Information Systems Research
ISR-2024-1140

Content Moderation with Shadowbanning

Afrouz Hojati
Haskayne School of Business, University of Calgary, Canada, afrouz.hojati@ucalgary.ca
2500 University Dr NW, Calgary, AB T2N 1N4

Barrie R. Nault
Haskayne School of Business, University of Calgary, Canada, nault@ucalgary.ca
2500 University Dr NW, Calgary, AB T2N 1N4

Online Appendix

Appendix A: Shadowbanning with Common User Beliefs

A.1. Proof of Lemma 1

From (3), we have:

$$E(U^S(x)) = \begin{cases} \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + \left[\alpha x + v - \int_x^y z\,dz\right]\theta & if x \leq y \\ \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + [-c+v]\,\theta & if x > y. \end{cases} \tag{A1}$$

$E(U^S(x))$ is a piece-wise function with a discontinuity at $x = y$. We refer to the first segment as less-extreme users in the range $[0, y]$ with an indifferent user, denoted by $\hat{x}_1^S(\theta, y)$, and the second segment as more-extreme users in the range $(y, 1]$, with an indifferent user, denoted by $\hat{x}_2^S(\theta)$. $E(U^S(x))$ is increasing in x on both ranges of $[0, y]$ and $(y, 1]$, thus, users participating on the platform belong to the range of $[\hat{x}_1^S(\theta, y), y]$ and $[\hat{x}_2^S(\theta), 1]$, respectively.

We start with the second segment, or users in $(y, 1]$. We solve for $\hat{x}_2^S(\theta)$ from $E(U^S(x_2^S(\theta))) = 0$, that is

$$E(U^S(x_2^S(\theta))) = \alpha x\,[1-\theta] + v - \theta c - [1-\theta]\int_{x_2^S(\theta)}^1 z\,dz = 0. \tag{A2}$$

Then, $\hat{x}_2^S(\theta)$ is given by

$$\hat{x}_2^S(\theta) = -\alpha + \sqrt{\alpha^2 + 1 - 2v + \frac{2\theta\,[c-v]}{1-\theta}}. \tag{A3}$$

The location of indifferent more-extreme user, $\hat{x}_2^S(\theta)$, falls in range $[0, 1]$. $\hat{x}_2^S(\theta) = 1$ implies that no users with $x > y$ participate on the platform. For some users with $x > y$ to participate, $\hat{x}_2^S(\theta) < 1$

should hold. This leads to the condition $\theta < \frac{v+\alpha}{c+\alpha} = \bar{\theta}$. In other words, when $\theta < \frac{v+\alpha}{c+\alpha} = \bar{\theta}$, all users in range $[\hat{x}_2^S(\theta), 1]$ participate. When $\theta \geq \frac{v+\alpha}{c+\alpha} = \bar{\theta}$, no user with $x > y$ participates on the platform and $\hat{x}_2^S(\theta) = 1$. Thus, $\hat{x}_2^S(\theta)$ is given by

$$\hat{x}_2^S(\theta) = \begin{cases} -\alpha + \sqrt{\alpha^2 + 1 - 2v + \frac{2\theta[c-v]}{1-\theta}} & if \quad \theta < \frac{v+\alpha}{c+\alpha} = \bar{\theta} \\ 1 & if \quad \theta \geq \frac{v+\alpha}{c+\alpha} = \bar{\theta}. \end{cases} \tag{A4}$$

Therefore, (8) holds.

Now, we consider the first segment, or less-extreme users in $[0, y]$. The marginal user $\hat{x}_1^S(\theta, y)$ is given by solving for $x_1^S(\theta, y)$ from $E(U^S(x_1^S(\theta, y))) = 0$ and depends on whether $y < \hat{x}_2^S(\theta)$ or $y \geq \hat{x}_2^S(\theta)$. Thus

$$E(U^S(x_1^S(\theta, y))) = 0 \Rightarrow \begin{cases} \alpha x + v - [1-\theta]\int_x^1 z\,dz - \theta\int_x^y z\,dz = 0 & if \quad y \geq \hat{x}_2^S(\theta) \\ \alpha x + v - [1-\theta]\int_x^y z\,dz - [1-\theta]\int_{\hat{x}_2^S(\theta)}^1 z\,dz - \theta\int_x^y z\,dz = 0 & if \quad y < \hat{x}_2^S(\theta). \end{cases} \tag{A5}$$

Solving (A5) and considering that $0 \leq x_1^S(\theta, y) \leq y$, we have

$$\hat{x}_1^S(\theta, y) = \begin{cases} min\{y, -\alpha + \sqrt{\alpha^2 + 1 - 2v + \theta\,[y^2 - 1]}\} & if \quad y \geq \hat{x}_2^S(\theta) \\ max\{0, -\alpha + \sqrt{\alpha^2 + 1 - 2v - \hat{x}_2^S(\theta)^2 + \theta\,[\hat{x}_2^S(\theta)^2 - 1] + y^2}\} & if \quad y < \hat{x}_2^S(\theta). \end{cases} \tag{A6}$$

Therefore, (9) holds. We can simply show that $\partial\hat{x}_2^S(\theta)/\partial\theta > 0$, and $\partial\hat{x}_1^S(\theta, y)/\partial\theta \leq 0$. □

A.2. Proof of Lemma 2

Case I. We begin with the case when $\theta > \bar{\theta}$ and according to (A4), $\hat{x}_2^S(\theta) = 1$. From (A6), $\hat{x}_1^S(\theta, y) = -\alpha + \sqrt{\alpha^2 - 2v + y^2}$. Because $\hat{x}_1^S(\theta, y) \geq 0$, then $y \in [\sqrt{2v}, 1]$. The user base of the platform is $y - \hat{x}_1^S(\theta, y)$, the profit is given by $\pi = \zeta\,[y - \hat{x}_1^S(\theta, y)]$, and

$$FOC: \frac{\partial\pi(y)}{\partial y} = \zeta\left[1 - \frac{y}{\sqrt{\alpha^2 - 2v + y^2}}\right] = 0. \tag{A7}$$

If $\alpha \geq \sqrt{2v}$, then $\partial\pi(y)/\partial y \geq 0$. Therefore, the optimal moderation threshold is $y^{S*}(\theta) = 1$. If $\alpha < \sqrt{2v}$, then $\partial\pi(y)/\partial y < 0$ and the optimal moderation threshold is $y^{S*} = \sqrt{2v}$.

Case II. Depending on whether y is greater or less than $\hat{x}_2^S(\theta)$, we have two segments of users (shown in Figure 2). If $y \geq \hat{x}_2^S(\theta)$, then the platform's user base is $1 - \hat{x}_1^S(\theta, y)$ and the profit is given by $\pi = \zeta\,[1 - \hat{x}_1^S(\theta, y)]$. From (A6), $\hat{x}_1^S(\theta, y)$ is given by $min\{y, -\alpha + \sqrt{\alpha^2 + 1 - 2v + \theta\,[y^2 - 1]}\}$, and is increasing in y. Thus, any $y > \hat{x}_2^S(\theta)$ is not optimal. If $y \leq \hat{x}_2^S(\theta)$, then note that $\hat{x}_1^S(\theta, y) \geq 0$ is equivalent to $y \geq \hat{y}^S(\theta)$ where $\hat{y}^S(\theta) = \sqrt{\theta - 1 + 2v + \hat{x}_2^S(\theta)^2 - \theta\hat{x}_2^S(\theta)^2}$ (determined by solving for y from $\hat{x}_1^S(\theta, y) = 0$). Any $y < \hat{y}^S(\theta)$ cannot be an optimal choice because if $y < \hat{y}^S(\theta)$ then $\hat{x}_1^S(\theta, y) = 0$, so the user base is $1 - \hat{x}_2^S(\theta) + y$ which is increasing in y.

Therefore, we only consider $y \in [\hat{y}^S(\theta), \hat{x}_2^S(\theta)]$. The platform profit is $\pi(y) = \zeta\left[1 - \hat{x}_2^S(\theta) + y - \hat{x}_1^S(\theta, y)\right]$, and we have

$$FOC: \frac{\partial \pi(y)}{\partial y} = \zeta \left[1 - \frac{y}{\sqrt{\alpha^2 + 1 - 2v - \hat{x}_2^S(\theta)^2 + \theta\left[\hat{x}_2^S(\theta)^2 - 1\right] + y^2}}\right] = 0. \tag{A8}$$

There is no interior solution for y. There exists a θ_0 where $\partial \pi(y)/\partial y = 0$ $\forall y \in [\hat{y}^S(\theta), \hat{x}_2^S(\theta)]$. θ_0 is given by

$$\theta_0 = \frac{\alpha^4 - \alpha^2\left[-2 + c + 2v\right] + \sqrt{\alpha^8 + \alpha^6\left[1 - 2c - 4v\right] + 4\alpha^2 c^2\left[1 - 2v\right] + 4\alpha^4\left[c^2 + c\left[-1 + v\right] + v^2\right]}}{2\left[\alpha^2 - c^2\right]}. \tag{A9}$$

If $\theta < \theta_0$, then $\partial \pi(y)/\partial y > 0$, and if $\theta \geq \theta_0$, then $\partial \pi(y)/\partial y \leq 0$. As $y \in [\hat{y}^S(\theta), \hat{x}_2^S(\theta)]$, the optimal moderation threshold is $y^{S*}(\theta) = \hat{x}_2^S(\theta)$ if $\theta < \theta_0$, and $y^{S*}(\theta) = \hat{y}^S(\theta)$ if $\theta \geq \theta_0$.

Case III. We now consider when $\hat{y}^S(\theta) \geq \hat{x}_2^S(\theta)$. The moderation threshold y should be in the range $[\hat{x}_2^S(\theta), \hat{y}^S(\theta)]$. Every user with $x \leq y$ participates on the platform when $y \leq \hat{y}^S(\theta)$ and every user with $x \geq \hat{x}_2^S(\theta)$ participates regardless of y. This means that all users participate and the platform covers the market by choosing any $y^{S*}(\theta) \in [\hat{x}_2^S(\theta), \hat{y}^S(\theta)]$. As a tie-breaking rule, we assume the platform selects the lowest optimal moderation threshold, $y^{S*}(\theta) = \hat{x}_2^S(\theta)$, to make the platform as moderated as possible. □

A.3. Proof of Lemma 3

In continuation of the proof for Lemma 2, we proceed to prove Lemma 3. Thus far, we established that the optimal moderation threshold for a platform implementing shadowbanning, denoted as $y^{S*}(\theta)$, may take on one of the following values: $\sqrt{2v}$, 1, $\hat{x}_2^S(\theta)$, or $\hat{y}^S(\theta)$. Because $y^{S*}(\theta) = 1$ implies no moderation, if a platform implements shadowbanning, then the optimal moderation threshold lies within the set $\{\hat{x}_2^S(\theta), \hat{y}^S(\theta)\}$. We can easily show that

$\partial \hat{x}_2^S(\theta)/\partial \alpha < 0$, $\partial \hat{x}_2^S(\theta)/\partial v < 0$, and $\partial \hat{x}_2^S(\theta)/\partial c > 0$,

$\partial \hat{y}^S(\theta)/\partial \alpha < 0$, $\partial \hat{y}^S(\theta)/\partial v > 0$, and $\partial \hat{y}^S(\theta)/\partial c > 0$. □

A.4. Proof of Lemma 4 and Proposition 4

In *Case III* of Lemma 2, we show that if $\hat{y}^S(\theta) \geq \hat{x}_2^S(\theta)$ and $\theta < \bar{\theta}$, then all users participate on the platform implementing shadowbanning. The condition $\hat{y}^S(\theta) \geq \hat{x}_2^S(\theta)$ is equivalent to the condition $c \leq \bar{c}$, where

$$\bar{c} = \frac{2v + \theta - 1 + 2\alpha[1 - \theta]\sqrt{2v\theta - \theta + \theta^2}}{2\theta^2}. \tag{A10}$$

Therefore, if $c \le \bar{c}$, then all users participate on the platform. If $c > \bar{c}$, then users in the range $[0, \hat{x}_1^S(\theta, y)]$ do not participate on the platform and the platform cannot cover the market. Thus, a small enough cost of moderation for users is a necessary condition for a platform to cover the market. We have similar conditions for the participation of all users when the user beliefs that the platform implements shadowbanning are heterogeneous.

We clarify why, in *Case I* or *Case II* of Lemma 2, not all users participate. In these scenarios, we consider $\hat{y}^S(\theta) < \hat{x}_2^S(\theta)$. We demonstrate that if $\theta \ge \theta_0$, then $y^{S*}(\theta) = \hat{y}^S(\theta)$. Consequently, the user base is given by $X^{S*} = [0, \hat{y}^S(\theta)] \cup [\hat{x}_2^S(\theta), 1]$, indicating that users within the range $[0, \hat{y}^S(\theta)] \cup [\hat{x}_2^S(\theta), 1]$ participate on the platform, leaving the market uncovered. However, if $\theta < \theta_0$, then $y^{S*}(\theta) = \hat{x}_2^S(\theta)$ and the platform's user base is $X^{S*} = [\hat{x}_1^{S*}(\theta), 1]$. Due to the condition $\hat{y}^S(\theta) < \hat{x}_2^S(\theta)$ we have $\hat{x}_1^{S*}(\theta) = -\alpha + \sqrt{\alpha^2 - \hat{y}^S(\theta)^2 + \hat{x}_2^S(\theta)^2} > 0$. Thus, the lower bound of the range $[\hat{x}_1^{S*}(\theta), 1]$ is always greater than zero, resulting in an uncovered market. However, when a platform uses content removal, its user base is defined as $X^{R*} = [0, \sqrt{2v}]$, indicating that not all users participate on the platform as $v < 1/2$. □

A.5. Proof of Proposition 2

According to Lemma 2, we know that if $\theta \ge \bar{\theta}$ and $\alpha \ge \sqrt{2v}$, then the optimal moderation threshold is $y^{S*}(\theta) = 1$, corresponding to *Case I.a* of Lemma 2. This represents the only scenario where the moderation threshold equals one, indicating that the platform does not shadowban content. In other cases, the moderation threshold is less than one, and the platform shadowbans content with the extremeness degree greater than $y^{S*}(\theta)$. We need to show that its profit in these cases is greater than no moderation profit.

From (10), (A6), and Lemma 2, the optimal profit of a platform implementing shadowbanning when it chooses to shadowban is either $\pi_{II.a}^{S*}(\theta, y^{S*}) = \zeta[1 - \hat{x}_1^{S*}(\theta, y^{S*})]$, $\pi_{II.b}^{S*}(\theta, y^{S*}) = \zeta[1 - \hat{x}_2^{S*}(\theta) + \hat{y}^S(\theta)]$ or $\pi_{III}^{S*}(\theta, y^{S*}) = \zeta$. We know that the profit of a platform implementing shadowbanning when it covers the market is greater than no moderation profit due to the participation of all users. Thus, to prove the proposition, we only need to show that $\pi_{II.a}^{S*}(\theta, y^{S*})$ and $\pi_{II.b}^{S*}(\theta, y^{S*})$ are greater than $\pi^{N*}(1) = \zeta[1 + \alpha - \sqrt{1 + \alpha^2 - 2v}]$ (refer to Section **??** for detail about $\pi^{N*}(1)$). As $\hat{x}_2^S(\theta)^2 < 1$, the condition $\theta\,[\hat{x}_2^S(\theta)^2 - 1] < 0$ holds. As a result, $\pi_{II.a}^{S*}(\theta, y^{S*}) = \zeta[1 + \alpha - \sqrt{1 + \alpha^2 - 2v + \theta\,[\hat{x}_2^S(\theta)^2 - 1]}] > \zeta[1 + \alpha - \sqrt{1 + \alpha^2 - 2v}]$. Therefore, $\pi_{II.a}^{S*}(\theta, y^{S*}) > \pi^{N*}(1)$. Additionally, utilizing a numerical approach, we can demonstrate that $\pi_{II.b}^{S*}(\theta, y^{S*}) > \pi^{N*}(1)$. Therefore, the profit of a platform implementing shadowbanning is always greater than that of no moderation. Further details on the numerical analysis are provided in Appendix A.9. □

A.6. Proof of Proposition 3

In Proposition 2, we compare the profit of a platform implementing shadowbanning with that of no moderation. Thus, we only need to compare the profit of a platform from shadowbanning with that of content removal. In equilibrium, if $\alpha \geq \sqrt{2v}$, then a platform employing content removal chooses not to remove any content. Consequently, its profit equals that of a platform with no moderation. As stated in Proposition 2, when $\alpha \geq \sqrt{2v}$, the shadowbanning profit is greater than no moderation profit and hence greater than the content removal profit.

If $\alpha < \sqrt{2v}$, then a platform implementing content removal chooses to remove content because its profit is greater than that of having no moderation. Therefore, we compare the profit of a platform implementing shadowbanning with that of content removal when $\alpha < \sqrt{2v}$. In equilibrium, the optimal profit of a platform implementing shadowbanning, as derived from (10), (A6), and Lemma 2, can take one of four forms: $\pi^{S*}_{I.b}(\theta, y^{S*}) = \zeta\sqrt{2v}$, $\pi^{S*}_{II.a}(\theta, y^{S*}) = \zeta[1 - \hat{x}^{S*}_1(\theta, y^{S*})]$, $\pi^{S*}_{II.b}(\theta, y^{S*}) = \zeta[1 - \hat{x}^S_2(\theta) + \hat{y}^S(\theta)]$, or $\pi^{S*}_{III}(\theta, y^{S*}) = \zeta$. Because shadowbanning profit is greater than content removal profit when the platform covers the market, we only need to compare $\pi^{S*}_{II.a}(\theta, y^{S*})$ and $\pi^{S*}_{II.b}(\theta, y^{S*})$ with $\pi^{R*}(y^{R*}) = \zeta\sqrt{2v}$. However, analytically comparing the optimal profits in equilibrium is challenging due to the complexity of expressions. Utilizing a numerical approach, we can demonstrate that $\pi^{S*}_{II.a}(\theta, y^{S*})$ and $\pi^{S*}_{II.b}(\theta, y^{S*})$ are greater than $\pi^{R*}(y^{R*})$. Further details are provided in Appendix A.9. □

A.7. Proof of Proposition 5

According to Lemma 2, the optimal moderation threshold for a platform implementing shadowbanning, $y^{S*}(\theta)$, is one of the elements in the set $\{\hat{x}^S_2(\theta), \hat{y}^S(\theta), \sqrt{2v}\}$. The characteristics of each shadowbanning moderation threshold are explained in Lemma 3. We begin by outlining the conditions under which each threshold applies, followed by a comparison with the moderation threshold for content removal, or $y^{R*} = \sqrt{2v}$. If $\theta \geq \bar{\theta}$, then the moderation threshold for shadowbanning is given by $y^{S*}(\theta) = \sqrt{2v}$, which is equivalent to the moderation threshold for content removal.

The moderation threshold for shadowbanning equals $\hat{y}^S(\theta)$ under two conditions: 1) when the moderation cost for users is small (i.e., the market is covered), and 2) when the moderation cost for users is large (i.e., the market is not covered) and $\theta_0 \geq \theta$. In these scenarios, the shadowbanning moderation threshold $y^{S*}(\theta) = \hat{y}^S(\theta)$ is always smaller than that of content removal. This is because the value of $\hat{y}^S(\theta) = \sqrt{2v + [1-\theta][\hat{x}^S_2(\theta)^2 - 1]}$ is smaller than $\sqrt{2v}$, due to $[1-\theta][\hat{x}^S_2(\theta)^2 - 1] < 0$.

Additionally, the moderation threshold for shadowbanning equals $\hat{x}^S_2(\theta)$ when the market is not covered and $\theta < \theta_0$. However, $\hat{x}^S_2(\theta)$ can be either larger or smaller than $\sqrt{2v}$. Solving the inequality $\hat{x}^S_2(\theta) > \sqrt{2v}$ using (8), we find that $\hat{x}^S_2(\theta) > \sqrt{2v}$ when either α or v is small, and $\hat{x}^S_2(\theta) < \sqrt{2v}$ when either α or v is large. A summary of these conditions is provided in Table 1. □

A.8. Proof of Propositions 6 and 7

User surplus is the sum of utilities for all participating users given by (12). We follow the structure of Lemma 2 in order to compute the consumer surpluses in equilibrium. Notably, consumer surplus in *Case I.a* equals the consumer surplus for no moderation, $CS^{S*}_{I.a}(\theta, y^*) = CS^{N*}(1)$. The expressions for consumer surpluses with content removal and shadowbanning are as follows:

$$CS^{R*}(y^*) = \int_0^{\sqrt{2v}} \left[\alpha x + v - \int_x^{\sqrt{2v}} z\,dz\right] dx = \alpha v + \frac{1}{3} v\sqrt{2v}, \tag{A11}$$

$$\begin{aligned} CS^{S*}_{I.a}(\theta, y^*) &= \int_{-\alpha+\sqrt{\alpha^2-2v+1}}^{1} \left[\alpha x + v - \int_x^1 z\,dz\right] dx \\ &= \frac{1}{3}\left[-\alpha^3 + \alpha^2\sqrt{\alpha^2-2v+1} + \sqrt{\alpha^2-2v+1} - 2v\sqrt{\alpha^2-2v+1} + 3v + 3\alpha v - 1\right], \end{aligned} \tag{A12}$$

$$\begin{aligned} CS^{S*}_{I.b}(\theta, y^*) &= \int_0^{\sqrt{2v}} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta\left[\alpha x + v - \int_x^{\sqrt{2v}} z\,dz\right]\right] dx \\ &= \alpha v + \frac{1}{3} v\sqrt{2v} - [1-\theta][1-2v], \end{aligned} \tag{A13}$$

$$\begin{aligned} CS^{S*}_{II.a}(\theta, y^*) = &\int_{\hat{x}_1^{S*}(\theta, y^*)}^{\hat{x}_2^{S*}(\theta)} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta\left[\alpha x + v - \int_x^{x_2^{S*}(\theta)} z\,dz\right]\right] dx \\ &+ \int_{\hat{x}_2^{S*}(\theta)}^{1} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta[v-c]\right] dx, \end{aligned} \tag{A14}$$

$$\begin{aligned} CS^{S*}_{II.b}(\theta, y^*) = &\int_0^{\hat{y}^{S*}(\theta)} \left[[1-\theta]\left[\alpha x + v - \int_x^{\hat{y}^{S*}(\theta)} z\,dz - \int_{\hat{x}_2^{S*}(\theta)}^1 z\,dz\right] + \theta\left[\alpha x + v - \int_x^{\hat{y}^{S*}(\theta)} z\,dz\right]\right] dx \\ &+ \int_{\hat{x}_2^{S*}(\theta)}^{1} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta[v-c]\right] dx, \end{aligned} \tag{A15}$$

$$\begin{aligned} CS^{S*}_{III}(\theta, y^*) = &\int_0^{\hat{y}^{S*}(\theta)} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta\left[\alpha x + v - \int_x^{\hat{y}^{S*}(\theta)} z\,dz\right]\right] dx \\ &+ \int_{\hat{y}^{S*}(\theta)}^{1} \left[[1-\theta]\left[\alpha x + v - \int_x^1 z\,dz\right] + \theta[v-c]\right] dx. \end{aligned} \tag{A16}$$

From (12), social welfare is computed using the above expressions of consumer surpluses and the platform profit from (10). For brevity, we do not provide the expressions for social welfare.

We can easily show that $CS^{S*}_{I.b}(\theta, y^*) < CS^{R*}(y^*)$, however, analytically comparing the remaining consumer surpluses and social welfare in equilibrium is challenging due to the complexity of expressions. Utilizing a numerical approach, we can demonstrate our claims in Propositions 6 and 7 in Section 6. Further details are provided in Appendix A.9. □

A.9. Numerically Comparing the Equilibrium Quantities (e.g., profits)

We perform numerical computations to determine equilibrium quantities for all moderation strategies. For each strategy, we construct a dataframe where each row corresponds to a combination of parameters α, v, c, and θ (or β). Our computation process involves an exhaustive enumeration of all feasible parameter values within defined ranges of each parameter. To achieve this, we employ nested for-loops, incrementing each parameter value by 0.2 at each iteration. As a result, our approach generates 2.8 million rows in each dataframe, representing a comprehensive exploration of the parameter space. The columns of the dataframe record equilibrium quantities such as profit, consumer surplus, social welfare, and moderation threshold. This yields four separate dataframes:

1. Dataframe N (No moderation): α, v, c, θ, y^{N*}, π^{N*}, CS^{N*}, W^{N*}.
2. Dataframe R (Content Removal): α, v, c, θ, y^{R*}, π^{R*}, CS^{R*}, W^{R*}.
3. Dataframe S (Shadowbanning with common beliefs, θ): α, v, c, θ, y^{S*}, π^{S*}, CS^{S*}, W^{S*}.
4. Dataframe B (Shadowbanning with heterogeneous beliefs, β): α, v, c, β, y^{S*}, π^{S*}, CS^{S*}, W^{S*}.

We then verify whether the specific conditions associated with each comparison of equilibrium quantities, detailed in Lemma 2, are satisfied. These conditions guide our analysis, allowing us to numerically confirm the claims presented in the propositions. Details of computations and results are available from the authors.

Appendix B: Shadowbanning with Heterogeneous User Beliefs

B.1. User Decision

With heterogeneous user beliefs, the expected utility that users gain is as follows:

$$E(U^S(x)) = \begin{cases} \left[\alpha x + v - \int_x^1 z\,dz\right][1-\beta x] + \left[\alpha x + v - \int_x^y z\,dz\right]\beta x & if x \le y \\ \left[\alpha x + v - \int_x^1 z\,dz\right][1-\beta x] + [-c+v]\,\beta x & if x > y. \end{cases} \tag{B1}$$

We start by characterizing the equilibrium configuration for users for any given content moderation threshold, y. With heterogeneous user beliefs, we have three indifferent users denoted by $\hat{x}^S_{2,1}(\beta)$, $\hat{x}^S_{2,2}(\beta)$ and $\hat{x}^S_1(\beta, y)$. The first two, $\hat{x}^S_{2,1}(\beta)$ and $\hat{x}^S_{2,2}(\beta)$ are the locations of the indifferent users for the group of users that the degree of extremeness of their content is greater than the platform's moderation threshold, $x > y$, and are thus targeted for shadowbanning. User $\hat{x}^S_{2,1}(\beta)$ participate on the platform and any user with a degree of extremeness greater than $\hat{x}^S_{2,1}(\beta)$ participates until reaching $\hat{x}^S_{2,2}(\beta)$. However, users with $x > \hat{x}^S_{2,2}(\beta)$ refrain from platform participation due to negative utility. The next indifferent user is denoted by $\hat{x}_1^S(\beta, y)$ and represents the location for the indifferent user whose degree of extremeness is smaller than the moderation threshold decided by the platform, $x < y$. Any user located in the range of $[\hat{x}_1^S(\beta, y), y]$ participates, and the platform does not shadowban their content.

Lemma B1. *For any $y \in [0,1]$, there exists $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$ such that, in equilibrium, all users in the range $[\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$ participate on the platform; and there exists $\hat{x}_1^S(\beta, y) \in [0, y]$ such that, all users in range $[\hat{x}_1^S(\beta, y), y]$ participate on the platform. Therefore, the user base of the platform is $X^S = [\hat{x}_1^S(\beta, y), y] \cup [\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$.*

B.1.1. Proof of Lemma B1 We first consider the second segment, or more-extreme users in range $[y, 1]$. Following (3), the $E(U^S(x))$ is not monotonic in x on the range $[y, 1]$ and is first increasing in x and then decreasing in x for $x > 0$. That is why we have two indifferent more-extreme users, $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$. Users in the range $[\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$ participate on the platform because their utility is positive. We solve for $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$ from $E(U^S(x_2^S(\beta))) = 0$.

$$E(U^S(x_2^S(\beta))) = \left[\alpha x + v - \int_x^1 z\,dz\right][1 - \beta x] + [-c + v]\,\beta x = 0 \tag{B2}$$

is equivalent to

$$-\frac{\beta x^3}{2} + \frac{x^2}{2} - \alpha\beta x^2 + \alpha x - \beta c x + \frac{\beta x}{2} + v - \frac{1}{2} = 0. \tag{B3}$$

This cubic polynomial equation can have three real roots and two of them are positive.[1] Thus, $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$ are given by[2]

$$\hat{x}_{2,1}^S = \frac{2}{3\beta}\sqrt{K}\cos\left(\frac{1}{3}\arccos\left(\frac{A}{K\sqrt{K}}\right)\right), \qquad \text{and} \qquad \hat{x}_{2,2}^S = \frac{2}{3\beta}\sqrt{K}\cos\left(\frac{1}{3}\pi + \frac{1}{3}\arccos\left(\frac{A}{K\sqrt{K}}\right)\right) \tag{B4}$$

where $K = 1 + 2\alpha\beta + \beta^2[4\alpha^2 + 3 - 6c]$, and $A = 1 - 3\beta^2[3 + \alpha + 2\alpha^2 + 3c - 9v] + \alpha\beta^3[-9 - 8\alpha^2 + 18c]$. We know that $\hat{x}_{2,1}^S(\beta) \le \hat{x}_{2,2}^S(\beta)$, $\hat{x}_{2,1}^S(\beta) \le 1$ and $\hat{x}_{2,2}^S(\beta) \le 1$ should hold. Therefore, considering the corner solutions, $\hat{x}_{2,1}^S(\beta)$, $\hat{x}_{2,2}^S(\beta)$, and $\hat{x}_1^S(\beta, y)$ are given by

$$\hat{x}_{2,1}^S(\beta) = \min\{1, \max\{0, \frac{2}{3\beta}\sqrt{K}\cos\left(\frac{1}{3}\arccos\left(\frac{A}{K\sqrt{K}}\right)\right)\}\}, \tag{B5}$$

[1] Assume p, q, and r are three roots for the equation $ax^3 + bx^2 + cx + d$. There exists a relation between roots and the coefficients of a cubic polynomial equation, which is $pqr = -d/a$. The value of pqr is $[-1 + 2v]/\beta$ for our equation and is negative. Thus, the three roots are either all negative or one of them is negative and the other two roots are positive.

[2] "When a cubic equation with real coefficients has three real roots, the formulas expressing these roots in terms of radicals involve complex numbers". In order to obtain purely real expressions of the solutions trigonometric functions should be used, specifically in terms of cosines and arccosines. (Source: `https://en.wikipedia.org/wiki/Cubib_equation`).
We expressed $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$ using trigonometric functions. However, they can also be expressed as: $x_k = -\frac{1}{3a}\left(b + \xi^k C + \frac{\Delta_0}{\xi^k C}\right)$ where $\xi = \frac{-1+\sqrt{-3}}{2}$ and $k \in \{0, 1\}$.

$$\hat{x}^S_{2,2}(\beta) = \min\{1, \frac{2}{3\beta}\sqrt{K}\cos(\frac{1}{3}\pi + \frac{1}{3}\arccos(\frac{A}{K\sqrt{K}}))\}, \tag{B6}$$

$$\hat{x}^S_1(\beta, y) = \begin{cases} \max\{0, -\alpha - \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - y^2\right] + \sqrt{\left[\alpha + \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - y^2\right]\right]^2 + \hat{x}^S_{2,2}(\beta)^2 - 2v}\} & \text{if } y \leq \hat{x}^S_{2,1}(\beta) \\ \max\{0, -\alpha - \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right] + \sqrt{\left[\alpha + \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right]\right]^2 + \hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2 + y^2 - 2v}\} & \\ & \text{if } \hat{x}^S_{2,1}(\beta) < y. \end{cases} \tag{B7}$$

We now consider the first segment, or less-extreme users in range $[0, y]$. The $E(U^S(x))$ is increasing in x on the range $[0, y]$ and the indifferent user $\hat{x}^S_1(\beta, y)$ is given by $E(U^S(x^S_1(\beta, y)) = 0$. Depending on whether $x^S_{2,1}(\beta, y) \geq y$ or $x^S_{2,1}(\beta, y) < y$ and from (3), the condition $E(U^S(x^S_1(\beta, y)) = 0$ is given by

$$E(U^S(x)) = 0 \Rightarrow \begin{cases} \left[\alpha x + v - \int_x^{\hat{x}^S_{2,2}(\beta)} z\,dz\right][1 - \beta x] + \left[\alpha x + v - \int_x^y z\,dz\right]\beta x = 0 & if \quad y > \hat{x}^S_{2,1}(\beta) \\ \left[\alpha x + v - \int_x^y z\,dz - \int_{\hat{x}^S_{2,1}}^{\hat{x}^S_{2,2}(\beta)} z\,dz\right][1 - \beta x] + \left[\alpha x + v - \int_x^y z\,dz\right]\beta x = 0 & if \quad y \leq \hat{x}^S_{2,1}(\beta). \end{cases} \tag{B8}$$

By solving for (B8) and taking corner solutions into account, $\hat{x}^S_1(\beta, y)$ is given as expressed in (B7). Therefore, Lemma B1 holds. □

The locations of the indifferent more-extreme users, $\hat{x}^S_{2,1}(\beta)$ and $x^S_{2,2}(\beta)$, change with β and do not change with y. That is, users in the range $[\hat{x}^S_{2,1}(\beta), x^S_{2,2}(\beta)]$ stay on the platform based on their beliefs that the platform implements shadowbanning, β, and not the platform's moderation threshold y. However, the location of the indifferent less-extreme user $\hat{x}^S_1(\beta, y)$, changes with both β and y. Stricter shadowbanning moderation (smaller value for y) decreases $\hat{x}^S_1(\beta, y)$, attracting more users with less extreme content to participate, thus increasing the platform's user base. These results closely resemble the results observed when the platform implements shadowbanning with common user beliefs.

Lemma B1 shows that the less-extreme users in the range $[\hat{x}^S_1(\beta, y), y]$ and more-extreme users in the range $[\hat{x}^S_{2,1}(\beta), \hat{x}^S_{2,2}(\beta)]$ participate on the platform. Two main configurations for the platform's user base are possible due to (B7), illustrated in Figure 4. Because $x^S_1(\beta, y) \geq 0$, the participation of less-extreme users is straightforward as users in range $[\hat{x}^S_1(\beta, y), y]$ participate as shown in Figure 4. But the participation of more-extreme users is more complicated. Depending on whether $\hat{x}^S_{2,1}(\beta)$ and $\hat{x}^S_{2,2}(\beta)$ exceed 1, the participation of more-extreme users can lead to different configurations, as illustrated in Figure B1. Defining the conditions under which some or none of the more-extreme users participate on the platform is challenging because the utility function for more-extreme users provided in (B1) is not monotonic with respect to x.

To determine these configurations, we utilize the value of the utility function at $x = 1$, denoted as $E(U^S(1))$, and the slope of the utility function at $x = 1$, denoted as $\partial E(U^S(1))/\partial x$. The values of $\bar{\beta}$ and $\hat{\beta}$ are derived by solving for $E(U^S(1)) = 0$ and $\partial E(U^S(1))/\partial x = 0$, respectively. Furthermore, the value of $\bar{c}$ helps to determine whether $\bar{\beta} \geq \hat{\beta}$ or $\bar{\beta} < \hat{\beta}$. The following lemma assists in specifying the conditions under which users with the degree of extremeness $x > y$ participate on the platform.

Lemma B2. *Extreme users whose degree of extremeness is greater than the platform's moderation threshold, $x > y$, participate on the platform as follows:*

Case I. If $\beta \geq max\{\bar{\beta}, \hat{\beta}\}$, then users in range $[\hat{x}^S_{2,1}(\beta), \hat{x}^S_{2,2}(\beta)]$ participate.

Case II. If $min\{\bar{\beta}, \hat{\beta}\} < \beta < max\{\bar{\beta}, \hat{\beta}\}$ and a) if $c \geq \bar{c}$, then $\hat{x}^S_{2,1}(\beta) = \hat{x}^S_{2,2}(\beta) = 1$ and none of more-extreme users with $x > y$ participate; b) if $c < \bar{c}$, then more-extreme users in range $[\hat{x}^S_{2,1}(\beta), 1]$ participate.

Case III. If $\beta \leq min\{\bar{\beta}, \hat{\beta}\}$, then users in range $[\hat{x}^S_{2,1}(\beta), 1]$ participate.

We first consider *Case I* in Lemma B2 with the user configuration shown in Figure B1b. If the probability of shadowbanning is sufficiently high, $\beta > \max\{\bar{\beta}, \hat{\beta}\}$, then users within the range $[\hat{x}^S_{2,2}(\beta), 1]$ choose not to participate on the platform. This is because, for users with a high degree of extremeness approaching 1, their beliefs that the platform implements shadowbanning increases significantly. Their beliefs are influenced by both β and their degree of extremeness, x. Consequently, the potential disutility they may incur if their content is shadowbanned is substantial enough to deter their participation. However, this is not the case for users in the range $[\hat{x}^S_{2,1}(\beta), \hat{x}^S_{2,2}(\beta)]$. Despite the high value of β and the potential for substantial disutility due to shadowbanning, their lower values of x mitigate their belief of the likelihood of being shadowbanned, as it is proportional to βx. Additionally, they highly expect to incur no disutility by reading more extreme content posted by users within the range $[\hat{x}^S_{2,2}(\beta), 1]$ because the latter choose not to participate. Therefore, the expected disutility users in the range $[\hat{x}^S_{2,1}(\beta), \hat{x}^S_{2,2}(\beta)]$ gain from both their lower beliefs of being shadowbanned and their reading more extreme content remains lower than the expected utility they derive from posting content, which leads to their participation on the platform.

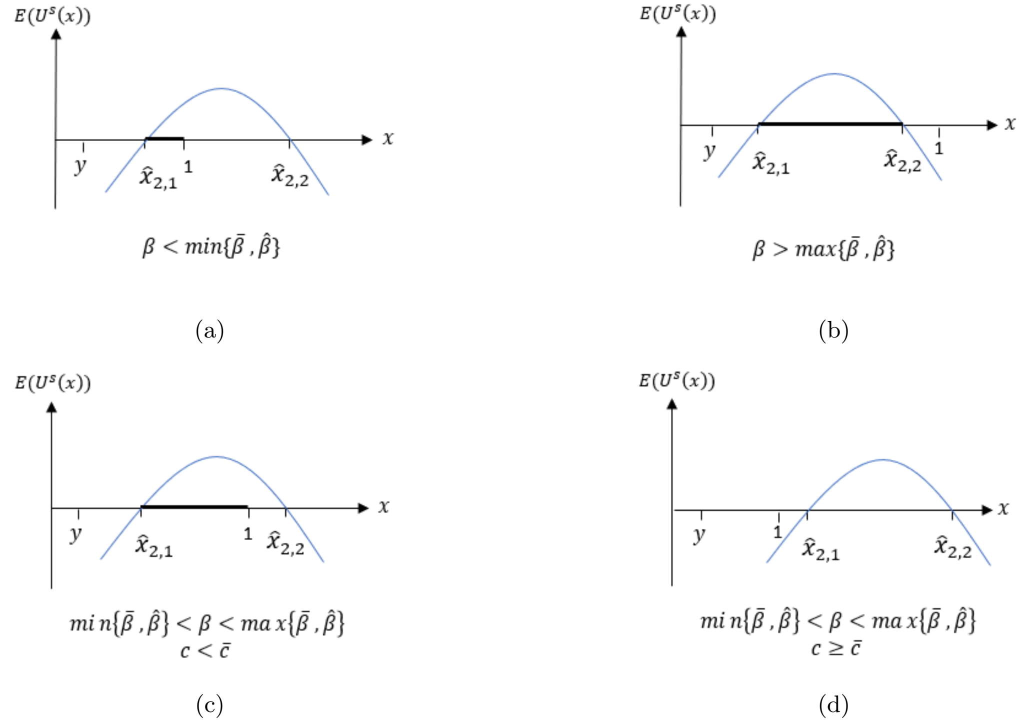
The *Case II* of Lemma B2 characterizes the participation of more-extreme users that hold a moderate belief about the likelihood of shadowbanning, $min\{\bar{\beta}, \hat{\beta}\} < \beta < max\{\bar{\beta}, \hat{\beta}\}$. If the cost to users subject to shadowbanning is low enough, $c < \bar{c}$, then more-extreme users in the range $[\hat{x}^S_{2,1}(\beta), 1]$ participate on the platforms (Figure B1c). This is in contrast to *Case I* where users within the range $[\hat{x}^S_{2,2}(\beta), 1]$ decide not to participate. However, if the cost is sufficiently high, $c \geq \bar{c}$, then more-extreme users choose not to participate (Figure B1d). This is because when the cost is sufficiently high, even with a moderate belief that the platform implements shadowbanning, more-extreme users face a higher expected disutility from being shadowbanned due to the larger value of c; therefore, they decide not to participate.

The *Case III* of Lemma B2 shows that when most extreme users in range $[\hat{x}^S_{2,1}(\beta), 1]$ have low beliefs that the platform implements shadowbanning, they participate (Figure B1a). This is because such users always derive the highest expected utility from the platform which is higher than the expected disutility they face either by their content being shadowbanned or by reading more extreme content.

In summary, the participation of both types of users, more-extreme users with $x > y$ and less-extreme users with $x < y$ are illustrated in Figure 4 in the main text. However, under some conditions, both $\hat{x}^S_{2,1}(\beta)$ and $\hat{x}^S_{2,2}(\beta)$ take the value equal to 1. This means that none of the users with $x > y$ participate on the platform. If $\hat{x}^S_{2,1}(\beta) < 1$ and $\hat{x}^S_{2,2}(\beta)$ takes the value equal to 1, then the user configurations for shadowbanning with heterogeneous user beliefs is similar to the user configurations for shadowbanning with common user beliefs, as illustrated in Figure 2.

B.1.2. Proof of Lemma B2

$E(U^S(x))$ is not monotonic in x, thus the conditions defining the scenario wherein some or none of the more-extreme users participate on the platform are not easily ascertainable because identifying those conditions using $\hat{x}^S_{2,1}(\beta)$ and $\hat{x}^S_{2,2}(\beta)$ is analytically challenging due to the complicated nature of their expressions. Thus, we use the two values of $E(U^S(x))$ and $\partial E(U^S(1))/\partial x$ for $x = 1$ to capture different scenarios. The former is the value of



(a) (b) (c) (d)

Figure B1 Illustration of more-extreme users participation - Shadowbanning with Heterogeneous User Beliefs

the utility function for the user $x=1$, and the latter is the slope of the utility function at the point $x=1$. Solving for $E(U^S(1))=0$ and $\partial E(U^S(1))/\partial x=0$, $\bar{\beta}$ and $\hat{\beta}$ are given by

$$\bar{\beta}=\frac{\alpha+v}{\alpha+c}, \qquad \text{and} \qquad \hat{\beta}=\frac{1+\alpha}{1+2\alpha+c}. \tag{B9}$$

If $\beta>\bar{\beta}$ ($\beta\leq\bar{\beta}$), then $E(U^S(1))<0$ ($E(U^S(1))>0$) meaning that the user $x=1$ has a negative (positive) utility. If $\beta>\hat{\beta}$ ($\beta\leq\hat{\beta}$), then $\partial E(U^S(1))/\partial x<0$ ($\partial E(U^S(1))/\partial x>0$) meaning the slope of utility function at $x=1$ is negative (positive). Another condition that impacts these circumstances is whether $\bar{\beta}\geq\hat{\beta}$ or $\bar{\beta}<\hat{\beta}$. The condition $\bar{\beta}\geq\hat{\beta}$ is equivalent to $c<\frac{\alpha^2+2\alpha v+v}{1-v}=\bar{c}$. Considering all these conditions, we can determine when the more-extreme users participate on the platform.

If the $E(U^S(1))<0$ and $\partial E(U^S(1))/\partial x<0$, then more-extreme users in range $[\hat{x}^S_{2,1}(\beta),\, \hat{x}^S_{2,2}(\beta)]$ participate as shown in Figure B1a. If the $E(U^S(1))>0$ and the slope of the utility function in x=1 is either positive or negative, then more-extreme users in range $[\hat{x}^S_{2,1}(\beta),\, 1]$ participate as shown in Figures B1b and B1c, respectively. If the $E(U^S(1))<0$ and $\partial E(U^S(1))/\partial x>0$, then none of the more-extreme users with $x>y$ participate as shown in Figure B1d. □

B.2. Platform's Moderation Decision

Using Lemma B2 and (5), with heterogeneous user beliefs the profit of the platform implementing shadowbanning is given by

$$\pi^S(y)=\zeta\left[\hat{x}^S_{2,2}(\beta)-\hat{x}^S_{2,1}(\beta)+y-\hat{x}_1(\beta,y)\right]. \tag{B10}$$

We can derive the optimal moderation threshold, and consequently, determine the platform's maximum profit. The following lemma characterizes the optimal moderation threshold for the platform.

Lemma B3. *The optimal moderation threshold for a platform implementing shadowbanning can be characterized by the following:*

Case I. If $\hat{x}^S_{2,1}(\beta)\geq 1$ then any user with $x>y$ does not participate on the platform and

a) if $\alpha\geq\sqrt{2v}$ then $y^{S*}=1$; b) if $\alpha<\sqrt{2v}$ then $y^{S*}=\sqrt{2v}$.

Case II. If $\hat{x}^S_{2,1}(\beta)<1$ and $y<\hat{x}^S_{2,1}(\beta)$ then

a) if $\beta\geq\beta_0$ then $y^{S*}=\hat{x}^S_{2,1}(\beta)$; b) if $\beta<\beta_0$ then $y^{S*}=\hat{y}^S(\beta)$.

Case III. If $\hat{x}^S_{2,1}(\beta)<1$ and $\hat{y}^S(\beta)\geq\hat{x}^S_{2,1}(\beta)$, then $y^{S*}\in[\hat{x}^S_{2,1}(\beta),\hat{y}^S(\beta)]$ and

a) if $\beta\leq\bar{\beta}$ then the market is covered; b) if $\beta>\bar{\beta}$ then the market is not covered.

Where $\hat{y}^S(\beta)=\sqrt{\hat{x}^S_{2,1}(\beta)^2-\hat{x}^S_{2,2}(\beta)^2+2v}$. The value of β_0 is obtained by solving $\partial\pi^S(y)/\partial y=0$, which helps with determining the sign of $\partial\pi^S(y)/\partial y$.

The *Case I* of Lemma B3 is equivalent to *Case II.a* of Lemma B2. That is, more-extreme users with moderate beliefs that the platform implements shadowbanning, $min\{\bar{\beta},\hat{\beta}\} < \beta < max\{\bar{\beta},\hat{\beta}\}$, do not participate on the platform when the cost of being shadowbanned is sufficiently high, $c \geq \bar{c}$. Therefore, a platform with shadowbanning behaves the same way as a platform with content removal. That is, if $\alpha \geq \sqrt{2v}$, then the platform does not shadowban content; if $\alpha < \sqrt{2v}$, then the platform shadowbans extreme content using the same moderation threshold as content removal, leading to equivalent profit.

However, when the user belief in the platform's shadowbanning implementation is either sufficiently high, or sufficiently low, or moderate with a small cost of moderation for users (*Case II* and *III* of Lemma B3), different groups of users participate on the platform and create different user base configurations as discussed earlier. The platform chooses the moderation threshold to maximize its profit and the optimal moderation threshold belongs to the set $\{\hat{x}_{2,1}^S(\beta), \hat{y}^S(\beta), \sqrt{2v}, 1\}$. Notably, if $\hat{y}^S(\beta) \geq \hat{x}_{2,1}^S(\beta)$ and $\beta < \bar{\beta}$ (Case III), then all users participate on the platform, leading to a covered market, $X^S = 1$. Consequently, the platform profit reaches the maximum possible value, specifically $\pi^S(y^{S*}) = \zeta$ with our specific forms, for all optimal moderation thresholds where $y^{S*} \in \left[\hat{x}_{2,1}^S(\beta), \hat{y}^S(\beta)\right]$.

We now elaborate on the two conditions under which a platform can cover the market. First, if $\beta > \bar{\beta}$, users in range $\left[\hat{x}_{2,2}^S(\beta), 1\right]$ do not participate on the platform, as they expect that their content will be shadowbanned. This is the reason why the market cannot be covered. However, if $\beta \leq \bar{\beta}$, then $\hat{x}_{2,2}^S(\beta) \geq 1$, implying that all the more-extreme users starting from $\hat{x}_{2,1}^S(\beta)$ up to 1, participate on the platform. This condition is necessary but not sufficient for a platform to cover the market. We should check the participation of other users, especially less-extreme users. The second condition $\hat{y}^S(\beta) \geq \hat{x}_{2,1}^S(\beta)$ is related to the participation of users in range $[0, \hat{x}_{2,d1}^S]$. If this condition holds, then all users up until the user with location $\hat{x}_{2,1}^S(\beta)$ participate on the platform.

B.2.1. Proof of Lemma B3 From Lemma B2, users participating in the platform belong to the range $[\hat{x}_1^S(\beta, y), y] \cup [\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$.[3] From (B10), the platform profit is given by $\pi^S(y) = \zeta\left[\hat{x}_{2,2}^S(\beta) - \hat{x}_{2,1}^S(\beta) + y - \hat{x}_1(\beta, y)\right]$. Depending on whether $\hat{x}_{2,1}^S(\beta)$ is greater or lower than 1 and whether y is greater or lower than $\hat{x}_{2,1}^S(\beta)$ different cases can happen:

Case I. If $x_{2,1}^S(\beta) \geq 1$, then none of extreme users $(x > y)$ participate on the platform, which is equivalent to $\hat{x}_{2,1}^S(\beta) = \hat{x}_{2,2}^S(\beta) = 1$. From (B7) we know that $\hat{x}_1^S(y) = -\alpha + \sqrt{\alpha^2 - 2v + y^2}$. Therefore, $\hat{y}^S = \sqrt{2v}$ is determined through solving for y from $\hat{x}_1^S(y) = 0$, and as a result, $y \in [\hat{y}^S, 1]$. The user base of the platform is $y - \hat{x}_1(\beta)$ and from (B10) the platform profit is given by

$$\pi^S(y) = \zeta\left[y - \hat{x}_1(\beta)\right] = \zeta\left[y + \alpha - \sqrt{\alpha^2 - 2v + y^2}\right], \tag{B11}$$

[3] Based on Lemma B1, the user base of the platform can also be $[\hat{x}_1^S(\beta, y), y] \cup [\hat{x}_{2,1}^S(\beta), 1]$. As this is the special case of $[\hat{x}_1^S(\beta, y), y] \cup [\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$ when $\hat{x}_{2,2}^S(\beta) = 1$, therefore we only consider $\hat{X}^S = [\hat{x}_1^S(\beta, y), y] \cup [\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$.

and

$$FOC: \frac{\partial \pi^S(y)}{\partial y} = \zeta \left[1 - \frac{y}{\sqrt{\alpha^2 - 2v + y^2}}\right] = 0. \tag{B12}$$

There is no interior solution for y. If $\alpha \geq \sqrt{2v}$, then $\partial \pi^S(y)/\partial y \geq 0$. and the platform profit is increasing in y. Thus, the platform sets the highest value for the optimal moderation threshold, or $y^* = 1$. If $\alpha < \sqrt{2v}$ then $\partial \pi^S(y)/\partial y < 0$ and the platform profit function is decreasing in y. Thus, the platform sets the lowest value for the optimal moderation threshold, or $y^* = \sqrt{2v}$.

Case II. If $\hat{x}^S_{2,1}(\beta) < 1$, then $\hat{x}^S_{2,1}(\beta) \geq 0$ is equivalent to $y > \hat{y}^S(\beta)$ where $\hat{y}^S(\beta) = \sqrt{\hat{x}^S_{2,1}(\beta)^2 - \hat{x}^S_{2,2}(\beta)^2 + 2v}$ ($\hat{y}^S(\beta)$ is determined through solving for y from $\hat{x}^S_1(\beta, y)$). Any $y < \hat{y}^S(\beta)$ cannot be optimal because if $y < \hat{y}^S(\beta)$, then $\hat{x}^S_1(\beta, y) = 0$, and then the platform profit is $\zeta[\hat{x}^S_{2,1}(\beta) - \hat{x}^S_{2,1}(\beta) + y]$ which is increasing in y. Therefore we should consider $y > \hat{y}^S(\beta)$. Moreover, depending on whether $\hat{y}^S(\beta) < \hat{x}^S_{2,1}(\beta)$ or $\hat{y}^S(\beta) \geq \hat{x}^S_{2,1}(\beta)$ the platform's user base varies. We start our analysis with the condition $\hat{y}^S(\beta) < \hat{x}^S_{2,1}(\beta)$. If $\hat{x}^S_{2,1}(\beta) < y$, then $\partial \pi^R(y)/\partial y < 0$, therefore any $y > \hat{x}^S_{2,1}(\beta)$ cannot be optimal.

If $\hat{y}^S(\beta) < y \leq \hat{x}^S_{2,1}(\beta)$, then the platform user base is $\hat{x}^S_{2,1}(\beta) - \hat{x}^S_{2,1}(\beta) + y - \hat{x}^S_1(\beta, y)$, where $\hat{x}^S_1(\beta)$ is given by (B7). Therefore, from (B10) the platform profit is given by

$$\begin{aligned}\pi^S(y) = \zeta \Bigg[&\hat{x}^S_{2,2}(\beta) - \hat{x}^S_{2,1}(\beta) + y \\ &- \left[-\alpha - \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right] + \sqrt{\left[\alpha + \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right]\right]^2 - \hat{y}^S(\beta)^2 + y^2}\right]\Bigg],\end{aligned} \tag{B13}$$

and

$$FOC: \frac{\partial \pi^S(y)}{\partial y} = \zeta \left[1 - \frac{y}{\sqrt{\left[\alpha + \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right]\right]^2 - \hat{y}^S(\beta)^2 + y^2}}\right] = 0. \tag{B14}$$

There is no interior solution for y. However there exists β_0 such that $\partial \pi^S(y)/\partial y > 0$ when $\beta < \beta_0$ and $\partial \pi^S(y)/\partial y < 0$ when $\beta \geq \beta_0$. Because $y \in [\hat{y}^S(\beta), \hat{x}^S_{2,1}(\beta)]$, the optimal moderation threshold is $y^{S*} = \hat{x}^S_{2,1}(\beta)$ when $\beta < \beta_0$ and $y^{S*} = \hat{y}(\beta)$ when $\beta \geq \beta_0$. Notably, β_0 can be found by solving for β from $\left[\alpha + \frac{\beta}{2}\left[\hat{x}^S_{2,2}(\beta)^2 - \hat{x}^S_{2,1}(\beta)^2\right]\right]^2 - \hat{y}^S(\beta)^2 = 0$.

Case III. If $\hat{y}^S(\beta) > \hat{x}^S_{2,1}(\beta)$, then $y \in [\hat{x}^S_{2,1}(\beta), \hat{y}]$. Users in the range $[0, \hat{x}^S_{2,1}]$ participate and users in range $[\hat{x}^S_{2,2}, 1]$ participate if and only if $\beta \leq \bar{\beta}$. Therefore, if $\hat{y}^S(\beta) > \hat{x}^S_{2,1}(\beta)$ and $\beta \leq \bar{\beta}$, then the market is covered. Notably, if $\bar{\beta} < \beta$, then users in range $[\hat{x}^S_{2,2}(\beta), 1]$ do not participate and the market cannot be covered. As a tie-breaking rule, for further analysis, we assume the platform selects the lowest optimal moderation threshold, i.e., $y^{S*} = \hat{x}^S_{2,1}(\beta)$ (representing the most stringent moderation policy), to make the platform as moderated as possible. □

B.2.2. Proof of the Platform's Optimal Profit The platform's optimal moderation threshold is provided in Lemma B3. We can easily determine the maximum profit of the platform using B11, as follows:

Case I. If $\hat{x}^S_{2,1}(\beta) \geq 1$, then

a) if $\alpha \geq \sqrt{2v}$, then $y^{S*} = 1$, $X^{S*} = [\alpha + \sqrt{1+\alpha^2 - 2v}, 1]$, and $\pi_1^{S*} = \zeta[1 + \alpha - \sqrt{1+\alpha^2-2v}]$.

b) if $\alpha < \sqrt{2v}$, then $y^{S*} = \sqrt{2v}$, $X^{S*} = [0, \sqrt{2v}]$, and $\pi_2^{S*} = \zeta\sqrt{2v}$.

Case II. If $\hat{x}^S_{2,1}(\beta) < 1$, and $\hat{y}^S(\beta) < y < \hat{x}^S_{2,1}(\beta)$, then

a) if $\beta \geq \beta_0$, then $y^{S*} = \hat{x}^S_{2,1}(\beta)$, $X^{S*} = [\hat{x}^S_1(\beta, y^*), \hat{x}^S_{2,2}(\beta)]$, and $\pi_3^{S*} = \zeta[\hat{x}^S_{2,2}(\beta) - \hat{x}^S_1(\beta, y^*)]$.

b) if $\beta < \beta_0$, then $y^{S*} = \hat{y}^S(\beta)$, $X^{S*} = [0, y^{S*}] \cup [\hat{x}^S_{2,1}(\beta), \hat{x}^S_{2,2}(\beta)]$, and $\pi_4^{S*} = \zeta[\hat{x}^S_{2,2}(\beta) - \hat{x}^S_{2,1}(\beta) + \hat{y}^{S*}(\beta))]$.

Case III. $\hat{y}^S(\beta) \geq \hat{x}^S_{2,1}(\beta)$, then $y^{S*} \in [\hat{x}^S_{2,1}(\beta), \hat{y}^S(\beta)]$ and

a) if $\beta \leq \bar{\beta}$, then $X^{S*} = 1$, and $\pi_5^{S*} = \zeta$. b) if $\beta > \bar{\beta}$, then $X^{S*} = [0, \hat{x}^S_{2,2}(\beta)]$, and $\pi_6^{S*} = \zeta \hat{x}^S_{2,2}(\beta)$.

The maximum profit of the platform implementing shadowbanning is $\pi^{S*} \in \{\pi_1^{S*}, \pi_2^{S*}, \pi_3^{S*}, \pi_4^{S*}, \pi_5^{S*}, \pi_6^{S*}\}$. We should compare it with the maximum profit of the platform using content removal or no moderation. Because the maximum profit of the platform implementing shadowbanning in *Case I* is exactly the same as that using content removal, we only analyze the maximum profits determined in *Case II* and *III*. Therefore, we only compare $\pi_3^{S*}, \pi_4^{S*}, \pi_6^{S*}$ with $\pi^{R*} = \zeta\sqrt{2v}$ and $\pi^{N*} = \zeta[1 + \alpha - \sqrt{\alpha^2 + 1 - 2v}]$.

We start with π_6^{S*}. The condition $\hat{y}^S(\beta) \geq \hat{x}^S_{2,1}(\beta)$ is equivalent to $\hat{x}^S_{2,2}(\beta) \leq \sqrt{2v}$. Thus, $\pi_6^{S*} = \zeta\hat{x}^S_{2,2}(\beta) \leq \zeta\sqrt{2v} = \pi^{R*}$. Moreover, we know that if $\alpha \geq \sqrt{2v}$, then $\pi^{R*} < \pi^{N*}$. Thus, if $\alpha \geq \sqrt{2v}$, then $\pi_6^{S*} < \pi^{N*}$. Now, we compare π_4^{S*} and π^{R*}. However, we first show that $\hat{x}^S_{2,2}(\beta) - \hat{x}^S_{2,1}(\beta) < \sqrt{2v} - \hat{y}^S(\beta)$. We should mention that the condition $\hat{y}^S(\beta) < \hat{x}^S_{2,1}(\beta)$ is equivalent to $\hat{x}^S_{2,2}(\beta) < \sqrt{2v}$.

Additionally, in the proof of Lemma B3, we determined $\hat{y}^S(\beta) = \sqrt{x^S_{2,1}(\beta)^2 - x^S_{2,2}(\beta)^2 + 2v}$ which is equivalent to $\hat{y}^S(\beta)^2 = \hat{x}^S_{2,1}(\beta)^2 - \hat{x}^S_{2,2}(\beta)^2 + 2v$. Rearranging this expression, we have $2v - \hat{y}^S(\beta)^2 = \hat{x}^S_{2,1}(\beta)^2 - \hat{x}^S_{2,2}(\beta)^2$. This is equivalent to

$$[\sqrt{2v} - \hat{y}^S(\beta)][\sqrt{2v} + \hat{y}^S(\beta)] = [\hat{x}^S_{2,1}(\beta) - \hat{x}^S_{2,2}(\beta)][\hat{x}^S_{2,1}(\beta) + \hat{x}^S_{2,2}(\beta)] \tag{B15}$$

and because $\hat{y}^S(\beta) < \hat{x}^S_{2,1}(\beta)$ and $\hat{x}^S_{2,2}(\beta) < \sqrt{2v}$, then

$$\frac{\sqrt{2v} - \hat{y}^S(\beta)}{\hat{x}^S_{2,1}(\beta) - \hat{x}^S_{2,2}(\beta)} = \frac{\hat{x}^S_{2,1}(\beta) + \hat{x}^S_{2,2}(\beta)}{\sqrt{2v} + \hat{y}^S(\beta)} > 1. \tag{B16}$$

So, we conclude that $\hat{x}^S_{2,2}(\beta) - \hat{x}^S_{2,1}(\beta) < \sqrt{2v} - \hat{y}^S(\beta)$. Considering this and the condition $\hat{x}^S_{2,2}(\beta) < \sqrt{2v}$, then

$$\pi_4^{S*} = \hat{x}^S_{2,2}(\beta) - \hat{x}^S_{2,1}(\beta) + \hat{y}^S(\beta) < \sqrt{2v} - \hat{y}^S(\beta) + \hat{y}^S(\beta) < \sqrt{2v} = \pi^{R*}. \tag{B17}$$

However, if $\hat{x}_{2,2}^S(\beta) = 1$, then $\pi_4^{S*} > \pi^{R*}$ according to the numerical analysis.

Now, we need to compare $\pi_3^{S*} = \zeta[\hat{x}_{2,2}^S(\beta) - \hat{x}_1^S(\beta, y^*)]$ with π^{R*} and π^{N*}. We know than $\hat{x}_1^N = -\alpha + \sqrt{\alpha^2 + 1 - 2v}$ and if $\hat{x}_{2,2}^S(\beta) = 1$, then from (B7)

$$\hat{x}_1^{S*}(\beta, y) = -\alpha - \frac{\beta}{2}\left[1 - \hat{x}_{2,1}^S(\beta)^2\right] + \sqrt{\left[\alpha + \frac{\beta}{2}\left[1 - \hat{x}_{2,1}^S(\beta)^2\right]\right]^2 + 1 - 2v}. \tag{B18}$$

Thus, we can show that $\hat{x}_1^{S*}(\beta, y) < \hat{x}_1^N$ and then $1 - \hat{x}_1^{S*}(\beta, y) > 1 - \hat{x}_1^N$. And if $\hat{x}_{2,2}^S(\beta) = 1$, then $\pi_3^{S*} > \pi^{N*}$. For our remaining comparison, as well as the comparison of social welfare and consumer surplus under shadowbanning with heterogeneous user beliefs with that of no moderation and content removal, we use a numerical approach as explained in Section A.9. A summary of results is given in Table 3. □

Our results indicate that almost always a platform chooses to shadowban content except for two situations: 1) when users have a moderate heterogeneous belief about the platform shadowbanning implementation, with high posting utility and a high moderation cost for users, and 2) when heterogeneous user beliefs are sufficiently high, $\beta > max\{\hat{\beta}, \bar{\beta}\}$. This means that if β is low enough, then the platform shadowbans content, regardless of the cost of moderation for users, c. In addition, if $min\{\hat{\beta}, \bar{\beta}\} < \beta < max\{\hat{\beta}, \bar{\beta}\}$ and the cost of moderation for users is small enough, then the platform shadowbans content as well. However, if $min\{\hat{\beta}, \bar{\beta}\} < \beta < max\{\hat{\beta}, \bar{\beta}\}$ and the cost of moderation for users is sufficiently high, then the platform shadowbans content if and only if reading utility is sufficiently high, $\alpha < \sqrt{2v}$. To facilitate understanding of these complex conditions, Table 3 provides a summary of the results for shadowbanning with heterogeneous user beliefs.

Table 3 provides interesting insights. First, the platform is better off when the heterogeneous user beliefs about the platform shadowbanning implementation are low enough. In particular, when β is small enough ($\beta < max\{\bar{\beta}, \hat{\beta}\}$) the platform has higher profits with shadowbanning than with content removal or no moderation. Second, the results of shadowbanning with heterogeneous user beliefs are in accordance with the results of shadowbanning with common user beliefs, with slight variations explained in Section 6.

Appendix C: Shadowbanning with Imperfect Technology

Following the approach outlined by Liu et al. (2022), we consider two types of errors in detection of extreme content. First, the platform might fail to shadowban all the extreme content, leaving some of it visible. Second, the technology might unintentionally shadowban content that the platform intended to keep visible. To account for both errors, the content moderation technology can be described as a probability function, denoted by $q_k(x \mid y)$ which represents the likelihood that content with extremeness level x is shadowbanned, given the platform's goal of moderating all content above threshold y. We define the accuracy level of the technology by parameter k, where $k \in [0, 1/2]$. Specifically, we have

$$q_k(x \mid y) = \begin{cases} \frac{1}{2} - k, & \text{if } x \leq y, \\ \frac{1}{2} + k, & \text{if } x > y. \end{cases} \tag{C1}$$

The imperfect technology described in (C1) identifies extreme content (i.e., $x > y$) with a probability of $1/2 + k$, while it mistakenly identifies less extreme content ($x < y$) as extreme with a probability of $1/2 - k$. This means the platform is more likely to correctly shadowban extreme content than to erroneously shadowban less extreme content. When $k = 1/2$, the model aligns with a scenario of perfect accuracy, where extreme content is always detected. Conversely, when $k = 0$, the detection is completely random, with no distinction between extreme and less extreme content. Therefore, a higher value of k reflects more accurate moderation. The total utility $E(U_k^S(x))$ for a user with a given extremeness level x can be expressed as:

$$E(U_k^S(x)) = \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + \left[\alpha x[1 - q_k(x \mid y)] - c q_k(x \mid y) + v - \int_x^1 [1 - q_k(x \mid y)] z\,dz\right]\theta, \tag{C2}$$

which is equivalent to

$$E(U_k^S(x)) = \begin{cases} \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + \left[\alpha x[\frac{1}{2}+k] - c[\frac{1}{2}-k] + v - \int_x^y [\frac{1}{2}+k] z\,dz - \int_y^1 [\frac{1}{2}-k] z\,dz\right]\theta & if x \leq y \\ \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + \left[\alpha x[\frac{1}{2}-k] - c[\frac{1}{2}+k] + v - \int_x^1 [\frac{1}{2}-k] z\,dz\right]\theta & if x > y. \end{cases} \tag{C3}$$

We follow similar steps as in the proofs of Lemma 1 and 2, omitting the details here for brevity. There are two indifferent users, denoted by $\hat{x}_{2,k}^S(\theta)$ and $\hat{x}_{1,k}^S(\theta, y)$, which are determined by solving $E(U_k^S(x)) = 0$ for x when $x > y$ and $x < y$, respectively. Because $E(U_k^S(x))$ is increasing in x over both the intervals $[0, y]$ and $(y, 1]$, users participating on the platform belong to the ranges $[\hat{x}_{1,k}^S(\theta, y), y]$ and $[\hat{x}_{2,k}^S(\theta), 1]$, respectively.

Note that if $\hat{x}_{2,k}^S(\theta) \geq 1$, then no user with $x > y$ participates on the platform. This leads to the following conditions: $k \geq \bar{k} = [2\alpha - \alpha\theta - c\theta + 2v]/[2\alpha\theta + 2c\theta]$ and $\theta \geq \bar{\theta} = [\alpha + v]/[\alpha + c]$. In other words, if $k \geq \bar{k}$ and $\theta \geq \bar{\theta}$, then $\hat{x}_{2,k}^S(\theta) = 1$, meaning no user with $x > y$ participates. Otherwise, users in range $[\hat{x}_{2,k}^S(\theta), 1]$ participate. The expressions for $\hat{x}_{2,k}^S(\theta)$ and $\hat{x}_{1,k}^S(\theta, y)$ are given as:

$$\hat{x}_{2,k}^S(\theta) = \begin{cases} 1 & if \quad k \geq \bar{k} = \frac{2\alpha - \alpha\theta - c\theta + 2v}{2\alpha\theta + 2c\theta} \quad \& \quad \theta \geq \bar{\theta} = \frac{\alpha+v}{\alpha+c}, \\ -\alpha + \sqrt{1 + \alpha^2 + \frac{-2c\theta - 4ck\theta + 4v}{-2+\theta+2k\theta}} & if \quad otherwise. \\ \end{cases} \tag{C4}$$

$$\hat{x}^S_{1,k}(\theta,y)=\begin{cases} max\{0,-\alpha+\sqrt{\alpha^2+\frac{2-4v+\theta[-1+2c-2k-4ck+4ky^2]}{2-\theta+2k\theta}}\} & if \;\; y\geq \hat{x}^S_{2,k}(\theta) \\ max\{0,-\alpha+\sqrt{\alpha^2+\frac{2-4v+\theta[-1+2c-2k-4ck+[2k-1]\hat{x}^S_{2,k}(\theta)^2+[2k+1]y^2]}{2-\theta+2k\theta}}\} & if \;\; y< \hat{x}^S_{2,k}(\theta). \end{cases} \tag{C5}$$

The profit of the platform implementing shadowbanning with imperfect technology is given by

$$\pi^S_k(y)=\zeta\left[1-\hat{x}^S_{2,k}(\theta)+y-\hat{x}^S_{1,k}(\theta,y)\right]. \tag{C6}$$

To further analyze the platform's decision-making process, we now examine the optimal shadowbanning threshold in the presence of imperfect detection technology. The following lemma characterizes the platform's moderation strategy across different levels of technology accuracy and user belief:

Lemma C1. *The optimal moderation threshold for a platform implementing shadowbanning can be characterized as follows:*

Case I. If $k\geq\bar{k}$ and $\theta\geq\bar{\theta}$, then $\hat{x}^S_{2,k}(\theta)=1$ and a) if $k<k_1=\frac{\alpha^2\theta-2\alpha^2-2c\theta+4v}{2\theta[\alpha^2-2c]}$, then $y^{S*}_k(\theta)=1$; b) if $k\geq k_1$, then $y^{S*}_k(\theta)=\sqrt{\frac{2c\theta-4ck\theta-4v}{2-\theta+2k\theta}}$.

Case II. If $k<\bar{k}$ and $\theta\geq\bar{\theta}$, or if $\theta<\bar{\theta}$ (regardless of the value of k), then $\hat{x}^S_{2,k}(\theta)<1$ and

a) if $\hat{y}^S_k(\theta)<\hat{x}^S_{2,k}(\theta)$, and if $k<k_2$, then $y^{S*}_k(\theta)=\hat{x}^S_{2,k}(\theta)$, and if $k\geq k_2$, then $y^{S*}_k(\theta)=\hat{y}^S_k(\theta)$.[4]

b) if $\hat{y}^S_k(\theta)\geq\hat{x}^S_{2,k}(\theta)$, then $y^{S*}_k(\theta)\in\left[\hat{x}^S_{2,k}(\theta),\hat{y}^S_k(\theta)\right]$ and market is covered.

Where $\hat{y}^S_k(\theta)=\sqrt{\frac{4v-2+\theta[1+2k]+\theta[1-2k][-2c+\hat{x}^S_{2,k}(\theta)^2]}{\theta[1+2k]}}$ is determined by solving for y from $\hat{x}^S_{1,k}(\theta)=0$.

Considering Lemma C1, a platform does not implement shadowbanning in *Case I.a*, when $\bar{k}\leq k\leq k_1$, and $\theta\geq\bar{\theta}$ as the optimal shadowbanning threshold is 1. In all other cases, the platform implements shadowbanning, provided that the profit from shadowbanning exceeds the profit from no moderation.

C.1. Proof of Proposition 8

Based on *Case I* of Lemma C1, if user belief is high, $\theta\geq\bar{\theta}$, and technology is sufficiently accurate, $k\geq k_1$, then the platform may shadowban content. We now explore the platform's moderation decision under the conditions of *Case II* of Lemma C1. From (C6), the platform's profit is continuous in k. Thus, to demonstrate that the technology must be sufficiently accurate for the platform to implement shadowbanning, we need to show that there exists $\delta>0$ such that, when $k<\delta$, the profit

[4] There is a solution for $y^S_k(\cdot)$ when solving $\partial\pi^{S*}_k(\theta)/\partial y=0$. However, because the second derivative of profit with respect to y is positive, as a result, the optimal shadowbanning threshold is at one of the corner solutions: either $\hat{x}^S_{2,k}(\theta)$ or $\hat{y}^S_k(\theta)$. We use Mathematica's `FindInstance` function to validate this claim. Our findings show that, considering the the constraints in Lemma C1, there are no parameter values for α, v, c, θ, y, and k where $\partial^2\pi^S_k(\theta)/\partial y^2$ is negative.
In addition, it can be shown that there exists a threshold, denoted by k_2, such that $\partial\pi^{S*}_k(\theta)/\partial y$ is negative for $k\geq k_2$, and positive for $k<k_2$. However, we do not derive an explicit expression for k_2, as our results do not depend on its exact value.

under shadowbanning is less than the profit under no content moderation. Therefore, it suffices to show that the platform's shadowbanning profit with the random technology ($k=0$) is lower than that of no moderation. When $k=0$, the expected utility of the most extreme user, $x=1$, is as follows:

$$E(U_k^S(1)) = \alpha - \frac{1}{2}[\alpha + c]\theta + v, \tag{C7}$$

which is positive, $E(U_k^S(1)) > 0$, when $\theta < \bar{\theta}$. Additionally, from (C3), we derive the indifferent user as

$$\hat{x}_{1,k}^S(\theta, y) = -\alpha + \sqrt{\alpha^2 + 1 + \frac{4v - 2c\theta}{\theta - 2}}. \tag{C8}$$

The maximum profit under shadowbanning when $k=0$ is given by $1 - \hat{x}_{1,k}^S(\theta,y) = 1 + \alpha - \sqrt{\alpha^2 + 1 + \frac{4v-2c\theta}{\theta-2}}$. Comparing this with the profit under no moderation, $1 + \alpha - \sqrt{\alpha^2 + 1 - 2v}$, we find that the shadowbanning profit when $k=0$ is always lower than the profit from no moderation.[5] This implies that, for the platform to shadowban content with imperfect technology, the technology must be sufficiently accurate. This completes the proof for Proposition 8. □

Better Technology and Less Content Moderation (Section 7.2). According to Lemma C1 the optimal shadowbanning threshold is either $\sqrt{\frac{2c\theta-4ck\theta-4v}{2-\theta+2k\theta}}$, $\hat{x}_{2,k}^S(\theta)$, or $\hat{y}_k^S(\theta)$. We can easily show that the derivative of $\sqrt{\frac{2c\theta-4ck\theta-4v}{2-\theta+2k\theta}}$ and $\hat{x}_{2,k}^S(\theta)$ with respect to k are positive. To determine the sign of $\partial\hat{y}^S k(\theta)/\partial k$, we used Mathematica's `FindInstance` function to explore the parameter space. Our results indicate that $\partial\hat{y}_k^S(\theta)/\partial k$ is positive for all values of the parameters α, v, c, k, and θ that satisfy the conditions of Lemma C1.

We also investigate the impact of a higher user belief (larger θ) on the optimal shadowbanning threshold. In the base model, a larger θ leads to a higher optimal shadowbanning threshold, as indicated by the positive comparative statics: $\partial\hat{y}^{S*}(\theta)/\partial\theta > 0$. Similarly, the derivative of the optimal shadowbanning threshold with imperfect technology with respect to θ is positive in all cases of Lemma C1. Therefore, when the technology is imperfect, the platforms implement a less strict shadowbanning as the user belief increases.

[5] The following code in Mathematica returns False.

$$\text{Reduce}\left[\left\{1+\alpha-\sqrt{\alpha^2+1-2v} < 1+\alpha-\sqrt{\alpha^2+1+\frac{4v-2c\theta}{\theta-2}}, \alpha>0, c>v, 1>\theta>0, v>0\right\}, \{v\}\right]. \tag{C9}$$

Incentive for Imperfect Content Moderation Technology (Section 7.3). Similar to the approach of Liu et al. (2022), we use numerical analysis to demonstrate these results. Our findings indicate that if the cost of moderation for users, c, is small, then the platform always chooses perfect technology. However, if the cost of moderation for users is large, then the platform has an incentive to maintain imperfect technology.

Appendix D: Shadowbanning with Mandated Content Removal Policy (SR)

D.1. Proof of Proposition 9

We examine a platform's content moderation decision that employs both content removal and shadowbanning, defined by two distinct thresholds: y_r for mandated content removal and y_s for shadowbanning. Both thresholds are bounded within [0,1], with $y_s \le y_r$. The platform removes content when its degree of extremeness x exceeds y_r and shadowbans content when its extremeness falls within the range $y_s \le x < y_r$. Although y_r is public due to regulatory policy, users remain unaware of the shadowbanning threshold y_s. All other assumptions from our main model remain unchanged.

Using the superscript SR to denote this setting, the total utility $E(U^{SR}(x))$ for a user with a given extremeness level x can be expressed as:

$$E(U^{SR}(x)) = \begin{cases} \left[\alpha x + v - \int_x^{y_r} z\,dz\right][1-\theta] + \left[\alpha x + v - \int_x^{y_s} z\,dz\right]\theta & if x \le y_s, \\ \left[\alpha x + v - \int_x^{y_r} z\,dz\right][1-\theta] + [-c+v]\,\theta & if y_s \le x \le y_r, \\ -c+v & if y_r \le x. \end{cases} \tag{D1}$$

Following the analytical approach used in Lemmas 1 and 2, we identify two indifferent users, represented by $\hat{x}_2^{SR}(\theta)$ and $\hat{x}_1^{SR}(\theta, y)$. These values are determined by solving $E(U^{SR}(x)) = 0$ for x in the ranges $y_s < x \le y_r$ and $x \le y_s$, respectively. Given that $E(U^{SR}(x))$ increases with x in both intervals $[0, y_s]$ and $(y_s, 1]$, platform users fall within the ranges $[\hat{x}_1^{SR}(\theta, y), y_s]$ and $[\hat{x}_2^{SR}(\theta), 1]$. For some users with $x > y$ to participate, the condition $\hat{x}_2^{SR}(\theta) < 1$ must hold. This leads to the following condition: $y_r > [v - c\theta]/\alpha[1-\theta]$. In other words, if $y_r \le [v - c\theta]/\alpha[1-\theta]$, then $\hat{x}_2^{SR}(\theta) = 1$, meaning no user with $x > y$ participates on the platform. Otherwise, users in the range $[\hat{x}_2^{SR}(\theta), 1]$ participate. The expressions for the threshold users are:

$$\hat{x}_2^{SR}(\theta) = \begin{cases} 1 & if \quad y_r \le \frac{v-c\theta}{\alpha[1-\theta]}, \\ -\alpha + \sqrt{\alpha^2 + y_r^2 + \frac{2c\theta - 2v}{1-\theta}} & if \quad y_r > \frac{v-c\theta}{\alpha[1-\theta]}. \end{cases} \tag{D2}$$

$$\hat{x}_1^{SR}(\theta, y) = \begin{cases} min\{y, -\alpha + \sqrt{\alpha^2 - 2v + \theta y_s^2 + [1-\theta] y_r^2}\} & if \quad y_s > \hat{x}_2^{SR}(\theta) \\ max\{0, -\alpha + \sqrt{\alpha^2 - 2v + y_s^2 + [1-\theta][y_r^2 - \hat{x}_2^{SR}(\theta)^2]}\} & if \quad y_s \le \hat{x}_2^{SR}(\theta). \end{cases} \tag{D3}$$

The platform profit is expressed as:

$$\pi^{SR}(y) = \zeta\left[y_r - \hat{x}_2^{SR}(\theta) + y_s - \hat{x}_1^{SR}(\theta, y)\right]. \tag{D4}$$

The following lemma characterizes the platform's shadowbanning implementation decision and threshold.

Lemma D1. *The optimal moderation threshold for a platform implementing both shadowbanning and content removal can be characterized as follows:*

Case I. If $y_r < \frac{v-c\theta}{\alpha[1-\theta]}$*, then* $\hat{x}_2^{SR}(\theta) = y_r$ *and a) if* $\alpha \geq \sqrt{2v}$*, then* $y_s^{SR*}(\theta) = y_r$*; b) if* $\alpha < \sqrt{2v}$*, then* $y_s^{SR*}(\theta) = \sqrt{2v}$*.*

Case II. If $y_r \geq \frac{v-c\theta}{\alpha[1-\theta]}$ *and* $\theta < \theta_0^{SR}$*, then* $y_s^{SR*}(\theta) = \hat{x}_2^{SR}(\theta)$*.*

Case III. If $y_r \geq \frac{v-c\theta}{\alpha[1-\theta]}$ *and* $\theta \geq \theta_0^{SR}$*, then* $y_s^{SR*}(\theta) = \hat{y}^{SR}(\theta)$*.*

where $\hat{y}^{SR}(\theta) = \sqrt{2v - [1-\theta][y_r^2 - \hat{x}_2^{SR}(\theta)^2]}$ *is derived by solving for* y_s *from* $\hat{x}_1^{SR}(\theta, y) = 0$*. Solving for* θ *from* $\partial\pi(y)/\partial y = 0$*,* θ_0^{SR} *is given.*

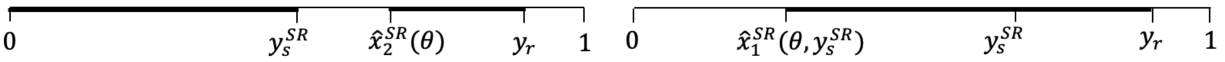
Based on Lemma D1, the platform refrains from shadowbanning when $y_r \leq [v - c\theta]/\alpha[1-\theta]$ and $\alpha \geq \sqrt{2v}$, as the optimal shadowbanning threshold equals y_r. In all other cases, shadowbanning is implemented. Because the profit from shadowbanning exceeds the profit from no moderation. This is verified through a numerical approach, similar to that in Section A.9, which compares the platform's profit with shadowbanning to its profit with no moderation. This completes the proof of Proposition 9. □

Lemma D2. *We have:* $\partial\hat{y}^{SR}(\theta)/\partial y_r < 0$*,* $\partial\hat{x}_2^{SR}(\theta)/\partial y_r > 0$*,* $\partial\hat{x}_1^{SR}(\theta, y)/\partial y_r > 0$*,* $\partial\hat{y}^{SR}(\theta)/\partial\theta > 0$*,* $\partial\hat{x}_2^{SR}(\theta)/\partial\theta > 0$*,* $\partial\hat{x}_1^{SR}(\theta, y)/\partial\theta < 0$*.*

The proof is straightforward and is omitted for brevity. □

D.2. The Platform's Optimal Shadowbanning Threshold and User Segmentation in the Equilibrium (Section 8.2)

We now examine how changes in y_r affect the platform's optimal shadowbanning threshold and user segmentation. From Lemma D2, we know that $\partial\hat{y}^{SR}(\theta)/\partial y_r < 0$, $\partial\hat{x}_2^{SR}(\theta)/\partial y_r > 0$, and



(a) $\theta \geq \theta_0^{SR}$; $y_s^{SR*}(\theta) = \hat{y}^{SR}(\theta)$.

(b) $\theta < \theta_0^{SR}$; $y_s^{SR*}(\theta) = \hat{x}_2^{SR}(\theta)$.

Figure D1 **User Segmentation in Equilibrium - Simultaneous Implementation of Shadowbanning and Content Removal (SR)**

$\partial \hat{x}_1^{SR}(\theta, y)/\partial y_r > 0$. Also, the equilibrium user segmentation is illustrated in Figure D1 using Lemma D1.

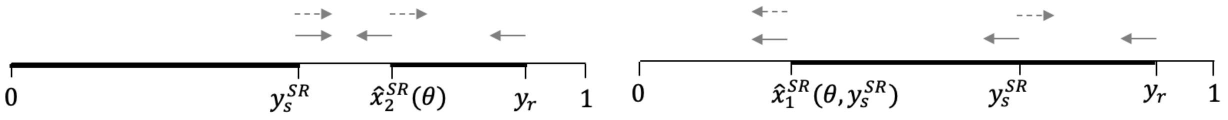
If user belief regarding shadowbanning implementation is low ($\theta < \theta_0^{SR}$), then $\hat{y}^{SR}(\theta) = \hat{x}_2^{SR}(\theta)$. In this case, the user segmentation is continuous, as shown in Figure D1b. An increase in y_r shifts the user segmentation to the right, while a decrease shifts it to the left. This implies that if policymakers adopt a more lax content removal policy, then the platform responds with more lax shadowbanning. Conversely, if policymakers implement a stricter content removal policy, then the platform adopts stricter shadowbanning.

If user belief regarding shadowbanning implementation is high, $\theta \geq \theta_0^{SR}$, then $y_s^{SR*}(\theta) = \hat{y}^{SR}(\theta)$. In this case, the user segmentation is disjoint, as shown in Figure D1a. The impact of changes in y_r on $y_s^{SR*}(\theta)$ reveals interesting findings. An increase in y_r leads to an increase in $\hat{x}_2^{SR}(\theta)$ and a decrease in $y_s^{SR*}(\theta)$. If policymakers adopt a more lax content removal policy (decrease in y_r), then fewer moderate users participate on the platform. In other words, under this lax content removal policy, the platform opts for stricter shadowbanning, which results in reduced participation of moderate users. Specifically, the platform shadowbans both moderate users and extreme users (whose content is not otherwise subject to removal) in favor of retaining the least extreme users.

Conversely, if policymakers adopt a stricter content removal policy (larger y_r), then moderate user participation increases. This is because the platform adopts more lax shadowbanning to attract moderate users, thereby compensating for the user base reduction caused by stricter content removal policies.

We also find that, given a specific removal threshold set by policymakers, if user belief increases (larger θ), then the platform chooses a higher shadowbanning threshold (implementing more lax shadowbanning), which results in a reduced volume of shadowbanned content. This result is in line with the findings of our main model. Additionally, the impact of the increase in user belief on user participation depends on their prior beliefs: if the initial user belief about shadowbanning is low, then an increase in user belief leads to participation primarily from the least extreme users; however, if the initial belief is high, then the platform's more lax shadowbanning attracts moderate users. Moreover, a substantial increase in user belief can transform user segmentation from continuous to disjoint, leading to moderate users opting out of platform participation (as illustrated in the transition from Figure D2b to Figure D2a).

In addition, we find that the joint impact of higher user belief (larger θ) and a stricter content removal policy (smaller y_r) on the shadowbanning threshold depends on the initial user beliefs. When the initial belief is high, they reinforce each other, leading the platform to choose more lax shadowbanning. However, when the initial belief is low, these factors have opposing effects, and the net impact on the shadowbanning threshold is unclear. Furthermore, although the net



(a) High initial belief; $y_s^*(\theta) = \hat{y}^{SR}(\theta)$.

(b) Low initial belief; $y_s^*(\theta) = \hat{x}_2^{SR}(\theta)$.

Dotted arrow indicates the effect of an increase in θ (higher user belief about shadowbanning).
Solid arrow indicates the effect of a decrease in y_r (stricter removal policy).

Figure D2 User Segmentation in Equilibrium (SR)

effects of these combined forces on the platform's total user base is ambiguous, their impact on user composition is clear and depends on the initial user belief: if the initial belief is high, then the combined forces encourage moderate users to participate; if the initial belief is low, then they encourage less extreme users.

D.3. Discussion on the assumption of the removal threshold defined by policymakers

Content moderation regulations worldwide have evolved to define specific categories of illegal content that must be removed. These typically include child sexual abuse material, terrorist content, hate speech inciting violence, non-consensual intimate imagery, and content facilitating serious criminal activities like human trafficking or illegal drug sales. Over time, these definitions have become more precise and comprehensive.

For instance, in the European Union, the Digital Services Act (DSA), defines illegal content as any information that is not compliant with EU law or the law of a Member State. This includes terrorist content, child sexual abuse material, illegal hate speech, commercial scams and fraud, or breaches of intellectual property rights (European Commission 2018). The DSA mandates that online platforms swiftly remove such illegal content. Although DSA forces platforms to remove specific types of content, it does not provide a list of content that could be moderated using other moderation strategies (e.g., shadowbanning). As you noted, it requires platforms to conduct risk assessments and implement measures to mitigate systemic risks (European Commission (2022), Articles 34 and 35).

Similarly, Germany's Network Enforcement Act (NetzDG) requires social media platforms to remove "clearly illegal" content within 24 hours. For cases where the illegality is not obvious, companies have up to seven days to investigate and decide on moderation (Wikipedia 2024, De Streel et al. 2020). The law itself does not provide a specific statutory definition of "clearly illegal" content. Instead, it references a list of offenses already defined in the German Criminal Code-such as incitement to hatred, dissemination of symbols of unconstitutional organizations, defamation, threats of violence, and others (Human Rights Watch 2018, De Streel et al. 2020).

Another example is Canada where the government has identified five categories of illegal content that must be moderated: child sexual exploitation content, terrorist content, content that incites violence, hate speech, and non-consensual sharing of intimate images (Benmoussa et al. 2024). In 2024, this list was expanded to include two additional categories: child bullying, and inducement of self-harm in children (Canadian Heritage 2024, Langevin et al. 2024).

These regulatory frameworks are moderation thresholds defined by policymakers, where platforms are required to remove content that exceeds a certain level of harm. For content falling below this threshold, platforms decide whether and how to moderate the content.

To illustrate, we refer to Meta's regulations and policies. As stated on their transparency website,[6] Meta takes a three-part approach to content enforcement on Facebook and Instagram: *remove, reduce, and inform.* In addition to removing content that violates their policies, they "often reduce the distribution of [problematic] content, even when it doesn't quite meet the standard for removal under [their] policies." They provide a list of problematic content categories,[7] and for each category, they apply the same three-part approach, although details about the "reduce" action are limited. For example, in the category of Fraud, Scams, and Deceptive Practices,[8] Meta removes content related to loan, gambling, and investment fraud/scams, as well as misleading health practices, while it is explicitly stated that they reduce the visibility of content that "promises specific weight-loss results in specific time with no qualifying or disclaimer language."

Appendix E: Shadowbanning with Updated Beliefs

We examine a two-period game where users that were shadowbanned in the first period adjust their belief about being shadowbanned in the second period. For users that were not shadowbanned in the first period, their belief remains unchanged at θ. However, users that were shadowbanned increase their belief by a small amount $\epsilon \in [0,1)$. This adjustment means their belief in the second period becomes $\theta+\epsilon<1$. We explore how this change in belief influences user segmentation and the platform's decisions.

We follow similar steps as in the proofs of Lemma 1, omitting the details here for brevity. For the first period, there are two indifferent users, denoted by $\hat{x}^S_{2,d1}(\theta)$ and $\hat{x}^S_{1,d1}(\theta,y)$, determined by solving $E(U^S_d(x))=0$ for $x>y$ and $x<y$, respectively. Similarly, for the second period, the two indifferent users are denoted by $\hat{x}^S_{2,d2}(\theta,\epsilon)$ and $\hat{x}^S_{1,d2}(\theta,\epsilon,y)$. Because $E(U^S_d(x))$ is increasing in x

[6] `https://transparency.meta.com/enforcement/taking-action/`

[7] `https://transparency.meta.com/policies/community-standards/`

[8] `https://transparency.meta.com/policies/community-standards/fraud-scams`

over both the intervals $[0, y]$ and $(y, 1]$ in both periods, users participating on the platform belong to the ranges $[\hat{x}^S_{1,d1}(\theta, y), y]$ and $[\hat{x}^S_{2,d1}(\theta), 1]$ for the first period, and $[\hat{x}^S_{1,d2}(\theta, \epsilon, y), y]$ and $[\hat{x}^S_{2,d2}(\theta, \epsilon), 1]$ for the second period.

Solving for θ from $E(U^S_{d1}(x)) = 0$ when $x = 1$ gives $\bar{\theta}$. Because $E(U^S_{d1}(x))$ increases in x, the value of $E(U^S_{d1}(1))$ is negative when $\theta \geq \bar{\theta}$, implying that all users in the range $(y, 1]$ have negative expected utilities. Consequently, if $\theta \geq \bar{\theta}$, then no users in the range $(y, 1]$ participate on the platform, leading to $\hat{x}^S_{2,d1}(\theta) = 1$. Similarly, $\bar{\theta}_d$ is determined by solving for θ from $E(U^S_{d2}(x)) = 0$ when $x = 1$.

The expressions for $\hat{x}^S_{2,d1}(\theta)$, $\hat{x}^S_{1,d1}(\theta, y)$, $\hat{x}^S_{2,d2}(\theta, \epsilon)$, and $\hat{x}^S_{1,d2}(\theta, \epsilon, y)$ are given as:

$$\hat{x}^S_{2,d1}(\theta) = \begin{cases} -\alpha + \sqrt{\alpha^2 + 1 - 2v + \frac{2\theta[c-v]}{1-\theta}} & if \quad \theta < \frac{v+\alpha}{c+\alpha} = \bar{\theta} \\ 1 & if \quad \theta \geq \frac{v+\alpha}{c+\alpha} = \bar{\theta}, \end{cases} \tag{E1}$$

$$\hat{x}^S_{1,d1}(\theta, y) = \begin{cases} min\{y, -\alpha + \sqrt{\alpha^2 + 1 - 2v + \theta\,[y^2 - 1]}\} & if \quad y \geq \hat{x}^S_{2,d1}(\theta) \\ max\{0, -\alpha + \sqrt{\alpha^2 + 1 - 2v - \hat{x}^S_{2,d1}(\theta)^2 + \theta\left[\hat{x}^S_{2,d1}(\theta)^2 - 1\right] + y^2}\} & if \quad y < \hat{x}^S_{2,d1}(\theta), \end{cases} \tag{E2}$$

$$\hat{x}^S_{2,d2}(\theta, \epsilon) = \begin{cases} -\alpha + \sqrt{\alpha^2 + 1 - 2v + \frac{2[\theta c + \theta\epsilon - v]}{1-\theta-\epsilon}} & if \quad \theta < \frac{1-\epsilon+2c\epsilon-2v\alpha}{1-2c} = \bar{\theta}_d \\ 1 & if \quad \theta \geq \frac{1-\epsilon+2c\epsilon-2v\alpha}{1-2c} = \bar{\theta}_d, \end{cases} \tag{E3}$$

$$\hat{x}^S_{1,d2}(\theta, \epsilon, y) = \begin{cases} min\{y, -\alpha + \sqrt{\alpha^2 + 1 - 2v + \theta\,[y^2 - 1]}\} & if \quad y \geq \hat{x}^S_{2,d2}(\theta, \epsilon) \\ max\{0, -\alpha + \sqrt{\alpha^2 + 1 - 2v - \hat{x}^S_{2,d2}(\theta, \epsilon)^2 + \theta\left[\hat{x}^S_{2,d2}(\theta, \epsilon)^2 - 1\right] + y^2}\} & if \quad y < \hat{x}^S_{2,d2}(\theta, \epsilon). \end{cases} \tag{E4}$$

We note that the indifferent users in the first period, $\hat{x}^S_{2,d1}(\theta)$ and $\hat{x}^S_{1,d1}(\theta, y)$, are identical to those in our static model (refer to Lemma 2 in Section 5.1). However, this equivalence does not hold for the second period. Additionally, from the above expressions, we observe that user participation in our dynamic model closely resembles that in our static model, as illustrated in Figure 2 of the main text.

We assume that $\bar{\theta}$ is greater than $\bar{\theta}_d$ in (E1) and (E3), as the condition $\bar{\theta} < \bar{\theta}_d$ contradicts the logic of our setting. Specifically, if $\bar{\theta} < \theta < \bar{\theta}_d$, then certain extreme users would refrain from participating on the platform in the first period but subsequently update their beliefs in the second period. This scenario is inconsistent with the assumptions and rationale underlying our model. Additionally, we assume that ϵ is sufficiently small ($\epsilon < 1 - \theta$) to ensure that $\hat{x}^S_{2,d2}(\theta, \epsilon)$ is greater than or equal to $\hat{x}^S_{2,d1}(\theta)$.

Lemma E1. *The optimal moderation threshold for a platform implementing shadowbanning can be characterized by the following:*

Case I. If $\bar{\theta}_d \le \theta$, then $\hat{x}^S_{2,d1}(\theta) \le 1$, and $\hat{x}^S_{2,d2}(\theta) = 1$ and a) if $\alpha \ge \hat{y}^S_{d1}(\theta)$, then $y^{S*}_d(\theta,\epsilon) = 1$; b) if $\alpha < \hat{y}^S_{d1}(\theta)$, then $y^{S*}_d(\theta,\epsilon) = \sqrt{2v}$.

Case II. If $\theta < \bar{\theta}_d$, then $\hat{x}^S_{2,d1}(\theta) < 1$ and $\hat{x}^S_{2,d2}(\theta) < 1$. If $\hat{y}_2(\theta,\epsilon) < \hat{x}^S_{2,d2}(\theta,\epsilon)$ and a) if $\alpha \ge \hat{y}^S_{d2}(\theta,\epsilon)$ then $y^{S*}_d(\theta,\epsilon) = \hat{x}^S_{2,d2}(\theta,\epsilon)$; b) if $\alpha < \hat{y}^S_{d2}(\theta,\epsilon)$ then $y^{S*}_d(\theta,\epsilon) = \hat{y}^S_{d2}(\theta,\epsilon)$.

Case III. If $\theta < \bar{\theta}_d$ and $\hat{y}_2(\theta,\epsilon) \ge \hat{x}^S_{2,d2}(\theta,\epsilon)$, then $y^{S*} \in [\hat{x}^S_{2,d2}(\theta,\epsilon), \hat{y}^S_{d2}(\theta,\epsilon)]$ and the market is covered in both periods.

Where $\hat{y}^S_{d1}(\theta) = \sqrt{2v + [1-\hat{x}^S_{2,d1}(\theta)^2][1-\theta]}$ and $\hat{y}^S_{d2}(\theta,\epsilon) = \sqrt{2v - [1-\hat{x}^S_{2,d2}(\theta,\epsilon)^2][1-\theta]}$, and are determined through solving for y from $\hat{x}^S_{1,d1}(\theta,y) = 0$ and $\hat{x}^S_{1,d2}(\theta,\epsilon,y) = 0$, respectively.

E.1. Proof of Lemma E1:

The profit of the platform implementing shadowbanning in our dynamic model is the sum of the profits of two periods and is given by

$$\pi^S_d(y) = \zeta\left[\underbrace{1-\hat{x}^S_{2,d2}(\theta,\epsilon)+y-\hat{x}^S_{1,d2}(\theta,\epsilon,y)}_{\text{Period 2's profit}}+\underbrace{1-\hat{x}^S_{2,d1}(\theta)+y-\hat{x}^S_{1,d1}(\theta,y)}_{\text{Period 1's profit}}\right]. \tag{E5}$$

We derive the optimal moderation threshold, and consequently, determine the platform's maximum profit.

Case I. We begin with the case when $\theta > \bar{\theta}_d$ and according to (E3), $\hat{x}^S_{2,d2}(\theta,\epsilon) = 1$. Notably, because we consider that $\bar{\theta} > \bar{\theta}_d$, therefore if $\theta > \bar{\theta}$, then $\hat{x}^S_{2,d1}(\theta) = 1$, and if $\theta < \bar{\theta}$, then $\hat{x}^S_{2,d1}(\theta) < 1$. Therefore, if $\theta > \bar{\theta}_d$, then $\hat{x}^S_{2,d1}(\theta) \le 1$.

From (E4), $\hat{x}^S_{1,d2}(\theta,\epsilon,y) = -\alpha + \sqrt{\alpha^2 - 2v + y^2}$. Because $\hat{x}^S_{1,d2}(\theta,\epsilon,y) \ge 0$, then $y \in [\sqrt{2v}, 1]$. The user base of the platform in period 1 and 2 are $[\hat{x}^S_{1,d1}(\theta,y), y] \cup [\hat{x}^S_{2,d1}(\theta), 1]$ and $[\hat{x}^S_{1,d2}(\theta,\epsilon,y), y]$, respectively. Thus the profit is given by $\pi(y) = \zeta\left[1-\hat{x}^S_{2,d1}(\theta)+y-\hat{x}^S_{1,d1}(\theta,y)+y-\hat{x}^S_{1,d2}(\theta,\epsilon,y)\right]$, and

$$FOC: \frac{\partial \pi(y)}{\partial y} = \zeta\left[2 - \frac{y}{\sqrt{\alpha^2 - 2v + [1-\hat{x}^S_{2,d1}(\theta)^2][1-\theta] + y^2}} - \frac{y}{\sqrt{\alpha^2 - 2v + y^2}}\right] = 0. \tag{E6}$$

We substitute $2v - [1-\hat{x}^S_{2,d1}(\theta)^2][1-\theta]$ with $\hat{y}^S_{d1}(\theta)$ in (E6) ($\hat{y}^S_{d1}(\theta)$ is determined by solving for y from $\hat{x}^S_{1,1}(\theta,y) = 0$). This substitution yields the following first-order condition:

$$FOC: \frac{\partial \pi(y)}{\partial y} = \zeta\left[1 - \frac{y}{\sqrt{\alpha^2 - \hat{y}^S_{d1}(\theta)^2 + y^2}} + 1 - \frac{y}{\sqrt{\alpha^2 - 2v + y^2}}\right] = 0. \tag{E7}$$

There is no interior solution for y. We know that $\hat{y}^S_{d1}(\theta) \le \sqrt{2v}$. Thus from (E7) for $\alpha < \hat{y}^S_{d1}(\theta)$, $\partial \pi(y)/\partial y$ is negative because the values of $y/\sqrt{\alpha^2 - \hat{y}^S_{d1}(\theta)^2 + y^2}$ and $y/\sqrt{\alpha^2 - 2v + y^2}$ are both greater than 1. Similarly, if $\alpha \geq \sqrt{2v}$, then $\partial\pi(y)/\partial y > 0$. If $\hat{y}^S_{d1}(\theta) < \alpha < \sqrt{2v}$, then using Mathematica's FindInstance function, we could not find any combination of parameter values where $\partial\pi(y)/\partial y < 0$. Therefore, if $\hat{y}^S_{d1}(\theta) < \alpha < \sqrt{2v}$, then $\partial\pi(y)/\partial y$ is always positive. Consequently, if $\alpha \geq \hat{y}^S_{d1}(\theta)$, then the optimal moderation threshold is $y^{S*}_d(\theta,\epsilon) = 1$, and if $\alpha < \hat{y}^S_{d1}(\theta)$, then $y^{S*}_d(\theta,\epsilon) = \sqrt{2v}$.

Case II. If $y > \hat{x}^S_{2,d2}(\theta,\epsilon)$, then the platform's user base is $[\hat{x}^S_{1,d1}(\theta,y),1]$ in the first period and $[\hat{x}^S_{1,d2}(\theta,\epsilon,y),1]$ in the second period. Thus, the platform profit is given by $\pi = \zeta\left[1-\hat{x}^S_{1,d1}(\theta,y)+1-\hat{x}^S_{1,d2}(\theta,\epsilon,y)\right]$. From (E2) and (E4), $\hat{x}^S_{1,d1}(\theta,y)$ and $\hat{x}^S_{1,d2}(\theta,\epsilon,y)$ are increasing in y. Thus, $\partial\pi(y)/\partial y < 0$ and any $y > \hat{x}^S_{2,d2}(\theta,\epsilon)$ is not optimal. If $y \leq \hat{x}^S_{2,d2}(\theta,\epsilon)$, then note that $\hat{x}^S_{1,2}(\theta,\epsilon,y) \geq 0$ is equivalent to $y \geq \hat{y}^S_{d2}(\theta)$ where $\hat{y}^S_{d2}(\theta,\epsilon) = \sqrt{2v-[1-\hat{x}^S_{2,d2}(\theta,\epsilon)^2][1-\theta]}$ (determined by solving for y from $\hat{x}^S_{1,2}(\theta,\epsilon,y) = 0$). Any $y < \hat{y}^S_{d2}(\theta,\epsilon)$ cannot be an optimal choice because if $y < \hat{y}^S_{d2}(\theta,\epsilon)$, then $\hat{x}^S_{2,1}(\theta,\epsilon,y) = 0$, so the platform profit is $\pi(y) = \zeta\left[1-\hat{x}^S_{2,d2}(\theta,\epsilon)+y-0+1-\hat{x}^S_{2,d1}(\theta)+y-\hat{x}^S_{1,d1}(\theta,y)\right]$. Because $\hat{x}^S_{2,d2}(\theta,\epsilon) > \hat{x}^S_{2,d1}(\theta)$, then $\hat{y}^S_{d1}(\theta) < \hat{y}^S_{d2}(\theta,\epsilon)$. As a result $\hat{x}^S_{1,d1}(\theta,y) \geq 0$.

The value of $\partial\pi(y)/\partial y = 2 - y/\sqrt{\alpha^2-\hat{y}^S_{d1}(\theta)^2+y^2}$ is always positive (by using Mathematica's FindInstance function in a similar approach as in Case I). Thus, any $y < \hat{y}^S_{d2}(\theta,\epsilon)$ cannot be optimal. Therefore, we only consider $y \in [\hat{y}^S_{d2}(\theta,\epsilon), \hat{x}^S_{2,d2}(\theta,\epsilon)]$. In this case, the platform profit is $\pi(y) = \zeta\left[1-\hat{x}^S_{2,d2}(\theta,\epsilon)+y-\hat{x}^S_{1,d2}(\theta,\epsilon,y)+1-\hat{x}^S_{2,d1}(\theta)+y-\hat{x}^S_{1,d1}(\theta,y)\right]$, and we have

$$FOC: \frac{\partial\pi(y)}{\partial y} = \zeta\left[1-\frac{y}{\sqrt{\alpha^2-\hat{y}^S_{d2}(\theta,\epsilon)^2+y^2}}+1-\frac{y}{\sqrt{\alpha^2-\hat{y}^S_{d1}(\theta)^2+y^2}}\right] = 0. \qquad \text{(E8)}$$

There is no interior solution for y. Considering the fact that $\hat{y}^S_{d1}(\theta) < \hat{y}^S_{d2}(\theta,\epsilon)$ – because $\hat{x}^S_{2,d2}(\theta,\epsilon) > \hat{x}^S_{2,d1}(\theta)$, with a similar approach as in Case I, we have the following condition: if $\alpha > \hat{y}^S_{d2}(\theta,\epsilon)$, then $\partial\pi(y)/\partial y$ is positive and then $y^{S*}_d(\theta,\epsilon) = \hat{x}^S_{2,d2}(\theta,\epsilon)$, and if $\alpha < \hat{y}^S_{d2}(\theta,\epsilon)$, then $\partial\pi(y)/\partial y$ is negative and then $y^{S*}_d(\theta,\epsilon) = \hat{y}^S_{d2}(\theta,\epsilon)$.

So far we considered that y is less than $\hat{x}^S_{2,d1}(\theta)$. If $y > \hat{x}^S_{2,d1}(\theta)$, then the platform profit is $\pi(y) = \zeta\left[1-\hat{x}^S_{2,d2}(\theta,\epsilon)+y-\hat{x}^S_{1,d2}(\theta,\epsilon,y)+1-\hat{x}^S_{2,d1}(\theta,\epsilon)+y-\hat{x}^S_{1,d1}(\theta,y)\right]$, however, the first-order condition is given by

$$FOC: \frac{\partial\pi(y)}{\partial y} = \zeta\left[1-\frac{y}{\sqrt{\alpha^2-\hat{y}^S_{d2}(\theta,\epsilon)^2+y^2}}+1-\frac{\theta y}{\sqrt{\alpha^2+1-2v+\theta[y^2-1]}}\right] = 0. \qquad \text{(E9)}$$

Using Mathematica's FindInstance function we can show that $\partial\pi(y)/\partial y$ is always positive. Thus, $y^{S*}_d(\theta,\epsilon) = \hat{x}^S_{2,d2}(\theta,\epsilon)$. For clarity, we do not include this scenario in the lemma.

Case III. We now consider when $\hat{y}^S_{d2}(\theta,\epsilon) \geq \hat{x}^S_{2,d2}(\theta,\epsilon)$. The moderation threshold y should be in the range $[\hat{x}^S_{2,d2}(\theta,\epsilon), \hat{y}^S_{d2}(\theta,\epsilon)]$. Every user with $x \leq y$ participates on the platform in both periods when $y \leq \hat{y}^S_{d2}(\theta,\epsilon)$ and every user with $x \geq \hat{x}^S_{2,d2}(\theta,\epsilon)$ or $x \geq \hat{x}^S_{2,d1}(\theta)$ participates in the second and first period, respectively. This means that all users participate in both periods, and the platform covers the market in both periods by choosing any $y^{S*}_d(\theta,\epsilon) \in [\hat{x}^S_{2,d1}(\theta), \hat{x}^S_{2,d2}(\theta,\epsilon)]$. As a tie-breaking rule, we assume the platform selects the lowest optimal moderation threshold, $y^{S*}_d(\theta,\epsilon) = \hat{x}^S_{2,d2}(\theta,\epsilon)$, to make the platform as moderated as possible. □

E.2. The Platform's Optimal Shadowbanning Threshold and User Segmentation in the Equilibrium (Section 9)

The platform does not implement shadowbanning when user posting utility and belief are sufficiently high (*Case I* of Lemma E1), and regardless of the increase in the beliefs of shadowbanned users (ϵ), there are conditions under which our dynamic and static models yield identical outcomes in terms of profitability, shadowbanning thresholds, and user segmentation. In addition, the increase in shadowbanned user beliefs, ϵ, reduces the threshold belief above which the platform may choose not to implement shadowbanning ($\partial\bar{\theta}_d/\partial\epsilon < 0$). Although the platform avoids shadowbanning when user beliefs are high in our static model, in our dynamic model it may refrain from shadowbanning even with moderate or low initial beliefs. Furthermore, under certain conditions, the platform covers the market across both periods in our dynamic model, just as it does in our static model (*Case III* of Lemma E1).

The moderation threshold in our dynamic model is one of the values from the set $\{1, \sqrt{2v}, \hat{y}^S_{d2}(\theta,\epsilon), \hat{x}^S_{2,d2}(\theta,\epsilon)\}$. Comparing these with the moderation threshold from our static model, we find that the platform always selects a shadowbanning threshold that is greater than or equal to the threshold in our static model. This is because for any ϵ, $\hat{y}^S_{d2}(\theta,\epsilon)$ and $\hat{x}^S_{2,d2}(\theta,\epsilon)$ are greater than $\hat{y}^S_{d1}(\theta)$ and $\hat{x}^S_{2,d1}(\theta)$ respectively. Note that the indifferent users in the first period of our dynamic model, $\hat{y}^S_{d1}(\theta)$ and $\hat{x}^S_{2,d1}(\theta)$, are identical to those in our static model.

This higher shadowbanning threshold benefits the platform in two ways: a) By refraining from shadowbanning some extreme users in the first period, the platform reduces the likelihood that participants from the first period update their belief and, as a result, do not participate in the second period. b) By attracting moderate users in the first period, these users maintain their initial beliefs and are less likely to leave the platform due to belief updates in the second period. However, a higher shadowbanning threshold could also lead to a reduction in the participation from less-extreme users. This is because $\hat{x}^S_{1,d1}(\theta, y)$ and $\hat{x}^S_{1,d2}(\theta,\epsilon, y)$ are increasing in y. Thus, in our dynamic model, if the platform chooses a higher shadowbanning threshold, then the position of these indifferent users increases; consequently, certain less-extreme users do not participate on the platform.

Comparing the profit of the platform in our dynamic and static models, we find that the platform's long-term profit in our dynamic model is not necessarily higher than that of our static model. This is because, although a higher shadowbanning threshold encourages participation from moderate users, it may also result in reduced participation from less-extreme users. Furthermore, even with an increased shadowbanning threshold, some extreme users may choose not to participate depending on their initial beliefs, θ, and the increase in their beliefs, ϵ. Overall, depending on other parameters, the profit in our dynamic model could be either higher or lower than that of our static model.

References

Benmoussa M, Chénier I, Keenan-Pelletier M, Mason R, Robichaud M, Tanguay L, Valiquet D, Walker J (2024) Legislative summary of Bill C-63: An act to enact the online harms act, to amend the criminal code, the canadian human rights act and an act respecting the mandatory reporting of internet child pornography by persons who provide an internet service and to make consequential and related amendments to other acts. Technical Report 44-1-C63-E, Library of Parliament, Canada, URL `https://lop.parl.ca/staticfiles/PublicWebsite/Home/ResearchPublications/LegislativeSummaries/PDF/44-1/PV_44-1-C63-E.pdf`, accessed: 2025-04-22.

Canadian Heritage (2024) Backgrounder – government of canada introduces legislation to combat harmful content online, including the sexual exploitation of children. `https://www.canada.ca/en/canadian-heritage/news/2024/02/backgrounder--government-of-canada-introduces-legislation-to-combat-harmful-content-online-includi`
`html`, accessed: 2025-04-22.

De Streel A, Defreyne E, Jacquemin H, Ledger M, Michel A, Innesti A, Goubet M, Ustowski D (2020) Online platforms' moderation of illegal content online. *Law, Practices and Options for Reform* .

European Commission (2018) Commission recommendation on measures to effectively tackle illegal content online. `https://ec.europa.eu/commission/presscorner/detail/en/memo_18_1170`, press memo, Accessed: 2025-04-22.

European Commission (2022) Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (digital services act). `https://eur-lex.europa.eu/eli/reg/2022/2065`, official Journal of the European Union, L 277/1, Accessed: 2025-04-22.

Human Rights Watch (2018) Germany: Flawed social media law. *Human Rights Watch* URL `https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law`, accessed: 2025-04-29.

Langevin L, Lenz J, Poitras C, Handa S (2024) Canada's bill c-63: Online harms act targets harmful content on social media. URL `https://www.blakes.com/insights/canada-bill-c63-online-harms-act-targets-harmful-content-on-social-media/`, accessed: 2025-04-22.

Liu Y, Yildirim P, Zhang ZJ (2022) Implications of revenue models and technology for content moderation strategies. *Marketing Science* 41(4):831–847.

Wikipedia (2024) Network enforcement act. URL `https://en.wikipedia.org/wiki/Network_Enforcement_Act`, accessed: 2025-04-22.