} = -\frac{\Delta}{\alpha + x^{NT}} < 0,$$

so x^{NT} is decreasing as z, i.e., x^{NT} is indeed moved to the left as $[z-\Delta, z]$ is pushed to the right.

Note that it is possible that at some point x^{NT} can be exactly 0, then further moving the target region $[z-\Delta, z]$ to the right may not increase the user base, but it will not do any harm either, so we can continue to move $[z-\Delta, z]$ to the right anyway until $z = z'$ so that the "blocks" of target zones merge together and a new larger target zone "block" containing x^T is obtained. Denote now the left border of this new block as y_1, the right border of this block as y_2, and the left border of the target zone block next to this block as w. We also denote the improved content moderation strategy as $\mathcal{C}_{X_2}$, as illustrated in Figure A7.

Figure A7: Content moderation strategy $\mathcal{C}_{X_2}$



Step 3. So far we have shown that any content moderation strategy $\mathcal{C}_X$ is weakly dominated by a strategy $\mathcal{C}_{X_2}$ where X_2 has the following structure: $\exists y_1 \in [x^{NT}, x^T]$ such that $[y_1, x^T] \subset X_2$ and $[0, y_1) \not\subset X_2$. In fact, y_1 is a free parameter that the platform can choose for this content moderation strategy $\mathcal{C}_{X_2}$. By definition, users within region $[x^{NT}, y_1)$ incur a lower probability of being removed and choose to participate in the platform, while users within region $[y_1, x^T]$ incur a higher probability of being removed and choose not to participate. Thus, all else being equal, choosing a y_1 closer to x^T, although preserving more extreme users next to x^T, will also increase the negative utility that moderate users incur and thus push x^{NT} rightward. This is similar to the tradeoff of the platform that we have seen in Proposition 1. In this step, we shall show the following claim: all else being equal, the platform will optimally choose y_1 such that either one of the following conditions is satisfied: (a) $y_1 = x^T$ or (b) $y_1 = \underline{y_1}$ where $\underline{y_1}$ is chosen such that $x^{NT} = 0$.

First, it is easy to see that the platform has no incentive to choose any y_1 greater than x^T or less than $\underline{y_1}$. This is because choosing a y_1 greater than x^T does not preserve more extreme users compared to setting y_1 right at x^T, but only increases the negative utility that moderate users incur. Similarly, choosing a y_1 less than $\underline{y_1}$ will only make the user base smaller compared to setting y_1 at $\underline{y_1}$.

Consider any $y_1 \in [\underline{y_1}, x^T]$. By definition, we have

$$\alpha x^{NT}(\frac{1}{2} + k) - c(\frac{1}{2} - k) + v - \int_{x^{NT}}^{y_1} (\frac{1}{2} + k)\tilde{x} d\tilde{x} - A = 0, \tag{A9}$$

where A is all the negative utility that a user at x^{NT} suffers from reading content with extremeness index $> x^T$, and thus A is independent of y_1. Equation (A9) can be reduced to

$$\frac{(x^{NT})^2}{2} + \alpha x^{NT} - \frac{1}{2} y_1^2 + B = 0, \tag{A10}$$

where $B = \frac{-c(\frac{1}{2}-k)+v-A}{\frac{1}{2}+k}$, independent of y_1. Solving for x^{NT}, we get

$$x^{NT} = -\alpha + \sqrt{\alpha^2 + y_1^2 - 2B}. \tag{A11}$$

Note that $y_1 \in [\underline{y_1}, x^T]$ ensures that users at x^{NT} always get a non-positive utility. Thus, the left hand side of Equation (A10) is non-positive when $x^{NT} = 0$, which implies that $B \leq \frac{1}{2} y_1^2$. Thus, the term within the square root is always non-negative.

The user base, which is the objective function of the platform, is $X^A = y_1 - x^{NT} + 1 - x^T$, so

$$\begin{aligned} \frac{\partial X^A}{\partial y_1} &= 1 - \frac{\partial x^{NT}}{\partial y_1} \\ &= 1 - \frac{\partial(-\alpha + \sqrt{\alpha^2 + y_1^2 - 2B})}{\partial y_1} \end{aligned}$$

$$= \frac{\sqrt{\alpha^2 + y_1^2 - 2B} - y_1}{\sqrt{\alpha^2 + y_1^2 - 2B}}.$$

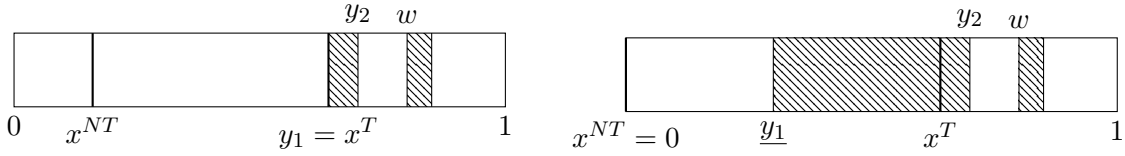
Thus,

$$\begin{aligned}
\text{sign}\left(\frac{\partial X^A}{\partial y_1}\right) &= \text{sign}\left(\sqrt{\alpha^2 + y_1^2 - 2B} - y_1\right) \\
&= \text{sign}\left((\sqrt{\alpha^2 + y_1^2 - 2B})^2 - y_1^2\right) \\
&= \text{sign}(\alpha^2 - 2B),
\end{aligned}$$

which is independent of y_1. Therefore, the sign of $\frac{\partial X^A}{\partial y_1}$ does not depend on y_1, i.e., X^A is monotonic in y_1 when $y_1 \in [\underline{y_1}, x^T]$, and thus the optimal y_1 must be at the corner – either x^T or $\underline{y_1}$.

Step 3 actually proves that $\mathcal{C}_X$ is dominated by either one of the following strategies illustrated in Figure A8, denoted as $\mathcal{C}_{X_{3a}}$ ($y_1 = x^T$) and $\mathcal{C}_{X_{3b}}$ ($y_1 = \underline{y_1}$), respectively.

Figure A8: Content moderation strategies $\mathcal{C}_{X_{3a}}$ and $\mathcal{C}_{X_{3b}}$



(a) $\mathcal{C}_{X_{3a}}$ ($y_1 = x^T$)

(b) $\mathcal{C}_{X_{3a}}$ ($y_1 = \underline{y_1}$ s.t. $x^{NT} = 0$)

Step 4. In this step, we will prove that for both $\mathcal{C}_{X_{3a}}$ and $\mathcal{C}_{X_{3b}}$, ceteris paribus, it is better for the platform to move y_2 to the right such that $y_2 = w$ and thus the two target zone blocks are connected. That is, we want to show $\frac{\partial X^A}{\partial y_2}$ is non-negative when $y_2 \leq w$. In the following, we prove this separately for $\mathcal{C}_{X_{3a}}$ and $\mathcal{C}_{X_{3b}}$.

(a) For $\mathcal{C}_{X_{3a}}$:

By definition, the utility of x^T and x^{NT} from participation is zero:

$$\alpha x^T(\frac{1}{2} - k) - c(\frac{1}{2} + k) + v - \int_{x^T}^{y_2}(\frac{1}{2} - k)\tilde{x}d\tilde{x} - \int_{y_2}^{w}(\frac{1}{2} + k)\tilde{x}d\tilde{x} - D = 0, \tag{A12}$$

and

$$\alpha x^{NT}(\frac{1}{2} + k) - c(\frac{1}{2} - k) + v - \int_{x^{NT}}^{x^T}(\frac{1}{2} + k)\tilde{x}d\tilde{x} - \int_{x^T}^{y_2}(\frac{1}{2} - k)\tilde{x}d\tilde{x} - \int_{y_2}^{w}(\frac{1}{2} + k)\tilde{x}d\tilde{x} - D = 0, \tag{A13}$$

where D is the negative utility that a user at x^T or x^{NT} suffers from reading content with extremeness index $> w$, and thus D is independent of y_2.

Taking the first-order derivative w.r.t. y_2 on both sides of Equations (A12) and (A13) yields

$$\alpha(\frac{1}{2}-k)\frac{\partial x^T}{\partial y_2}-(\frac{1}{2}-k)(y_2-x^T\frac{\partial x^T}{\partial y_2})+(\frac{1}{2}+k)y_2=0$$

and

$$\alpha(\frac{1}{2}+k)\frac{\partial x^{NT}}{\partial y_2}-(\frac{1}{2}+k)(x^T\frac{\partial x^T}{\partial y_2}-x^{NT}\frac{\partial x^{NT}}{\partial y_2})-(\frac{1}{2}-k)(y_2-x^T\frac{\partial x^T}{\partial y_2})+(\frac{1}{2}+k)y_2=0,$$

which, since $k<1/2$ imply

$$\frac{\partial x^T}{\partial y_2}=-\frac{4ky_2}{(1-2k)(\alpha+x^T)}<0,$$

and

$$\frac{\partial x^{NT}}{\partial y_2}=-\frac{4k(y_2-x^T\frac{\partial x^T}{\partial y_2})}{(1+2k)(\alpha+x^{NT})}=-\frac{4k(y_2+x^T\frac{4ky_2}{(1-2k)(\alpha+x^T)})}{(1+2k)(\alpha+x^{NT})}<0.$$

The user base, which is the objective function of the platform, is $X^A=1-x^T+x^T-x^{NT}=1-x^{NT}$, so

$$\frac{\partial X^A}{\partial y_2}=-\frac{\partial x^{NT}}{\partial y_2}>0.$$

Therefore, moving y_2 to the right can increase the user base if x^{NT} is interior ($x^{NT}>0$). It is possible that at some point x^{NT} can be exactly 0, then further moving y_2 to the right may not increase the user base, but it will not do any harm either ($\frac{\partial X^{NT}}{\partial y_2}=0$), so we can move y_2 to the right until $y_2=w$ and get a strategy dominating $\mathcal{C}_{X_{3a}}$.

(b) For $\mathcal{C}_{X_{3b}}$:

By definition of x^T and $\underline{y_1}$,

$$\alpha x^T(\frac{1}{2}-k)-c(\frac{1}{2}+k)+v-\int_{x^T}^{y_2}(\frac{1}{2}-k)\tilde{x}d\tilde{x}-\int_{y_2}^{w}(\frac{1}{2}+k)\tilde{x}d\tilde{x}-D=0, \tag{A14}$$

and

$$-c(\frac{1}{2}-k)+v-\int_0^{\underline{y_1}}(\frac{1}{2}+k)\tilde{x}d\tilde{x}-\int_{x^T}^{y_2}(\frac{1}{2}-k)\tilde{x}d\tilde{x}-\int_{y_2}^{w}(\frac{1}{2}+k)\tilde{x}d\tilde{x}-D=0, \tag{A15}$$

where D is the negative utility that a user at x^T or x^{NT} suffers from reading content with extremeness index $>w$, and thus D is independent of y_2.

Taking the first-order derivative w.r.t. y_2 on both sides of Equations (A14) and (A15) yields

$$\alpha(\frac{1}{2}-k)\frac{\partial x^T}{\partial y_2}-(\frac{1}{2}-k)(y_2-x^T\frac{\partial x^T}{\partial y_2})+(\frac{1}{2}+k)y_2=0$$

and

$$-(\frac{1}{2}+k)\underline{y_1}\frac{\partial \underline{y_1}}{\partial y_2}-(\frac{1}{2}-k)(y_2-x^T\frac{\partial x^T}{\partial y_2})+(\frac{1}{2}+k)y_2=0,$$

which imply

$$\frac{\partial x^T}{\partial y_2}=-\frac{4ky_2}{(1-2k)(\alpha+x^T)}<0,$$

and

$$\frac{\partial \underline{y_1}}{\partial y_2}=\frac{4ky_2+(1-2k)x^T\frac{\partial x^T}{\partial y_2}}{(1+2k)y_1}=\frac{4ky_2\alpha}{(1+2k)y_1(\alpha+x^T)}>0.$$

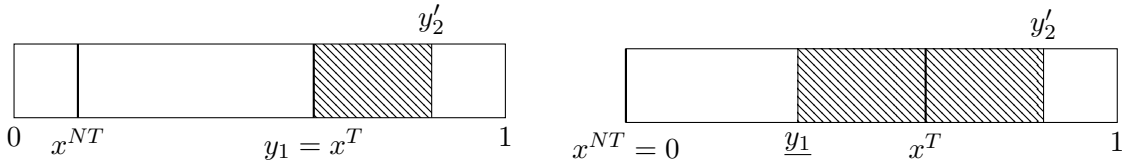
The user base, which is the objective function of the platform, is $X^A=1-x^T+\underline{y_1}$, so

$$\frac{\partial X^A}{\partial y_2}=-\frac{\partial x^{NT}}{\partial y_2}+\frac{\partial \underline{y_1}}{\partial y_2}>0.$$

Therefore, moving y_2 to the right can increase $\underline{y_1}$ and decrease x^T, and thus increase the user base if $\underline{y_1}<x^T$. It is possible that at some point $\underline{y_1}$ and x^T "bump" into each other, then all users in $[0,1]$ participate in the platform and thus further moving y_2 to the right may not increase the user base, but it will not do any harm either. Therefore, we can move y_2 to the right until $y_2=w$ and get a strategy dominating $\mathcal{C}_{X_{3b}}$.

In summary, for either $\mathcal{C}_{X_{3a}}$ or $\mathcal{C}_{X_{3b}}$, we can move y_2 to the right until $y_2=w$ so that the "blocks" of target zones merge together and a new larger target zone "block" containing x^T is obtained. Denote now the right border of this new block as a "new" y_2 (shown as y_2' in the figures below). The improved moderation strategies after Step 4 are illustrated in Figure A9.

Figure A9: Content moderation strategies after Step 4

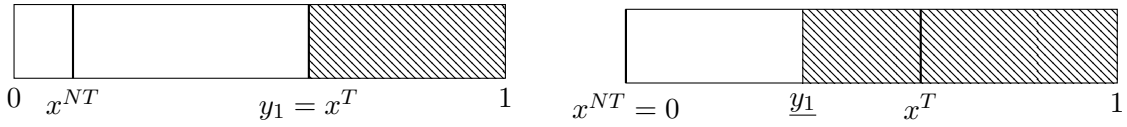


(a) $y_1=x^T$ (b) $y_1=\underline{y_1}$ s.t. $x^{NT}=0$

Step 5. Repeat Step 4 until we "merge" all blocks together on the right and get a threshold strategy (i.e., $y_2=1$). This threshold strategy dominates $\mathcal{C}_X$ in terms of the platform's revenue. In our illustrating example, we only need to repeat this once and get the threshold strategy as shown in either Figure A10a or A10b.

Thus far, we have found the threshold strategy that dominates $\mathcal{C}_X$ under advertising. The whole argument also works for a subscription-based platform, based on the same logic as we show in the

Figure A10: Content moderation strategies after Step 5 (becoming threshold strategies)



(a) $y_1 = x^T$ (b) $y_1 = \underline{y_1}$ s.t. $x^{NT} = 0$

perfect technology case: When the subscription fee p is given, the only thing that a platform cares about is the user base, so the procedure above for the advertising revenue case still applies. Therefore, for any given subscription fee p, any arbitrary content moderation strategy is dominated by a threshold strategy. In particular, when the optimal subscription fee is also chosen optimally, the optimal content moderation strategy should still be a threshold one.

□

A.2 Proofs of Lemmas and Propositions in Main Text

Proof of Lemma 1. When $x > y$, since $c > v$ by assumption, $U(x) = -c + v < 0$ holds and users with $x > y$ do not participate in the platform.

For users $x \leq y$, consider two users x_1, x_2 such that $x_1 < x_2 \leq y$. Then

$$\begin{aligned} U(x_1) - U(x_2) =& \alpha x_1 - \int_{\tilde{x} \in \hat{\mathcal{X}}, x_1 < \tilde{x} \leq y} \tilde{x} d\tilde{x} - (\alpha x_2 - \int_{\tilde{x} \in \hat{\mathcal{X}}, x_2 < \tilde{x} \leq y} \tilde{x} d\tilde{x}) \\ =& \alpha(x_1 - x_2) - \int_{\tilde{x} \in \hat{\mathcal{X}}, x_1 < \tilde{x} \leq x_2} \tilde{x} d\tilde{x} \\ <& 0, \end{aligned}$$

implying first that, if x_1 participates then all users in the range $[x_1, y]$ participate, and second that the utility of participating users is increasing in x. The former also implies that there exists $x^A \geq 0$ such that the user base is $[x^A, y]$. □

Proof of Proposition 1. Recall that the platform's profit maximization problem is $\max_y \pi^A = \zeta(y - x^A(y))$, where

$$x^A(y) = \begin{cases} -\alpha + \sqrt{\alpha^2 + y^2 - 2v} & \text{if } y \geq \sqrt{2v}, \\ 0 & \text{if } y < \sqrt{2v}, \end{cases}$$

from Equation (3).

When $y < \sqrt{2v}$, $x^A(y) = 0$ and the profit of the platform is ζy, which is maximized when $y^* = \sqrt{2v}$.

When $y \geq \sqrt{2v}$, the profit becomes $\zeta(y+\alpha-\sqrt{\alpha^2+y^2-2v})$ and $\frac{d\pi^A}{dy} = \zeta(1-\frac{y}{\sqrt{y^2+\alpha^2-2v}})$. Notice that the profit is increasing in y when $\frac{d\pi^A}{dy} > 0$, which holds iff $\alpha > \sqrt{2v}$. Therefore, when $\alpha \geq \sqrt{2v}$, profit maximizing moderation strategy is $y^{A*} = 1$. On the other hand, when $\alpha < \sqrt{2v}$, $\frac{d\pi^A}{dy} \leq 0$ and the profit maximizing content moderation strategy is $y^{A*} = \sqrt{2v}$. In other words, we find $\alpha^A \equiv \sqrt{2v}$ such that $y^{A*} = \sqrt{2v}$ when $\alpha < \alpha^A$ while $y^{A*} = 1$ when $\alpha \geq \alpha^A$.

□

Proof of Proposition 2. Under subscription, the profit maximization problem of the platform is $\max_{y,p} \pi^S = p(y - x^S(y,p))$, where

$$x^S(y,p) = \begin{cases} -\alpha + \sqrt{\alpha^2+y^2-2(v-p)} & \text{if } y \geq \sqrt{2(v-p)}, \\ 0 & \text{if } y < \sqrt{2(v-p)}. \end{cases}$$

from Equation (5).

- If $y \leq \sqrt{2(v-p)}$, then $p \leq v - y^2/2$ holds. Thus

$$\pi^S = py \leq (v - \frac{y^2}{2})y \leq \frac{2v}{3}\sqrt{\frac{2v}{3}}. \tag{A16}$$

  Therefore, π^S takes the maximum value of $\frac{2v}{3}\sqrt{\frac{2v}{3}}$, when both inequalities in Equation (A16) are equality, i.e., $p = v - \frac{y^2}{2}$ and $(v - \frac{y^2}{2})y = \frac{2v}{3}\sqrt{\frac{2v}{3}}$. These two conditions give the optimal price $p = \frac{2v}{3}$ and content moderation policy $y = \sqrt{\frac{2v}{3}}$ when $y \leq \sqrt{2(v-p)}$.

- If $y \geq \sqrt{2(v-p)}$, then $\pi^S = p(y+\alpha-\sqrt{\alpha^2+y^2-2(v-p)})$. First fix p as given. Taking the first order partial derivative w.r.t. y, we have $\frac{\partial \pi^S}{\partial y} = p\Big(1 - \frac{y}{\sqrt{y^2+\alpha^2-2(v-p)}}\Big)$. If $\alpha^2 - 2(v-p) > 0$, i.e., $p > v - \alpha^2/2$, $\frac{\partial \pi^S}{\partial y} = p\Big(1 - \frac{y}{\sqrt{y^2+\alpha^2-2(v-p)}}\Big) > p\Big(1 - \frac{y}{\sqrt{y^2+0}}\Big) = 0$. Therefore, π^S is increasing in y for any given $p > v - \alpha^2/2$, or the optimal level of y is 1. If $\alpha^2 - 2(v-p) \leq 0$, i.e., $p \leq v - \alpha^2/2$, $\frac{\partial \pi^S}{\partial y} \leq 0$, which means that π^S is decreasing in y, or the optimal y for any given $p \leq v - \alpha^2/2$ is $\sqrt{2(v-p)}$.

So the optimal level of moderation is either $y = \sqrt{2(v-p)}$ or $y = 1$. When $y = \sqrt{2(v-p)}$, we have seen that the optimal level of p and y should be $p = \frac{2v}{3}$ and $y = \sqrt{\frac{2v}{3}}$, which induces $\pi^S = \frac{2v}{3}\sqrt{\frac{2v}{3}}$. When $y = 1$, the optimal subscription fee $\hat{p}^{S*}$ should maximize

$$\hat{\pi^S}(p) = p(1 - x^S(1,p)) = p(1 + \alpha - \sqrt{\alpha^2+1-2(v-p)}).$$

Solving the first order condition (FOC) w.r.t. p gives

$$\hat{p}^{S*} = \frac{1}{9}\Big[(1+\alpha)\sqrt{2(2-3v+2\alpha^2+\alpha)} - 2(1-3v+\alpha^2-\alpha)\Big].$$

The second order condition (SOC) is clearly satisfied since

$$\frac{\partial^2 \hat{\pi^S}(p)}{\partial p^2} = \frac{-2(\alpha^2+1)-3p+4v}{(\alpha^2+2p-2v+1)^{3/2}} < \frac{-2+4v}{(\alpha^2+2p-2v+1)^{3/2}} < 0.$$

Therefore, $\pi^{S*} = \max\{\frac{2v}{3}\sqrt{\frac{2v}{3}}, \hat{\pi^S}(\hat{p}^{S*})\}$. By the envelope theorem, we know that

$$\frac{\partial \hat{\pi^S}(\hat{p}^{S*})}{\partial \alpha} = \frac{\partial \hat{\pi^S}(p)}{\partial \alpha}|_{p=\hat{p}^{S*}} = p(1 - \frac{\alpha}{\sqrt{\alpha^2+1-2(v-p)}})|_{p=\hat{p}^{S*}} = \hat{p}^{S*}(1 - \frac{\alpha}{\sqrt{\alpha^2+1-2(v-\hat{p}^{S*})}}).$$

Since by construction $1 \geq x^S(1,p) = -\alpha + \sqrt{\alpha^2+1-2(v-p)}$, we know that $1 - \frac{\alpha}{\sqrt{\alpha^2+1-2(v-p)}} \geq 0$ for any p, and specifically for $p = \hat{p}^{S*}$. Thus, $\frac{\partial \hat{\pi^S}(\hat{p}^{S*})}{\partial \alpha} \geq 0$, i.e., $\hat{\pi^S}(\hat{p}^{S*})$ is increasing in α. $\frac{2v}{3}\sqrt{\frac{2v}{3}}$ is independent of α. Therefore, proving that there exists $\alpha^S \in (0, \alpha^A)$ such that $\pi^{S*} = \frac{2v}{3}\sqrt{\frac{2v}{3}}$ (with the optimal content moderation strategy $y^{S*} = \sqrt{\frac{2v}{3}}$) when $\alpha < \alpha^S$ and $\pi^{S*} = \hat{\pi^S}(\hat{p}^{S*})$ (with the optimal content moderation strategy $y^{S*} = 1$) when $\alpha > \alpha^S$ requires $\hat{\pi^S}(\hat{p}^{S*}) < \frac{2v}{3}\sqrt{\frac{2v}{3}}$ when $\alpha = 0$ and $\hat{\pi^S}(\hat{p}^{S*}) > \frac{2v}{3}\sqrt{\frac{2v}{3}}$ when $\alpha = \sqrt{2v} \equiv \alpha^A$ for any $v \in (0, \frac{1}{2})$. To check this, we denote

$$H(v) = (\hat{\pi^S}(\hat{p}^{S*}) - \frac{2v}{3}\sqrt{\frac{2v}{3}})|_{\alpha=0}$$

and

$$J(v) = (\hat{\pi^S}(\hat{p}^{S*}) - \frac{2v}{3}\sqrt{\frac{2v}{3}})|_{\alpha=\sqrt{2v}}$$

and we want to show $H(v) < 0$ and $J(v) > 0$ for any $v \in (0, \frac{1}{2})$.

Since $H(v) = (\hat{\pi^S}(\hat{p}^{S*}) - \frac{2v}{3}\sqrt{\frac{2v}{3}})|_{\alpha=0}$, plugging in the expression of $\hat{\pi^S}(\hat{p}^{S*})$ and $\alpha = 0$ obtains

$$H(v) = \frac{1}{27}(2-\sqrt{4-6v})(6v+\sqrt{4-6v}-2) - \frac{2v}{3}\sqrt{\frac{2v}{3}},$$

$$H'(v) = \frac{1}{3}(2-\sqrt{6v}-\sqrt{4-6v})$$

and

$$H''(v) = \frac{1}{\sqrt{4-6v}} - \frac{1}{\sqrt{6v}}.$$

Note that $H''(v) \lesseqgtr 0$ if $v \lesseqgtr \frac{1}{3}$, so $H'(v)$ is first decreasing and then increasing on $(0, \frac{1}{2})$. Therefore,

$H'(v) < \max\{H'(0), H'(\frac{1}{2})\} = \max\{0, \frac{1}{3}(1-\sqrt{3})\} = 0$, so $H(v)$ is decreasing in v. Thus,

$$H(v) < H(0) = 0.$$

Since $J(v) = (\hat{\pi^S}(\hat{p}^{S*}) - \frac{2v}{3}\sqrt{\frac{2v}{3}})|_{\alpha=\sqrt{2v}}$, plugging in the expression of $\hat{\pi^S}(\hat{p}^{S*})$ and $\alpha = \sqrt{2v}$ obtains

$$J(v) = \frac{1}{27}(2\sqrt{2v} - \sqrt{2(v+\sqrt{2v}+2)}+2)(2v+2\sqrt{2v}+(2\sqrt{v}+\sqrt{2})\sqrt{v+\sqrt{2v}+2}-2),$$

$$J'(v) = \frac{(\sqrt{2}-3\sqrt{6})v + \sqrt{v+\sqrt{2v}+2} + \sqrt{v}(\sqrt{2(v+\sqrt{2v}+2)}+2) - \sqrt{2}}{9\sqrt{v}}.$$

Denote the numerator of $J'(v)$ as $J_1(v)$, then

$$J_1'(v) = \frac{4\sqrt{2}v + 8\sqrt{v} + 4\sqrt{v+\sqrt{2v}+2} + 5\sqrt{2}}{4\sqrt{v}\sqrt{v+\sqrt{2v}+2}} + \sqrt{2} - 3\sqrt{6}.$$

$J_1'(v) = 0$ has a unique solution $v = v_0$ where $v_0 \approx 0.2187 < \frac{1}{2}$. Furthermore, $J_1'(v) \gtrless 0$ when $v \lessgtr v_0$. Therefore, $J_1(v) > \min\{J_1(0), J_1(\frac{1}{2})\} = \min\{0, \frac{\sqrt{2}-3\sqrt{6}+2\sqrt{14}}{2}\} = 0$. So, $J'(v) > 0$ and then

$$J(v) > J(0) = 0.$$

Thus, by the fact that $\hat{\pi^S}(\hat{p}^{S*})$ is increasing in α while $\frac{2v}{3}\sqrt{\frac{2v}{3}}$ is independent of α, we claim that there exists $\alpha^S \in (0, \alpha^A)$ such that $\pi^{S*} = \frac{2v}{3}\sqrt{\frac{2v}{3}}$ ($y^{S*} = \sqrt{\frac{2v}{3}}$) when $\alpha < \alpha^S$, while $\pi^{S*} = \hat{\pi^S}(\hat{p}^{S*})$ ($y^{S*} = 1$) when $\alpha > \alpha^S$.

□

Proof of Proposition 3. We first consider the case where content moderation is allowed. When $\alpha < \alpha^S$, $\pi^{A*} = \zeta\sqrt{2v}$ and $\pi^{S*} = \frac{2v}{3}\sqrt{\frac{2v}{3}}$. Thus, $\overline{\zeta} = \frac{\frac{2v}{3}\sqrt{\frac{2v}{3}}}{\sqrt{2v}} = \frac{2v}{3\sqrt{3}}$ which is independent of α. When $\alpha^S \leq \alpha < \alpha^A$, $\pi^{A*} = \zeta\sqrt{2v}$ and $\pi^{S*} = \hat{\pi^S}(\hat{p}^{S*})$. Thus, $\overline{\zeta} = \frac{\hat{\pi^S}(\hat{p}^{S*})}{\sqrt{2v}}$. We have proved that $\hat{\pi^S}(\hat{p}^{S*})$ is increasing in α in the proof of Proposition 2, so $\overline{\zeta}$ is also increasing in α.

If content moderation is not allowed, the expressions for the platform's profits (denoted as π_0^A and π_0^S) are the same as those when no moderation is conducted, i.e., $\pi_0^A = \zeta(1+\alpha-\sqrt{\alpha^2+1-2v})$ and $\pi_0^S = \hat{\pi^S}(\hat{p}^{S*})$. Thus, $\hat{\zeta} = \frac{\hat{\pi^S}(\hat{p}^{S*})}{1+\alpha-\sqrt{\alpha^2+1-2v}}$. Clearly, when $\alpha^S \leq \alpha < \alpha^A$, $\overline{\zeta} < \hat{\zeta}$ since $\pi_0^S = \pi^{S*}$ but $\pi_0^A < \pi^{A*}$ (this is when the optimal strategy for an advertising-based platform is to conduct moderation but that of a subscription-based one is not to do so).

We calculate

$$\begin{aligned}\frac{\partial\hat{\zeta}}{\partial\alpha}=&\frac{\frac{\partial\hat{\pi^S}(\hat{p}^{S*})}{\partial\alpha}(1+\alpha-\sqrt{\alpha^2+1-2v})-\hat{\pi^S}(\hat{p}^{S*})(1-\frac{\alpha}{\sqrt{\alpha^2+1-2v}})}{(1+\alpha-\sqrt{\alpha^2+1-2v})^2}\\=&\frac{\hat{p}^{S*}}{(1+\alpha-\sqrt{\alpha^2+1-2v})^2}\Big((1-\frac{\alpha}{\sqrt{\alpha^2+1-2(v-\hat{p}^{S*})}})(1+\alpha-\sqrt{\alpha^2+1-2v})\\&-(1+\alpha-\sqrt{\alpha^2+1-2(v-\hat{p}^{S*})})(1-\frac{\alpha}{\sqrt{\alpha^2+1-2v}})\Big).\end{aligned}$$

Denote $A=\sqrt{\alpha^2+1-2(v-\hat{p}^{S*})}$ and $B=\sqrt{\alpha^2+1-2v}$, then $A\geq B>\alpha$.

$$\begin{aligned}\text{sign}(\frac{\partial\hat{\zeta}}{\partial\alpha})=&\text{sign}\Big((1-\frac{\alpha}{A})(1+\alpha-B)-(1+\alpha-A)(1-\frac{\alpha}{B})\Big)\\=&\text{sign}\Big(\frac{(A-B)(AB+(1-A-B)\alpha+\alpha^2)}{AB}\Big)\\=&\text{sign}(AB+(1-A-B)\alpha+\alpha^2).\end{aligned}$$

Note that

$$\begin{aligned}AB+(1-A-B)\alpha+\alpha^2)=&A(B-\alpha)-B\alpha+\alpha+\alpha^2\\\geq&B(B-\alpha)-B\alpha+\alpha+\alpha^2\\=&(B-\alpha)^2+\alpha\\>&0.\end{aligned}$$

Therefore, $\frac{\partial\hat{\zeta}}{\partial\alpha}>0$, i.e., $\hat{\zeta}$ is increasing in α. When $\alpha=\alpha^S$, it has been shown at the end of last paragraph that $\hat{\zeta}>\overline{\zeta}$. When $\alpha=0$,

$$\hat{\zeta}-\overline{\zeta}=\frac{1}{27}\Big(\frac{(3-(\sqrt{4-6v}+1))(6v+\sqrt{4-6v}-2)}{1-\sqrt{1-2v}}-6\sqrt{3}v\Big).$$

Proving $\hat{\zeta}<\overline{\zeta}$ requires that

$$G(v)=(3-(\sqrt{4-6v}+1))(6v+\sqrt{4-6v}-2)-6\sqrt{3}v(1-\sqrt{1-2v})<0$$

for any $v\in(0,\frac{1}{2})$. Note that

$$G'(v)=-\frac{3\sqrt{3}}{\sqrt{1-2v}}+9(\sqrt{3-6v}-\sqrt{4-6v})-6\sqrt{3}+18,$$

while both $-\frac{3\sqrt{3}}{\sqrt{1-2v}}$ and $\sqrt{3-6v}-\sqrt{4-6v}$ are decreasing in v, so $G'(v)$ is decreasing in v. Then

$G'(v) < G'(0) = 0$ so $G(v)$ is decreasing in v. Thus, $G(v) < G(0) = 0$. Therefore, $\hat{\zeta} < \overline{\zeta}$ when $\alpha = 0$. Thus, by the fact that $\hat{\zeta}$ is increasing in α, we can claim that there is $\alpha_1 \in (0, \alpha^S)$ such that $\hat{\zeta} \lesseqgtr \overline{\zeta}$ when $\alpha \lesseqgtr \alpha_1$ and finish the proof.

□

Proof of Proposition 4. Part (i): We first prove that a platform carries out content moderation only if the technology is sufficiently accurate, under both advertising and subscription revenues. To this end, we show that there exists $\epsilon > 0$ such that when $k < \epsilon$, the profit induced by optimal moderation strategy is less than that of no content moderation. Notice that when $k = \frac{1}{2}$, both under advertising and subscription revenue, a platform chooses to moderate content, under the assumption $\alpha < \alpha^S$ stated on page 18. Since the platform's profit

$$\pi_k^A = \zeta(1 - x_{2,k}^A + y - x_{1,k}^A(y))$$

or

$$\pi_k^S = p(1 - x_{2,k}^S(p) + y - x_{1,k}^S(y,p))$$

is obviously continuous in k, it suffices to show that the platform's profit if moderating content is lower than that if no moderation is conducted, when the technology accuracy is $k = 0$.

We start with the analysis of a platform with advertising revenues. Note that when $k = 0$, all content has the probability $\frac{1}{2}$ of being pruned, regardless of their extremeness index x and the platform's choice of y. A user at $x = 1$ receives utility $U(1) \equiv \frac{1}{2}\alpha - \frac{1}{2}c + v \geq 0$ by the assumption $c \leq \alpha + 2v$. Therefore, the user base for the platform will be $[\underline{x}, 1]$ where the marginal user $\underline{x}$ is the solution to $U(\underline{x}) \equiv \frac{1}{2}\alpha\underline{x} - \frac{1}{2}c + v - \frac{1}{2}\frac{1}{2}(1 - \underline{x}^2) = 0$. With some algebra, we know that the size of the platform's user base is

$$1 - \underline{x} = 1 + \alpha - \sqrt{\alpha^2 + 1 - 2(2v - c)} < 1 + \alpha - \sqrt{\alpha^2 + 1 - 2v},$$

which is the user base size if no moderation is conducted. The inequality comes from the fact that $v < c$ and thus $2v - c < v$. Therefore, the profit with content moderation is also less than that without moderation.

The proof when the platform earns revenues from subscription is similar to that under advertising revenues. With subscription revenues and lowest accuracy ($k = 0$), one can show that for any given subscription price p, the user base is smaller when the platform moderates content than when it does not: If p induces $U(1) \leq 0$, there is no user on the platform so the user base (zero) is trivially smaller when the platform moderates content than when it does not. If p induces $U(1) > 0$, the procedure to prove that the user base is smaller when the platform moderates content than when it does not is

exactly the same as for advertising case, except that we replace the terms v with $v - p$.

Therefore, we have proved that a platform will conduct content moderation only if technology is sufficiently accurate, for both advertising and subscription.

Part (ii): Next, we prove that if the platform is moderating content, it may prune more of the moderate content than it does of the extreme content, and moreover, the average extremeness index of the content on the platform may be lower than when it prunes more of the extreme content and when it does not moderate content. To prove the existence of an equilibrium content moderation strategy where these statements hold, it suffices to give an example.

First, consider a platform under advertising revenue. Based on Figure 5, when the content moderation policy is y, the amount of the extreme content (i.e., $x > y$) that is pruned in equilibrium, denoted as $M_{1,k}(y)$, is

$$M_{1,k}(y) = (\frac{1}{2} + k)(1 - \max\{y, x_{2,k}\}),$$

and the amount of the moderate content (i.e., $x < y$) that is pruned in equilibrium, denoted as $M_{2,k}(y)$, is

$$M_{2,k}(y) = (\frac{1}{2} - k)(y - x_{1,k}(y)).$$

In equilibrium, the platform prunes $M_{1,k}^{A*} \equiv M_{1,k}(y^{A*})$ unit of extreme content as well as $M_{2,k}^{A*} \equiv M_{2,k}(y^{A*})$ unit of moderate content. Based on the expression of the average extremeness index ($\hat{x}$) on page 20, in equilibrium, the average extremeness index ($\hat{x}_k^{A*}$) is

$$\begin{aligned}\hat{x}_k^{A*} =& \frac{\int_{\mathcal{X}} x(1-q(x))dx}{\int_{\mathcal{X}}(1-q(x))dx} \\ =& \begin{cases} \frac{\int_{x_{1,k}(y^{A*})}^{y_k^{A*}} x(\frac{1}{2}+k)dx + \int_{x_{2,k}}^{1} x(\frac{1}{2}-k)dx}{(\frac{1}{2}+k)(y_k^{A*}-x_{1,k}(y^{A*}))+(\frac{1}{2}-k)(1-x_{2,k})} & \text{if content moderation is conducted in equilibrium,} \\ \frac{-\alpha+\sqrt{\alpha^2+1-2v}+1}{2} & \text{if content moderation is not conducted in equilibrium.} \end{cases}\end{aligned}$$

Consider $\alpha = 0.05, v = 0.2, c = 0.25$. Plug in the numbers into the expressions for $M_{1,k}^{A*}$, $M_{2,k}^{A*}$, and $\hat{x}_k^{A*}$, we can plot out a figure with k as x-axis while $M_{1,k}^{A*}$, $M_{2,k}^{A*}$, and $\hat{x}_k^{A*}$ as y-axis, to find out whether there can be cases such that the following two claims hold:

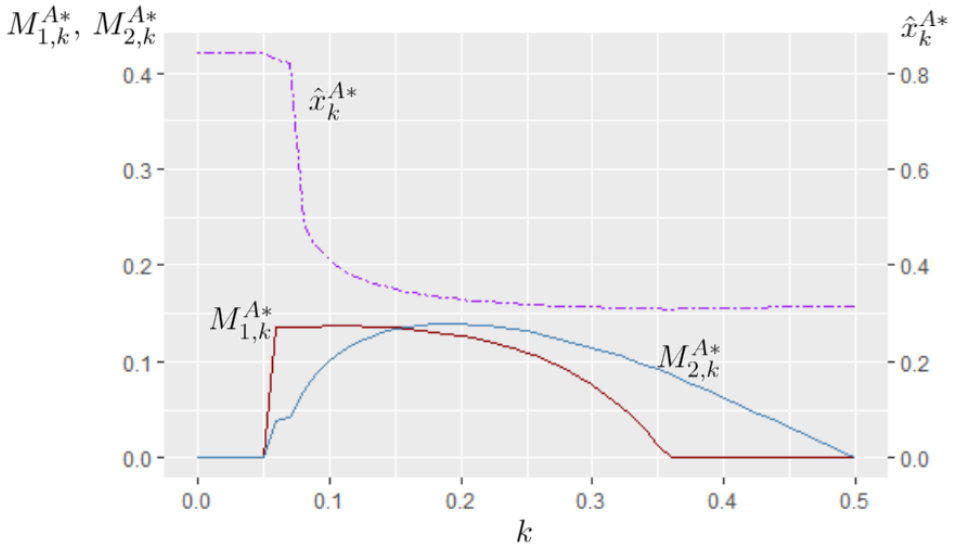
(1) the platform prunes more of the moderate content than it does of the extreme content, i.e., there exists $k_0 \in [0, \frac{1}{2}]$ such that $M_{1,k_0}^{A*} < M_{2,k_0}^{A*}$, and

(2) the average extremeness index of the content on the platform is lower than when it prunes more of the extreme content and when it does not moderate content, i.e., there exists $k_1, k_2 \in [0, \frac{1}{2}]$ such that $M_{1,k_1}^{A*} < M_{2,k_1}^{A*}$, $M_{1,k_2}^{A*} > M_{2,k_2}^{A*}$, but $\hat{x}_{k_1}^{A*} > \hat{x}_{k_2}^{A*}$.

Figure A11 illustrates the relationship between k and $M_{1,k}^{A*}$, $M_{2,k}^{A*}$, or $\hat{x}_k^{A*}$.

From Figure A11, we see that for k greater than around 0.15, we have $M_{1,k_0}^{A*} < M_{2,k_0}^{A*}$, so claim (1)

Figure A11: $M_{1,k}^{A*}$, $M_{2,k}^{A*}$, or $\hat{x}_k^{A*}$ and technology accuracy k ($v = 0.2, \alpha = 0.05$, $c = 0.25$)



holds. Also, consider $k_1 = 0.1$ and $k_2 = 0.3$, we see from the figure that $M_{1,k_1}^{A*} < M_{2,k_1}^{A*}$, $M_{1,k_2}^{A*} > M_{2,k_2}^{A*}$, but $\hat{x}_{k_1}^{A*} > \hat{x}_{k_2}^{A*}$, so claim (2) holds.

For a platform under subscription revenues, since the full solution including optimal pricing is analytically challenging (see Section A.3.2), we use numerical simulations which exhaust the parameter space of $\alpha \in [0, 1]$, $v \in [0, \frac{1}{2}]$, and $c \in [v, \alpha + 2v]$ with a grid of 0.05. The details of how to generate the equilibrium outcomes are in Appendix A.4 (especially Section A.4.2). Based on the outcomes stored in Dataframe S (described on page A32), we cannot find any examples where the platform prunes extreme content more than it does moderate content, and we find that the average extremeness index is lower when a platform conducts content moderation than that when it does not.

□

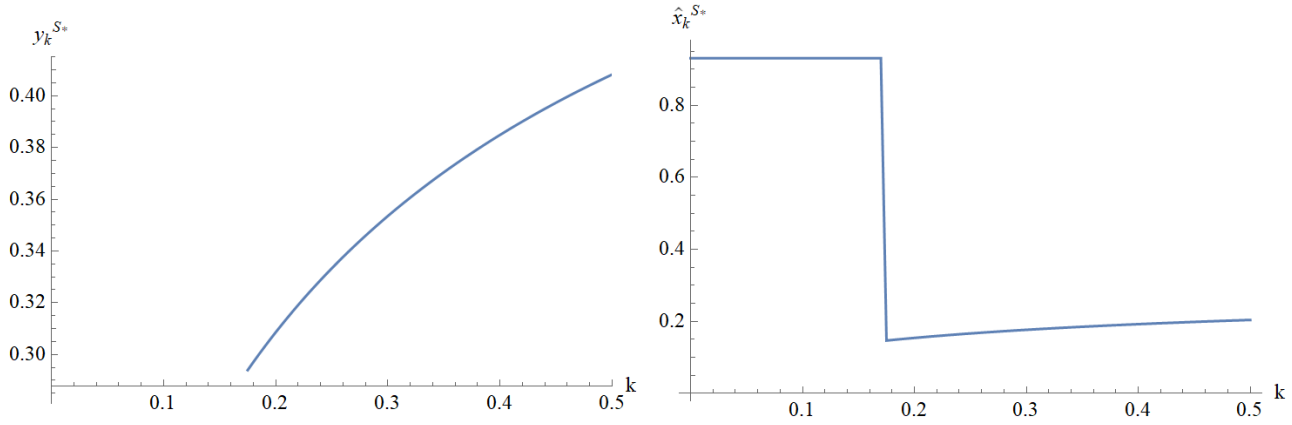
Proof of Proposition 5. First, consider a platform with advertising revenues. When $k > \overline{k}$, no users with $x > y$ will participate in the platform. Based on Lemma A.4, the only two candidates for the optimal content moderation y_k^{A*} are either $y_k^{A*} = 1$ or $y_k^{A*} = \sqrt{\frac{4v-(1-2k)2c}{1+2k}}$.

Let the profit of the platform when it chooses $y = 1$ and $y = \sqrt{\frac{4v-(1-2k)2c}{1+2k}}$ be π_1 and π_2, respectively. When $k = \frac{1}{2}$ (i.e., perfect technology), conducting content moderation is more profitable than not doing so, based on the assumption that $\alpha < \alpha^S < \alpha^A$. Therefore, $\pi_1 < \pi_2$ when $k = \frac{1}{2}$. Since the platform's profit, π_k^A or π_k^S, is continuous in k, there exists a $\hat{k} > \overline{k}$ such that $\pi_1 < \pi_2$ for any $k \in [\hat{k}, \frac{1}{2}]$. That is, the optimal content moderation strategy is $y_k^{A*} = \sqrt{\frac{4v-(1-2k)2c}{1+2k}}$ for any $k \in [\hat{k}, \frac{1}{2}]$. Note that $\frac{\partial(\frac{4v-(1-2k)2c}{1+2k})}{\partial k} = \frac{8(c-v)}{(1+2k)^2} > 0$, so y_k^{A*} is increasing in k, i.e., the platform adopts a more

relaxed standard for content moderation as technology further improves. The average extremeness index is $\hat{x}_k^{A*} = \frac{\int_{\mathcal{X}} x(1-q(x))dx}{\int_{\mathcal{X}}(1-q(x))dx} = \frac{\int_0^{y_k^{A*}} x(\frac{1}{2}+k)dx}{(\frac{1}{2}+k)y_k^{A*}} = \frac{y_k^{A*}}{2}$ is increasing in k since y_k^{A*} is increasing in k.

The solutions under subscription revenues are proven numerically since characterizing the equilibrium as a closed form solutions is analytically not tractable. We show the monotonicity between k and y_k^{S*} or $\hat{x}$ numerically by exhausting the parameter space of $\alpha \in [0,1]$, $v \in [0, \frac{1}{2}]$, and $c \in [v, \alpha+2v]$ with a grid of 0.05. The details of how to generate the equilibrium outcomes are in Appendix A.4 (especially Section A.4.2). Using the outcomes stored in Dataframe S (described on page A32), we can show the relationship between k and the optimal moderation strategy y_k^{S*}, as well as the relationship between k and the average extremeness index $\hat{x}_k^{S*} = \frac{\int_{\mathcal{X}} x(1-q(x))dx}{\int_{\mathcal{X}}(1-q(x))dx} = \frac{\int_{x_{1,k}^{S*}}^{y_k^{S*}} x(\frac{1}{2}+k)dx + \int_{x_{2,k}^{S*}}^{1} x(\frac{1}{2}-k)dx}{(\frac{1}{2}+k)(y_k^{S*}-x_{1,k}^{S*})+(\frac{1}{2}-k)(1-x_{2,k}^{S*})}$ numerically.[22] Figure A12 below is an example when $\alpha = 0$, $v = 0.25$, and $c = 0.5$.

Figure A12: Content moderation policy (y_k^{S*}) and avg. extremeness index ($\hat{x}_k^{S*}$) vs. technology accuracy (k) ($v = 0.25, \alpha = 0, c = 0.5$)



We can see that when k is large, y_k^{S*} and $\hat{x}_k^{S*}$ are increasing in k.[23] The same pattern is repeated for all other (α, v, c) combinations.

□

Proof of Proposition 6. For the advertising case, although Section A.3.1 provides the equilibrium for the imperfect technology case, expressions for the solution are too complicated to analytically derive comparative statics. Therefore, we exhaust the parameter space of $\alpha \in [0,1]$, $v \in [0, \frac{1}{2}]$, and $c \in [v, \alpha+2v]$, using a grid of 0.05, and for any combination of (α, v, c), we find out the maximum

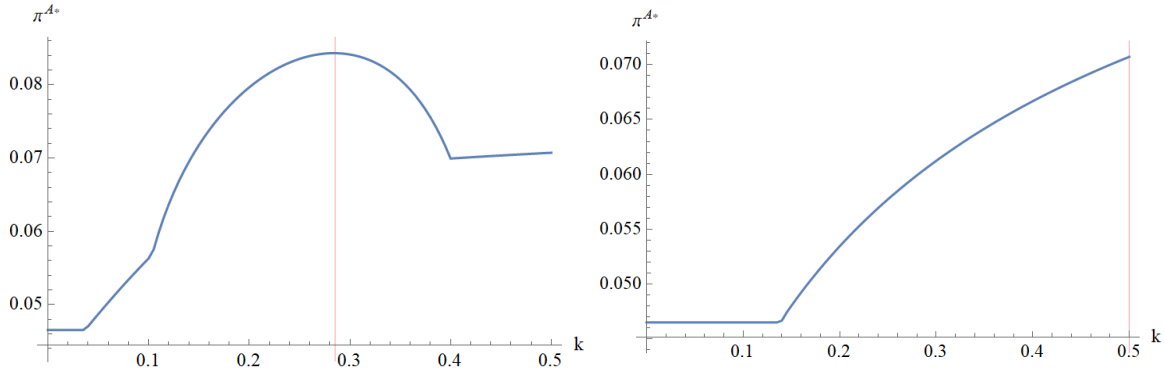
[22] When no content moderation is conducted, the average extremeness index $\hat{x}_k^{S*} = \hat{x}_0^S = \frac{x^S(1,p_1^*)+1}{2}$ when the expressions of $x^S(y,p)$ and p_1^* are given by Equation (5) and the last sentence of Proposition 2, respectively.

[23] In the left subfigure of Figure A12, there is no value of y_k^{S*} for small k, which means no content moderation is conducted in equilibrium when k is small.

profit across different k. Details of the numerical solution are provided in Section A.4.1 of Appendix A.4.

Then we compare the maximum profit ($\max_k \pi_k^{A*}$) with the profit when $k = \frac{1}{2}$ (i.e., perfect technology, $\pi_{k=\frac{1}{2}}^{A*}$). The numerical results confirm that for any α and v, if c is small, we have $\max_k \pi_k^{A*} > \pi_{k=\frac{1}{2}}^{A*}$; otherwise, $\max_k \pi_k^{A*} = \pi_{k=\frac{1}{2}}^{A*}$. Therefore, when c is small, imperfect technology with $k < \frac{1}{2}$ is optimal for a platform under advertising. Figure A13 below illustrates the relationship between k and π_k^{A*} when $\alpha = 0.2$, $v = 0.25$. Specifically, Figure A13a corresponds to the case when c is small while Figure A13b corresponds to the case when c is large. We see that the optimal technology is less than $\frac{1}{2}$ when c is small, but is exactly $\frac{1}{2}$ when c is large. Similar results can be seen for all other combinations of (α, v, c).

Figure A13: Platform profit π_k^{A*} and technology accuracy k ($v = 0.25, \alpha = 0.2$)



(a) $c = 0.3$ (small) (b) $c = 0.5$ (large)

For the subscription case, since the full solution including optimal pricing is analytically challenging (see Section A.3.2), we show for any $k \in [0, \frac{1}{2}]$, the optimal profit π_k^{S*} is weakly increasing in k numerically by exhaustive simulation. Details are provided in Appendix A.4, especially Section A.4.2. Figure A14 below illustrates the relationship between k and π_k^{S*} when $\alpha = 0.2$, $v = 0.25$, and $c = 0.3$. A similar pattern can be seen for all other combinations of (α, v, c).

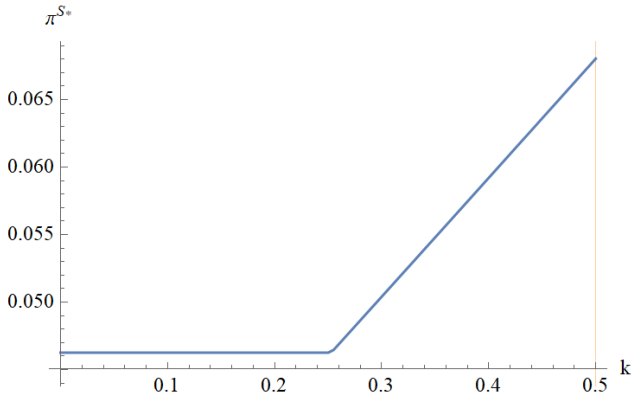
□

Proof of Proposition 7. First, notice that $x^P(y) = x^A(y)$ since the users' behavior is the same as the advertising case. Moreover, based on Equations (3) and (11), the social welfare $W(y)$ is given by

$$W(y) = \begin{cases} \int_0^y \left(\alpha x + v - \frac{1}{2}(y^2 - x^2)\right) dx & \text{if } y < \sqrt{2v}, \\ \int_{-\alpha+\sqrt{y^2+\alpha^2-2v}}^y \left(\alpha x + v - \frac{1}{2}(y^2 - x^2)\right) dx & \text{if } y \geq \sqrt{2v}. \end{cases}$$

If $y < \sqrt{2v}$, then the FOC with respect to y is $\frac{dW(y)}{dy} = \alpha y + v - y^2 = 0$, which yields $y^* = \frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})$. Notice that SOC is satisfied since $\frac{d^2W(y)}{dy^2}|_{y=y^*} = \alpha - 2y^* = -\sqrt{4v + \alpha^2} \leq 0$. The

Figure A14: Platform profit π_k^{S*} and technology accuracy k ($v = 0.25, \alpha = 0.2$, $c = 0.3$)



condition $y < \sqrt{2v}$ requires $\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2}) < \sqrt{2v}$, which is equivalent to $\alpha < \sqrt{\frac{v}{2}}$. In other words, if $\alpha \geq \sqrt{\frac{v}{2}}$, $W(y)$ is increasing in y on $[0, \sqrt{2v}]$.

If $y \geq \sqrt{2v}$, $W(y)$ is increasing in y if $\frac{dW(y)}{dy} = y\left(\sqrt{\alpha^2 - 2v + y^2} - y\right) + v \geq 0$, which is equivalent to $y \geq \frac{v}{\alpha}$. Note that $\frac{d^2W(y)}{dy^2} = \frac{\left(y - \sqrt{\alpha^2 - 2v + y^2}\right)^2}{\sqrt{\alpha^2 - 2v + y^2}} \geq 0$, i.e., $W(y)$ is convex in y. The optimal solution is then either $y^* = 1$ or $y^* = \sqrt{2v}$.

When $\alpha \geq \sqrt{\frac{v}{2}}$, we have $\frac{v}{\alpha} \leq \sqrt{2v}$, so $W(y)$ is increasing in y on $[0, 1]$, and thus $y^{P*} = 1$.

When $\alpha < \sqrt{\frac{v}{2}}$, we have $\frac{v}{\alpha} > \sqrt{2v}$, so $W(y)$ is first increasing in y until $y = \frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})$ when reaching the local optimum $W(\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2}))$, then decreasing until $y = \frac{v}{\alpha}$, and then again increasing in y. So the optimal social welfare is either $W(\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2}))$ or $W(1)$, depending on which one is higher. Content moderation is only conducted when $W(\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})) > W(1)$. To prove the existence of $\alpha^P < \sqrt{\frac{v}{2}}$ such that content moderation is only conducted when $\alpha < \alpha^P$, we define $\Delta W \equiv W(\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})) - W(1)$ and prove the following: (1) ΔW is decreasing in α, (2) $\Delta W > 0$ when $\alpha = 0$, and (3) $\Delta W < 0$ when $\alpha = \sqrt{\frac{v}{2}}$.

(1) ΔW is decreasing in α because

$$\frac{\partial \Delta W}{\partial \alpha} = \frac{1}{4}\left(\alpha\left(5\alpha + \sqrt{\alpha^2 + 4v} - 4\sqrt{\alpha^2 - 2v + 1}\right) - 2v\right) < 0.$$

(2) When $\alpha = 0$,

$$\begin{aligned}
\Delta W|_{\alpha=0} =& W(\sqrt{v}) - W(1) \\
=& \int_0^{\sqrt{v}} (v - \frac{1}{2}(v - x^2))dx - \int_{\sqrt{1-2v}}^1 (v - \frac{1}{2}(1 - x^2))dx \\
=& \frac{1}{3}\left(2v^{3/2} + \left(2\sqrt{1 - 2v} - 3\right)v - \sqrt{1 - 2v} + 1\right).
\end{aligned}$$

Taking derivative w.r.t. v yields

$$\frac{\partial(\Delta W|_{\alpha=0})}{\partial v} = \sqrt{v} + \sqrt{1-2v} - 1,$$

and setting it to zero gives that $v = 0$ or $v = \frac{4}{9}$. Therefore, $\Delta W|_{\alpha=0}$ is increasing in v on $v \in (0, \frac{4}{9})$ and decreasing in v on $v \in (\frac{4}{9}, \frac{1}{2})$. Thus, for any $v \in (0, \frac{1}{2})$

$$\Delta W|_{\alpha=0} > \min\{\Delta W|_{\alpha=0,v=0}, \Delta W|_{\alpha=0,v=\frac{1}{2}}\} = \min\{0, \frac{1}{3}(\frac{3}{2} + \frac{1}{\sqrt{2}})\} = 0.$$

(3) When $\alpha = \sqrt{\frac{v}{2}}$,

$$\begin{aligned}
\Delta W|_{\alpha=\sqrt{\frac{v}{2}}} =& W(\sqrt{2v}) - W(1) \\
=& \int_0^{\sqrt{2v}} (\sqrt{\frac{v}{2}}x + v - \frac{1}{2}(2v - x^2))dx - \int_{\frac{\sqrt{2-3v}-\sqrt{v}}{\sqrt{2}}}^1 (\sqrt{\frac{v}{2}}x + v - \frac{1}{2}(1-x^2))dx \\
=& \frac{1}{12}\left(5\sqrt{2}v^{3/2} + 3\left(\sqrt{4-6v} - 4\right)v - 2\sqrt{4-6v} + 4\right).
\end{aligned}$$

Taking derivative w.r.t. v yields

$$\frac{\partial(\Delta W|_{\alpha=\sqrt{\frac{v}{2}}})}{\partial v} = \frac{1}{8}\left(5\sqrt{2v} + 3\sqrt{4-6v} - 8\right),$$

and setting it to zero gives that $v = \frac{49}{338}$ or $v = \frac{1}{2}$. Therefore, $\Delta W|_{\alpha=\sqrt{\frac{v}{2}}}$ is decreasing in v on $v \in (0, \frac{49}{338})$ and increasing in v on $v \in (\frac{49}{338}, \frac{1}{2})$. Thus, for any $v \in (0, \frac{1}{2})$

$$\Delta W|_{\alpha=\sqrt{\frac{v}{2}}} < \max\{\Delta W|_{\alpha=\sqrt{\frac{v}{2}},v=0}, \Delta W|_{\alpha=\sqrt{\frac{v}{2}},v=\frac{1}{2}}\} = \max\{0, 0\} = 0.$$

Therefore, we claim that there exists $\alpha^P < \sqrt{\frac{v}{2}}$ such that $y^{P*} = \frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})$ if $\alpha < \alpha^P$ while no content moderation is conducted ($y^{P*} = 1$) otherwise.

Since α^P, α^S, and α^A are all single-variable functions of v, one can easily check their relative sizes. Using Mathematica's `FindInstance` function, we show that when $v \in (0, \frac{1}{2})$ and $\alpha \in (0, \sqrt{\frac{v}{2}})$, the intersection set of $\alpha < \alpha^P$ and $\alpha \geq \alpha^S$ is empty. This means $\alpha^P < \alpha^S$ for all $v \in (0, \frac{1}{2})$. We already know that $\alpha^S < \alpha^A$, so we have $\alpha^P < \alpha^S < \alpha^A$.

When the social planner moderates content ($\alpha < \alpha^P$), we have

$$y^{S*} = \sqrt{2v/3} < \sqrt{v} < y^{P*} = \frac{1}{2}(\alpha + \sqrt{4v + \alpha^2}) < \frac{1}{2}(\sqrt{\frac{v}{2}} + \sqrt{4v + (\sqrt{\frac{v}{2}})^2}) = \sqrt{2v} = y^{A*},$$

where the last inequality comes from $\alpha < \alpha^P < \sqrt{\frac{v}{2}}$. This completes the proof.

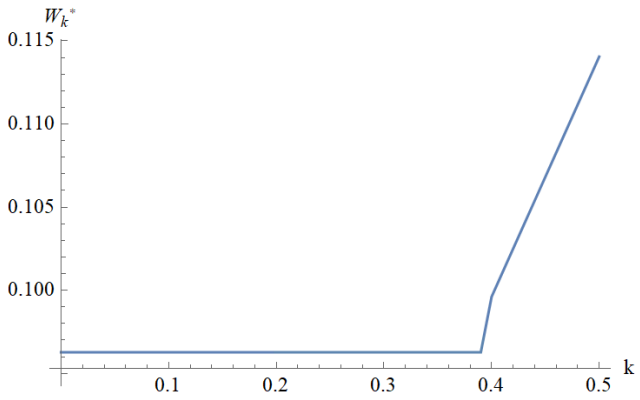
A final check about the claim that "the social planner only prunes users with a negative utility contribution to the society" (on page 24) is as follows. Consider the interior solution $y^{P*} = \frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})$, the net utility contribution of a user at y^{P*} is $\alpha y^{P*} + v - \frac{1}{2}(y^{P*})^2$ where the last term is the total negative utility this user imposes to all other users on the platform. Substituting y^{P*} with $\frac{1}{2}(\alpha + \sqrt{4v + \alpha^2})$, one can find that this net utility contribution is exactly zero.

□

Proof of Proposition 8. We show that for any $k \in [0, \frac{1}{2}]$, the optimal social welfare W_k^* is weakly increasing in k numerically by exhausting the parameter space of $\alpha \in [0, 1]$, $v \in [0, \frac{1}{2}]$, and $c \in [v, \alpha+2v]$ with an increment of 0.05. Details are provided in Appendix A.4, especially Section A.4.2.

Figure A15 below illustrates the relationship between k and W_k^* when $\alpha = 0.2$, $v = 0.25$, and $c = 0.3$. A similar pattern can be seen for all other combinations of (α, v, c).

Figure A15: Optimal social welfare W_k^* and technology accuracy k ($v = 0.25, \alpha = 0.2$, $c = 0.3$)



□

A.3 Equilibrium with Imperfect Technology

A.3.1 Advertising Revenue Model

Lemma A.3 characterizes the user equilibrium under advertising, given a content moderation policy y.

Lemma A.3. *For an ad-supported platform, when $k \in [0, \frac{1}{2}]$ and the platform does content moderation, there exists $x_{2,k}^A \in [0, 1]$ such that all users in range $[x_{2,k}^A, 1]$ participate in the platform. In particular,*

$$x_{2,k}^A = \min\{1, \sqrt{\alpha^2 + 1 + \frac{2c(1+2k) - 4v}{1-2k}} - \alpha\} \equiv \begin{cases} \sqrt{\alpha^2 + 1 + \frac{2c(1+2k)-4v}{1-2k}} - \alpha & \text{if } k < \overline{k}; \\ 1 & \text{if } k \geq \overline{k}, \end{cases} \tag{A17}$$

where $\overline{k}=\frac{\alpha+2v-c}{2(\alpha+c)}$. *For any* $y\in[0,1]$, *there exists* $x^A_{1,k}(y)\in[0,y]$ *such that if* $y<x^A_{2,k}$, *the user set of the platform* $\mathcal{X}^A$ *is* $[x^A_{1,k}(y),y]\cup[x^A_{2,k},1]$; *if* $y\geq x^A_{2,k}$, $\mathcal{X}^A$ *is* $[x^A_{1,k}(y),1]$. *In particular,*

$$x^A_{1,k}(y)=\begin{cases}\sqrt{\alpha^2+\max\left\{0,y^2+\min\{2\alpha y,\frac{(1-2k)(2c+1-(x^A_{2,k})^2)-4v}{1+2k}\}\right\}}-\alpha & \text{if } y<x^A_{2,k};\\ \sqrt{\alpha^2+\max\left\{0,y^2+\min\{2\alpha y,\frac{(1-2k)(2c+1-y^2)-4v}{1+2k}\}\right\}}-\alpha & \text{if } y\geq x^A_{2,k}.\end{cases}\tag{A18}$$

Furthermore, $x^A_{2,k}$ *is increasing in* k *and* $x^A_{1,k}(y)$ *is decreasing in* k.

Proof of Lemma A.3. Based on Equations (7) and (8), we have

$$U(x)=\begin{cases}\alpha x(\frac{1}{2}+k)-c(\frac{1}{2}-k)+v-\int_{\tilde{x}\in\hat{\mathcal{X}},x<\tilde{x}\leq y}\tilde{x}(\frac{1}{2}+k)d\tilde{x}-\int_{\tilde{x}\in\hat{\mathcal{X}},\tilde{x}>y}\tilde{x}(\frac{1}{2}-k)d\tilde{x} & \text{if } x\leq y;\\ \alpha x(\frac{1}{2}-k)-c(\frac{1}{2}+k)+v-\int_{\tilde{x}\in\hat{\mathcal{X}},\tilde{x}>x}\tilde{x}(\frac{1}{2}-k)d\tilde{x} & \text{if } x>y.\end{cases}\tag{A19}$$

Note that there is a discontinuity at $x=y$: $U(y^-)>U(y^+)$ as long as $k>0$. Also, similar to what is shown in the proof of Lemma 1, $U(x)$ is increasing in x on $[0,y]$ and also increasing in x on $(y,1]$. Therefore, there can be possibly two segments of participating users with $U(x)>0$: the moderate users $[x^A_{1,k},y]$ and the extreme users $[x^A_{2,k},1]$.

First consider the extreme users in $(y,1]$. If $U(1)<0$, no users in $(y,1]$ participate in the platform since $U(x)$ is increasing in x. $U(1)=\alpha(\frac{1}{2}-k)-c(\frac{1}{2}+k)+v<0$ is equivalent to $k>\frac{\alpha+2v-c}{2(\alpha+c)}=\overline{k}$. Otherwise, when $k\leq\overline{k}$, all users in $[x^A_{2,k},1]$ participate, where $x^A_{2,k}$ solves $U(x^A_{2,k})=\alpha x^A_{2,k}(\frac{1}{2}-k)-c(\frac{1}{2}+k)+v-\int_{x^A_{2,k}}^1\tilde{x}(\frac{1}{2}-k)d\tilde{x}=0$, which gives $x^A_{2,k}=\sqrt{\alpha^2+1+\frac{2c(1+2k)-4v}{1-2k}}-\alpha$. Therefore, Equation (A17) holds.

For the moderate users in $[0,y]$, $x^A_{1,k}$ can be given by the condition $U(x^A_{1,k})=0$ if the solution to this condition is interior ($0<x^A_{1,k}<y$). Denote the interior solution as $\tilde{x^A_{1,k}}$. Depending on whether $y<x^A_{2,k}$ or $y\geq x^A_{2,k}$, the condition $U(\tilde{x^A_{1,k}})=0$ is given by

$$\begin{cases}\alpha\tilde{x^A_{1,k}}(\frac{1}{2}+k)-c(\frac{1}{2}-k)+v-\int_{\tilde{x^A_{1,k}}}^y\tilde{x}(\frac{1}{2}+k)d\tilde{x}-\int_{x^A_{2,k}}^1\tilde{x}(\frac{1}{2}-k)d\tilde{x}=0 & \text{if } y<x^A_{2,k};\\ \alpha\tilde{x^A_{1,k}}(\frac{1}{2}+k)-c(\frac{1}{2}-k)+v-\int_{\tilde{x^A_{1,k}}}^y\tilde{x}(\frac{1}{2}+k)d\tilde{x}-\int_y^1\tilde{x}(\frac{1}{2}-k)d\tilde{x}=0 & \text{if } y\geq x^A_{2,k}.\end{cases}\tag{A20}$$

Solving Equation (A20), we have

$$\tilde{x^A_{1,k}}=\begin{cases}\sqrt{\alpha^2+y^2+\frac{(1-2k)(2c+1-(x^A_{2,k})^2)-4v}{1+2k}}-\alpha & \text{if } y<x^A_{2,k};\\ \sqrt{\alpha^2+y^2+\frac{(1-2k)(2c+1-y^2)-4v}{1+2k}}-\alpha & \text{if } y\geq x^A_{2,k}.\end{cases}\tag{A21}$$

If $0 < \tilde{x^A_{1,k}} < y$, $x^A_{1,k} = \tilde{x^A_{1,k}}$. If $\tilde{x^A_{1,k}} \leq 0$ or $\tilde{x^A_{1,k}} \geq y$, corner solutions apply. I.e.,

$$x^A_{1,k}(y) = \begin{cases} 0 & \text{if } \tilde{x^A_{1,k}} \leq 0; \\ \tilde{x^A_{1,k}} & \text{if } 0 < \tilde{x^A_{1,k}} < y; \\ y & \text{if } \tilde{x^A_{1,k}} \geq y. \end{cases} \tag{A22}$$

Rewriting Equations (A21) and (A22) in a dense format obtains Equation (A18).

With Equations (A17) and (A18), we see immediately that $x^A_{2,k}$ is increasing in k and $x^A_{1,k}(y)$ is decreasing in k, since $(1-2k)$ is decreasing in k and $(1+2k)$ is increasing in k.

□

The following lemma further investigates the platform's optimal content moderation strategy given users' response.

Lemma A.4. *Let* $\hat{y^A_k} \equiv \sqrt{\max\{0, \frac{4v-(1-2k)(2c+1-(x^A_{2,k})^2)}{1+2k}\}}$. *The optimal level of content moderation under advertising revenues* (y^{A*}_k) *can be characterized by the following:*

Case 1. If $k \geq \overline{k}$, *then* $x^A_{2,k} = 1$, *and* y^{A*}_k *is (a)* 1 *if* $k < k^A_1$, *and (b)* $\hat{y^A_k}$ *if* $k \geq k^A_1$.

Case 2. If $k < \overline{k}$ *and* $\hat{y^A_k} < x^A_{2,k}$, *then* $x^A_{2,k} < 1$, *and* y^{A*}_k *is (a)* $x^A_{2,k}$ *if* $k < k^A_2$, *and (b)* $\hat{y^A_k}$ *if* $k \geq k^A_2$.

Case 3. If $k < \overline{k}$ *and* $\hat{y^A_k} \geq x^A_{2,k}$, *then the market can be fully covered with any* $y^{A*}_k \in [x^A_{2,k}, \hat{y^A_k}]$,

where k^A_1, k^A_2 *are constant.*

In Case 3, there are multiple maximizers. As a tie-breaking rule, we assume that the platform will choose the lowest $y^{A*}_k = x^A_{2,k}$ (the most strict policy) to make the platform as moderate as possible.

Proof of Lemma A.4. Since Case 1 ($x^A_{2,k} = 1$) is just a special case of Case 2, we only need to show the following: When $\hat{y^A_k} < x^A_{2,k}$, $y^{A*}_k = x^A_{2,k}$ if $k < k_0$ and $y^{A*}_k = \hat{y^A_k}$ if $k > k_0$, where k_0 is a constant; when $\hat{y^A_k} \geq x^A_{2,k}$, any $y^{A*}_k \in [x^A_{2,k}, \hat{y^A_k}]$ can make the market fully covered ($X^{A*} = 1$).

If $y > x^A_{2,k}$, then

$$\begin{aligned} x^A_{1,k} &= \sqrt{\alpha^2 + \max\{0, y^2 + \min\{2\alpha y, \frac{(1-2k)(2c+1-y^2)-4v}{1+2k}\}\}} - \alpha \\ &= \sqrt{\alpha^2 + \max\{0, \min\{y(y+2\alpha), \frac{(2c+1)(1-2k)+4ky^2-4v}{1+2k}\}\}} - \alpha \end{aligned}$$

is increasing in y, and thus the user base $1 - x^A_{1,k}$ is decreasing in y, so any $y > x^A_{2,k}$ cannot be an optimal choice.

If $y < x^A_{2,k}$, note that $\tilde{x^A_{1,k}} \geq 0$ is equivalent to $y \geq \hat{y^A_k}$ where $\hat{y^A_k} = \sqrt{\frac{4v-(1-2k)(2c+1-(x^A_{2,k})^2)}{1+2k}}$ is determined through solving for y from $\tilde{x^A_{1,k}} = 0$ (which is also equivalent to $U(0) = 0$, by definition of $\tilde{x^A_{1,k}}$).[24] Any $y < \hat{y^A_k}$ also cannot be an optimal choice because if $y < \hat{y^A_k}$ then $x^A_{1,k} = 0$ so the user base is just $y + 1 - x^A_{2,k}$ which is increasing in y.

Therefore, we only consider $y \in [\hat{y^A_k}, x^A_{2,k}]$. In this case, $\frac{\partial \pi^A_k}{\partial y} = \frac{\partial \zeta(y - x^A_{1,k}(y) + 1 - x^A_{2,k})}{\partial y} = \zeta\Big(1 - \frac{y}{\sqrt{\alpha^2 + y^2 + \frac{(1-2k)(2c+1-(x^A_{2,k})^2)-4v}{1+2k}}}\Big) =: \zeta(1 - \frac{y}{L(k)})$, where

$$L(k) \equiv \sqrt{\alpha^2 + y^2 + \frac{(1-2k)(2c+1-(x^A_{2,k})^2) - 4v}{1+2k}}.$$

Note that $x^A_{2,k}$ is increasing in k and thus $L(k)$ is decreasing in k. Therefore, $\frac{\partial \pi^A_k}{\partial y}$ is decreasing in k. Furthermore, $\frac{\partial \pi^A_k}{\partial y}|_{k=\frac{1}{2}} < 0$ because $L(\frac{1}{2}) = \sqrt{y^2 + \alpha^2 - 2v} < y$ since $\alpha < \alpha^S < \alpha^A = \sqrt{2v}$. $\frac{\partial \pi^A_k}{\partial y}|_{k=0} > 0$ because

$$\begin{aligned} L(0) &= \sqrt{y^2 + \alpha^2 + 2c + 1 - 4v - (x^A_{2,k=0})^2} \\ &= \sqrt{y^2 + \alpha^2 + 1 + 2c - 4v - (\sqrt{\alpha^2 + 1 + 2c - 4v} - \alpha)^2} \\ &> \sqrt{y^2 + \alpha^2 + 1 + 2c - 4v - (\sqrt{\alpha^2 + 1 + 2c - 4v})^2} \\ &= y. \end{aligned}$$

So there exists k_0 such that $\frac{\partial \pi^A_k}{\partial y} > 0$ when $k < k_0$ and $\frac{\partial \pi^A_k}{\partial y} < 0$ when $k > k_0$. Since $y \in [\hat{y^A_k}, x^A_{2,k}]$, we know that the optimal $y^{A*}_k = x^A_{2,k}$ when $k < k_0$ and $y^{A*}_k = \hat{y^A_k}$ when $k > k_0$.

Note that if $\hat{y^A_k} > x^A_{2,k}$, it simply means that the market can be fully covered by choosing any $y^{A*}_k \in [x^A_{2,k}, \hat{y^A_k}]$. This is because every user with $x \leq y$ participates when $y \leq \hat{y^A_k}$, and every user with $x \geq x^A_{2,k}$ participates regardless of the choice of y.

□

Lemma A.4 indicates that unless the market is fully covered, there are generally two potential levels of content moderation that the platform can choose. As the technology becomes more accurate, the platform tends to choose the higher level of content moderation (a smaller y).[25] Meanwhile, it points out the possibility that the market is fully covered when the technology is imperfect, and thus the possibility that an imperfect technology may enlarge the market for the platform.

[24] If there is no real number solution to this equation, we set $\hat{y^A_k} = 0$ without loss of generality.

[25] Note that this statement is about the choice between two levels for a given k and that it does not mean y^{A*} is decreasing in k.

A.3.2 Subscription Revenue Model

Similar to the advertising case, Lemma A.5 gives the full characterization of the user equilibrium under advertising, given a content moderation policy y and subscription fee p.

Lemma A.5. *Suppose the subscription fee p is given. For a subscription-supported platform, when $k \in [0, \frac{1}{2}]$ and the platform does content moderation, there exists $x^S_{2,k}(p) \in (0,1]$ such that all users in $[x^S_{2,k}(p), 1]$ participate in the platform. In particular,*

$$x^S_{2,k}(p) = \min\{1, \sqrt{\alpha^2 + 1 + \frac{2c(1+2k) - 4(v-p)}{1-2k}} - \alpha\}. \tag{A23}$$

Furthermore, for any $y \in [0,1]$, there exists $x^S_{1,k}(y,p) \in [0,y]$ such that if $y < x^S_{2,k}(p)$, the user set of the platform $\mathcal{X}^S$ is $[x^S_{1,k}(y,p), y] \cup [x^S_{2,k}(p), 1]$; if $y \geq x^S_{2,k}(p)$, $\mathcal{X}^S$ is $[x^S_{1,k}(y,p), 1]$. In particular,

$$x^S_{1,k}(y,p) = \begin{cases} \sqrt{\alpha^2 + \max\{0, y^2 + \min\{2\alpha y, \frac{(1-2k)(2c+1-(x^S_{2,k}(p))^2)-4(v-p)}{1+2k}\}\}} - \alpha & \text{if } y < x^S_{2,k}(p); \\ \sqrt{\alpha^2 + \max\{0, y^2 + \min\{2\alpha y, \frac{(1-2k)(2c+1-y^2)-4(v-p)}{1+2k}\}\}} - \alpha & \text{if } y \geq x^S_{2,k}(p). \end{cases} \tag{A24}$$

Proof of Lemma A.5. The proof of Lemma A.5 is the same as that of Lemma A.3 except that we change v to $v - p$. □

Given the response of users, we next derive the optimal content moderation policy y^{S*} and pricing p^{S*}.

The following lemma describes the optimal level of content moderation, $y^{S*}(p)$, for any given p.

Lemma A.6. *Let $\hat{y^S_k}(p) \equiv \sqrt{\max\{0, \frac{4(v-p)-(1-2k)(2c+1-(x^S_{2,k}(p))^2)}{1+2k}\}}$. The optimal level of content moderation under subscription revenues ($y^{S*}_k(p)$) can be characterized by the following:*

Case 1. If $p > p_{1,k}$, then $x^S_{2,k}(p) = 1$, and $y^{S*}_k(p)$ is (a) 1 if $k < k^S_1$, and (b) $\hat{y^S_k}(p)$ if $k \geq k^S_1$.

Case 2. If $p_{2,k} < p \leq p_{1,k}$, then $x^S_{2,k}(p) < 1$, and $y^{S*}_k(p)$ is (a) $x^S_{2,k}(p)$ if $k < k^S_2$, and (b) $\hat{y^S_k}(p)$ if $k \geq k^S_2$.

Case 3. If $p \leq p_{2,k}$, then the market can be fully covered with any $y^{S*}_k(p) \in [x^S_{2,k}(p), \hat{y^S_k}(p)]$.

where k^S_1, k^S_2 are constants, and $p_{1,k}, p_{2,k}$ are constants when k is given.

Proof of Lemma A.6. The proof can be built on that of Lemma A.4. The only difference here is that p is another decision variable of the platform, so whether $x^S_{2,k}(p) < 1$ depends not only on the technology accuracy k but also on the pricing of the platform. Therefore, only the condition for each of the three cases will change.

If $x^S_{2,k}(p)=1$, which is equivalent to $\sqrt{\alpha^2+1+\frac{2c(1+2k)-4(v-p)}{1-2k}}-\alpha>1 \Leftrightarrow p>v+\alpha(\frac{1}{2}-k)-c(\frac{1}{2}+k)$, no users with $x>y$ participate in the platform. Let

$$p_{1,k} \equiv v+\alpha(\frac{1}{2}-k)-c(\frac{1}{2}+k), \tag{A25}$$

and we have found the condition for Case 1.

If $x^S_{2,k}(p)<1$, let $\hat{y^S_k}(p)$ be the content moderation strategy such that $U(0)-p=0$. Solving it, we have $\hat{y^S_k}(p)=\sqrt{\frac{4(v-p)-(1-2k)(2c+1-(x^S_{2,k}(p))^2)}{1+2k}}$.[26] We then only need to check whether $\hat{y^S_k}(p)<x^S_{2,k}(p)$. If so, it corresponds to Case 2; otherwise, it corresponds to Case 3. By Equation (A23), we know that when $x^S_{2,k}(p)<1$, $\frac{dx^S_{2,k}(p)}{dp}=\frac{2/(1-2k)}{\sqrt{\alpha^2+1+\frac{2c(1+2k)-4(v-p)}{1-2k}}}=\frac{2/(1-2k)}{x^S_{2,k}(p)+\alpha}>0$ so $x^S_{2,k}(p)$ is increasing in p. Furthermore, $\text{sign}(\frac{d\hat{y^S_k}(p)}{dp})=\text{sign}(\frac{d[4(v-p)-(1-2k)(2c+1-(x^S_{2,k}(p))^2)]}{dp})=\text{sign}(-4+(1-2k)2x^S_{2,k}(p)\frac{dx^S_{2,k}(p)}{dp})=\text{sign}(-4+4\frac{x^S_{2,k}(p)}{x^S_{2,k}(p)+\alpha})<0$, so $\hat{y^S_k}(p)$ is decreasing in p. Therefore, the condition $\hat{y^S_k}(p)\lesseqgtr x^S_{2,k}(p)$ is equivalent to $p\gtreqless p_{2,k}$, which gives the conditions for Cases 2 and 3. Solving for $p_{2,k}$ by setting $\hat{y^S_k}(p_{2,k})=x^S_{2,k}(p_{2,k})$, we have

$$p_{2,k}=v-\frac{1}{4}-\frac{c\,(12k^2+1)}{2(2k+1)}+\frac{2\alpha(1-2k)k\sqrt{8ck(2k+1)+\alpha^2(1-2k)^2}}{(2k+1)^2}+\frac{1}{2}k\left(1-4\alpha^2\left(1-\frac{8k}{(2k+1)^2}\right)\right). \tag{A26}$$

□

We can see from Lemma A.6 that similar to the advertising model case, there are two potential levels of content moderation and the one with more moderation (smaller y) is more preferred as k increases. Also, we can see that when price is too high, there are no users more extreme than y who participate in the platform. Denote the associated profits in Cases 1(a), 1(b), 2(a), 2(b), and 3 as $\pi^{1a}_k(p)$, $\pi^{1b}_k(p)$, $\pi^{2a}_k(p)$, $\pi^{2b}_k(p)$, and $\pi^3_k(p)$, respectively. It is clear that

$$\pi^{S*}_k=\max\{\max_{p>p_{1,k}}\pi^{1a}_k(p),\max_{p>p_{1,k}}\pi^{1b}_k(p),\max_{p_{2,k}<p\le p_{1,k}}\pi^{2a}_k(p),\max_{p_{2,k}<p\le p_{1,k}}\pi^{2b}_k(p),\max_{p\le p_{2,k}}\pi^3_k(p)\}. \tag{A27}$$

A.4 Numerically Solving for Equilibrium in Imperfect Technology Case

Lemmas A.3 to A.6 in Appendix A.3 give the analytical characterization of the equilibrium in the imperfect technology case. However, to fully solve the equilibrium (especially for the subscription case) and to carry out any further analysis based on the equilibrium are analytically challenging, since the expressions are so complicated that only implicit functions can be provided for equilibrium characterization. Note that the range of each exogenous variable in our model is bounded ($\alpha\in[0,1]$,

[26] If there is no real number solution to this equation, we set $\hat{y^S_k}(p)=0$ without loss of generality.

$v \in [0, \frac{1}{2}]$, and $c \in [v, \alpha + 2v] \subset [0, 2]$), so we can numerically compute the results exhaustively. For each variable, we discretize the range with a grid of 0.05, and enumerate all possible values within the given range. In other words, for α and v, we have values of 0, 0.05, 0.1, 0.15, ..., 0.95, and 1; for c, we have 0, 0.05, ..., 1.95, and 2. For each combination of the parameters, which is denoted as a tuple (α, v, c), we further check whether the following two conditions are satisfied: (1) $v \leq c \leq \alpha + 2v$ and (2) $\alpha \leq \alpha^S$, which is equivalent to $\pi_S(y = \sqrt{\frac{2v}{3}}) \geq \pi_S(y = 1)$. We only proceed if both of them are satisfied.

Then, for any given tuple (α, v, c), we can find out the optimal y^{A*} (in the advertising case) or y^{S*} and p^* (in the subscription case) as a function of k. For k, we also enumerate all possible values between $[0, \frac{1}{2}]$, with a grid of 0.01. To make the search algorithm more efficient, we leverage the analytical results in Appendix A.3.

A.4.1 Advertising

For the advertising case, given any (α, v, c) and k, we can directly get the optimal y_k^{A*} based on the analytical solution provided by Lemma A.4. Based on Lemma A.4, we know that the optimal content moderation policy y_k^{A*} is either

$$y_k^{A*} = x_{2,k}^A \equiv \min\{1, \sqrt{\alpha^2 + 1 + \frac{2c(1+2k) - 4v}{1-2k}} - \alpha\},$$

or

$$y_k^{A*} = \hat{y_k^A} \equiv \sqrt{\max\{0, \frac{4v - (1-2k)(2c + 1 - (x_{2,k}^A)^2)}{1+2k}\}},$$

whichever gives the largest user base. The user base is calculated as $X^{A*} \equiv 1 - x_{2,k}^A + y^{A*} - x_{1,k}^A(y^{A*})$, where the expression for $x_{1,k}^A(y)$ is given by Equation (A18).[27] We thus also determine the equilibrium user configuration (i.e., $x_{1,k}^A$ and $x_{2,k}^A$) and the equilibrium profit (user base) of the platform when it conducts content moderation. Note that we also need to compare this optimal user base (X_k^{A*}) when conducting content moderation with the user base when no moderation is conducted at all (denoted as X_0^A), to see whether the equilibrium strategy is y_k^{A*} or simply no content moderation. The technology accuracy does not matter when no content moderation is conducted, so the user base

$$X_0^A = 1 + \alpha - \sqrt{\alpha^2 + 1 - 2v},$$

which is given in Proposition 1.

So far, we have solved the imperfect technology equilibrium for a platform under advertising, and

[27] In Case 3 of Lemma A.4, there are multiple maximizers. Based on the tie-breaking rule, we know that if $x_{2,k}^A$ and $\hat{y_k^A}$ induce the same profit for the platform, it chooses $y_k^{A*} = \min\{x_{2,k}^A, \hat{y_k^A}\}$.

also calculated the equilibrium quantities.

A.4.2 Subscription

For the subscription case, given any (α, v, c) and k, solving for the equilibrium needs more work. Lemma A.6 only provides a partial equilibrium for any given subscription fee p. To find out the optimal p_k^* as well as the associated y_k^{S*}, we do the following numerical analysis.

There are 5 subcases in Lemma A.6 (Cases 1(a), 1(b), 2(a), 2(b), and 3). We numerically solve a constrained maximization problem over p for each subcase. Let $\pi_k^{1a}(p)$, $\pi_k^{1b}(p)$, $\pi_k^{2a}(p)$, $\pi_k^{2b}(p)$, and $\pi_k^3(p)$ denote the associated profits in Cases 1(a), 1(b), 2(a), 2(b), and 3, respectively. Based Lemma A.6, we can write out the expressions for the profit (objective function) as well as the constraints in each subcase:

- Case 1(a):
  - Objective function: $\pi_k^{1a}(p) = p(1 - x_{1,k}^S(1,p))$.
  - Constraints: $p > p_{1,k}$, $p > 0$.
- Case 1(b):
  - Objective function: $\pi_k^{1b}(p) = p\hat{y_k^S}(p) = p\sqrt{\max\{0, \frac{4(v-p)-(1-2k)2c}{1+2k}\}}$.
  - Constraints: $p > p_{1,k}$, $p > 0$.
- Case 2(a):
  - Objective function: $\pi_k^{2a}(p) = p(1 - x_{1,k}^S(x_{2,k}^S, p))$.
  - Constraints: $p \leq p_{1,k}$, $p > p_{2,k}$, $p > 0$.
- Case 2(b):
  - Objective function: $\pi_k^{2b}(p) = p(\hat{y_k^S}(p)+1-x_{2,k}^S(p)) = p(\sqrt{\max\{0, \frac{4(v-p)-(1-2k)(2c+1-(x_{2,k}^S(p))^2)}{1+2k}\}}+ 1 - x_{2,k}^S(p))$.
  - Constraints: $p \leq p_{1,k}$, $p > p_{2,k}$, $p > 0$.
- Case 3:
  - Objective function: $\pi_k^3(p) = p$.
  - Constraints: $p \leq p_{2,k}$, $p > 0$.

where the expressions for $x_{2,k}^S(p)$, $x_{1,k}^S(y,p)$, $p_{1,k}$ and $p_{2,k}$ are given by Equations (A23), (A24), (A25) and (A26), respectively. Each optimization problem is solved by a direct search algorithm, which is implemented by the `NMaximize` function in Mathematica.[28]

Then we find the maximum across all the subcases (see Equation (A27)), which gives the optimal y^{S*} and p^*, as well as the optimal profit when the platform conducts content moderation. Again, we compare this profit with the profit when no moderation is conducted. If the latter is larger, no content moderation is conducted in equilibrium. Based on Proposition 2, we know that the optimal profit when no content moderation is conducted, π_0^S, is given by

$$\pi_0^S = p_1^*(1+\alpha-\sqrt{\alpha^2+1-2(v-p_1^*)}),$$

where $p_1^* \equiv \frac{1}{9}\Big[(1+\alpha)\sqrt{2(2-3v+2\alpha^2+\alpha)}-2(1-3v+\alpha^2-\alpha)\Big]$.

A.4.3 Social Planner

For the social planner's case, users' response to a given content moderation y is the same as that for a platform under advertising, so all the results in Lemma A.3 hold for the social planner. In other words, the marginal users $x_{2,k}^P \equiv x_{2,k}^A$ and $x_{1,k}^P(y) \equiv x_{1,k}^A(y)$, where the expressions are given in Equations (A17) and (A18). The only difference for a social planner is the objective function, which is no longer the size of user base, but the total welfare of the users, which we denote as $W_k(y)$. As an extension of Equation (11) to the imperfect technology case, we have

$$W_k(y) = \begin{cases} \int_{x_{1,k}^P(y)}^{y} \Big((\frac{1}{2}+k)\alpha x-(\frac{1}{2}-k)c+v-(\frac{1}{2}+k)\frac{1}{2}(y^2-x^2)-(\frac{1}{2}-k)\frac{1}{2}(1-(x_{2,k}^P)^2)\Big)dx \\ \qquad +\int_{x_{2,k}^P}^{1}\Big((\frac{1}{2}-k)\alpha x-(\frac{1}{2}+k)c+v-(\frac{1}{2}-k)\frac{1}{2}(1-x^2)\Big)dx \qquad \text{if } y < x_{2,k}^P, \\ \int_{x_{1,k}^P(y)}^{y} \Big((\frac{1}{2}+k)\alpha x-(\frac{1}{2}-k)c+v-(\frac{1}{2}+k)\frac{1}{2}(y^2-x^2)-(\frac{1}{2}-k)\frac{1}{2}(1-y^2)\Big)dx \\ \qquad +\int_{y}^{1}\Big((\frac{1}{2}-k)\alpha x-(\frac{1}{2}+k)c+v-(\frac{1}{2}-k)\frac{1}{2}(1-x^2)\Big)dx \qquad \text{if } y \geq x_{2,k}^P. \end{cases}$$

The optimal moderation strategy under imperfect technology $y_k^{P*} \in [0,1]$ maximizes $W_k(y)$. Given any (α, v, c) and k, the optimization problem is solve by a direct search algorithm, which is implemented by the `NMaximize` function in Mathematica.

Note that we also need to compare this optimal optimal social welfare when conducting content moderation with the social welfare when no moderation is conducted at all (denoted as W_0), to see

[28] `NMaximize` function in Mathematica uses one of the four direct search algorithms (Nelder-Mead, differential evolution, simulated annealing, and random search), then fine-tunes the solution by using a combination of KKT solution, the interior point, and a penalty method. Source: `https://reference.wolfram.com/language/tutorial/ConstrainedOptimizationComparison.html`.

whether the equilibrium strategy is y_k^{P*} or simply no content moderation. The social welfare without content moderation is given by

$$\begin{aligned} W_0 =& \int_{\sqrt{\alpha^2-2v+1}-\alpha}^{1} \left(v - \frac{1}{2}\left(1-x^2\right) + \alpha x \right) dx \\ =& \frac{1}{3}\left(\alpha^2 \left(\sqrt{\alpha^2-2v+1} - \alpha \right) + \sqrt{\alpha^2-2v+1} + v\left(3\alpha - 2\sqrt{\alpha^2-2v+1} + 3 \right) - 1 \right), \end{aligned}$$

which comes from Equation (11) by plugging in $y = 1$.

So far, under each revenue model or the social planner's problem, we have now developed a dataframe where each row corresponds to a combination of α, v, c, and k. The columns in this dataframe document the equilibrium content moderation strategy y_k^*, the equilibrium profit π_k^* (or social welfare W_k^*), and the equilibrium user configuration (the marginal users $x_{1,k}^*$ and $x_{2,k}^*$). In other words, we obtain three dataframes with the following attributes:

- Dataframe A (advertising): α, v, c, k, y_k^{A*}, π_k^{A*}, $x_{1,k}^{A*}$, and $x_{2,k}^{A*}$;
- Dataframe S (subscription): α, v, c, k, y_k^{S*}, π_k^{S*}, $x_{1,k}^{S*}$, and $x_{2,k}^{S*}$;
- Dataframe P (social planner): α, v, c, k, y_k^{P*}, W_k^*, $x_{1,k}^{P*}$, and $x_{2,k}^{P*}$.

With these tables, we can numerically show all the claims in the Propositions in our main text.

A.5 Preliminary Empirical Evidence

In this appendix, we provide some preliminary empirical evidence for the results generated from our theoretical analysis. Since content moderation is an increasingly important topic getting attention from users, managers, and policy makers, content moderation policies of social media platforms are ever-evolving and dynamic. Therefore, it is difficult to collect a data set that is comprehensive of all platforms. In this analysis, we rely on a list of 103 social media platforms composed by *Influencer Marketing Hub*.[29]

For each social media platform, we collected the texts of their content moderation policy and also information about their major revenue models. The former was copied from a platform's community guidelines or terms of use. The latter was found by searching "[*name of platform*] business model" or "how does [*name of platform*] make money" on Google and referring to the relevant search results. Among the 103 platforms, we focus on those that are published in English language and precluded instant messaging platforms such as WhatsApp since they do not fit our modeling context. We also

[29] "101+ Social Media Sites You Need to Know in 2021." `https://influencermarketinghub.com/social-media-sites/`.

excluded platforms whose content moderation policy information could not be found and/or revenue models are not reported or ambiguous. This reduced the number of social media platforms we analyze to 67.

We hired independent graders from Mechanical Turk to read and decode the text of content moderation policy of each platform. Specifically, each grader was asked to give answers to a set of yes/no questions (given in Table A1) after reading the entire text of a platform's content moderation policy. The first question (Q1) asks whether a platform moderates content at all. Questions Q2-Q10 ask if the platform moderates particular types of content potentially offensive to some users[30].

Table A1: Questions for graders

	Does [*name of platform*]...
Moderation or not:	Q1: ... remove any content posted by users?
Content categories:	Q2: ... remove sexual/adult content such as nudity?
	Q3: ... remove illegal content such as terrorism, drug, arm selling, and etc.?
	Q4: ... remove hate content toward a group (based on race, gender, sexual orientation, and etc.)?
	Q5: ... remove content related to harassment/bullying/threats?
	Q6: ... remove content related to violence/blood/injury?
	Q7: ... remove spam or repeated content?
	Q8: ... remove content that violates others' privacy?
	Q9: ... remove promotional or self-promotional content?
	Q10: ... remove misleading information such as fake pictures, news, and etc.?

At least five graders were assigned to each platform's content moderation policy. An attention-check question saying "regardless of the true answer, check 'No' for this question" was also included in the survey. We precluded the responses which missed this attention-check question. To incentivize graders to give quality answers, we also gave bonus to graders if at least 80% of their responses were consistent with the rest of the graders.

In Propositions 1 and 2, we claimed that a platform under advertising is more likely to conduct content moderation and when conducting content moderation, a platform under subscription does so more aggressively. We show some preliminary evidence of these results based on our data.

Among the 67 platforms, only two platforms do not conduct content moderation (based on the majority response to Q1) and they both adopt subscription as revenue models. For the remaining 65 platforms, we count how many of the 9 categories of content each platform moderates by aggregating the graders' responses. We use two different ways of aggregating questions Q2-Q10:

1. *Percentage (PER)*: the score a platform gets for Qi ($i = 2, 3, ..., 10$) is the share of graders who respond "yes" to this question.

[30]These categories were summarized by the researchers after reading the content moderation policies of around 30 platforms.

2. *Majority rule (MR)*: the score a platform gets for Qi ($i = 2, 3, ..., 10$) is 1 if at least 50% of the graders respond "yes" to this question, and 0 otherwise.

Then we sum up all the scores a platform gets across questions Q2-Q10, which gives the number of content categories each platform moderates, denoted as $categories_{PER}$ or $categories_{MR}$. A higher $categories_{PER}$ or $categories_{MR}$ indicates a stricter content moderation policy. We also define a dummy variable AD for each platform which takes the value 1 if the platform's major revenue source is advertising, and 0 if subscription.

We run the following regressions across the 65 platforms which conducts content moderation:

$$categories_{PER} = \beta_0 + \beta_1 AD + \epsilon, \tag{A28}$$

$$categories_{MR} = \beta_0' + \beta_1' AD + \epsilon'. \tag{A29}$$

Based on our theoretical results, a platform under subscription moderates content more aggressively than one under advertising given that it moderates content, so we expect the signs of β_1 and β_1' to be negative. The regression results are shown in Table A2.

Table A2: Regression results

	Dependent variable:	
	$categories_{PER}$	$categories_{MR}$
	(1)	(2)
AD	-0.235	-0.751^{**}
	(0.294)	(0.365)
Constant	7.365^{**}	8.516^{**}
	(0.213)	(0.264)
Observations	65	65
R^2	0.010	0.063
Note:		**p<0.05

We see that the directions of estimated β_1 and β_1' are as expected, which provides preliminary support for our theoretical predictions. The estimate based on the majority rule aggregation is significant and the one based on percentage aggregation is in the same direction but less precisely estimated. We anticipate the $categories_{PER}$ measure to be noisier than $categories_{MR}$ since the latter focuses on the response on which the graders reach a consensus and the former does not require that. This would explain $categories_{PER}$ measure is less precisely estimated.