***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

The task of balancing online safety with preserving freedom of expression is increasingly challenging for social media platforms. We examine shadowbanning (making a user's content hidden without their knowledge) and content removal (removing content with a user's knowledge) to determine their relative impacts on market coverage, moderated content, platform profit, consumer surplus, and welfare. With shadowbanning, users assess a probability, or a belief, that a platform implements shadowbanning. We employ a stylized formulation from the literature where users receive utility from posting content and reading content, and may face disutility from reading extreme content. We find that if common user beliefs that the platform implements shadowbanning are not too high, then the platform prefers to shadowban over no moderation. Content removal helps the platform expand the user base across users with a low to moderate degree of extremeness, whereas shadowbanning expands the user base across users with any degree of extremeness. Consequently, the platform implementing shadowbanning has a larger user base and higher profit than one using content removal. In addition, the platform can cover the market with shadowbanning; however, it cannot do so with content removal. Even with a larger user base, shadowbanning results in a larger volume of moderated content compared to content removal. Finally, if users hold moderate common beliefs about shadowbanning, then shadowbanning increases consumer surplus and social welfare over content removal or no moderation. Our main findings remain mostly consistent when user beliefs about shadowbanning implementation are heterogeneous. However, unlike with common beliefs, highly extreme users may drop out. Additionally, results from our extended models—shadowbanning with imperfect technology, shadowbanning with mandated content removal, and shadowbanning with updated beliefs—suggest that shadowbanning can be beneficial, although the policy implications depend on which metric is prioritized: market coverage, volume of moderated content, platform profit, consumer surplus, or social welfare.

*Keywords*: Content Moderation, Social Media Platform, Content Removal, Shadowbanning

## 1. Introduction

Social media platforms have become an integral part of modern communication, with global daily usage averaging 151 minutes per person (Dixon 2023). Although their significant influence on daily life is well established (CIGI-IPSOS 2019), concerns about user safety and content moderation strategies are growing. Online users prefer to participate in a safe environment (e.g., one with less harassment, toxicity, or fake content); yet, platform content moderation strategies—such as content removal (deleting inappropriate content) and shadowbanning (making a user's content hidden without their knowledge)—face criticism related to freedom of speech and freedom of reach.

Content removal is the most prevalent content moderation strategy used by social media platforms; however, there is a growing tendency toward alternative strategies, in particular, shadowbanning (Fowler 2022). With shadowbanning, the user can continue to post content and interact on a platform, but their content is hidden from others. One of the earliest documented examples of shadowbanning is a description by Michael Pryor, president and co-founder of Fog Creek Software, Inc., regarding a content moderation mechanism in the project management software FogBugz: *"You take their post and make it invisible to everyone else, but they still see it. They won't know they've been deleted... They get silenced and eventually just go away"* (Walsh 2006, p. 183). Pryor's statement captures the essence of shadowbanning, which originally referred to restricting the visibility of a user's content without their knowledge or explicit notification (Cotter 2023). In more recent usage, "shadowbanning" often refers more broadly to alternative visibility approaches, such as delisting or downranking a user's content (Leerssen 2023). Shadowbanning is not limited to social media and is used in other areas as well. For instance, shadowbanning is commonly used for hiding comments on news and e-commerce websites (Apodaca and Uzcátegui-Liggett 2024). In addition, shadowbanning has been noted on Craigslist.com, an online platform primarily used for advertising. Despite advertisers receiving a confirmation email and having their ad visible on their Craigslist account page, the ad still may fail to be displayed publicly (Dstribute 2022).

Social media platforms engage in content moderation for multiple reasons, including providing a safe online community, ensuring content quality, attracting and retaining advertisers, maintaining reputation, and ensuring legal compliance. However, they encounter complex trade-offs in their content moderation design choices because they face competing incentives and have different stakeholders, including content creators, content consumers, advertisers, shareholders, and governments. As these groups often have divergent goals, a platform may prefer to give them incomplete information. Shadowbanning can help the platform convince all stakeholders that their interests are being met while also shielding itself from criticism, scrutiny, and consequences (Gillespie 2022, Pasquale 2015). For instance, in situations where a platform is uncertain if the content violates established laws, it prefers shadowbanning as a less restrictive alternative to content removal in order to legitimize moderation actions and avoid accusations of censorship (Brown 2021). Shadowbanning has additional benefits for a platform. First, shadowbanning can reduce the speed at which potentially harmful content goes viral. Second, shadowbanning obviates the need for an explicit policy and provides flexibility for intervention. Finally, shadowbanning reduces the chance of backlash by users, yielding fewer appeals and preventing reverse engineering of the automated content detection systems.

However, because shadowbanning is characterized by a particular form of secrecy, primarily resulting from the absence of clear rules as well as users' inability to detect if their posts are (or will be) shadowbanned (Leerssen 2023), the moderation strategy is subject to ongoing debate. Critics argue that it can be abused to suppress certain viewpoints or unfairly target users without transparency. Another concern is that shadowbanned content typically has reduced reach, lowering a user's engagement rate (i.e., how much other users view their content). As a result, a user's ability to attract sponsorship or advertising is negatively affected, which can have significant financial consequences for those attempting to monetize their content (Nicholas 2022). Users also experience a missed opportunity: with content removal, users are notified when their posts are taken down and can re-post them; with shadowbanning, they receive no such notification. Furthermore, social media platforms often deny engaging in shadowbanning and may mislead users about whether their content is being moderated in this way (Cook 2020, Constine 2019). Finally, platforms frequently depend on automated systems and artificial intelligence (AI)-driven content moderation. Given that one form of shadowbanning occurs through automated and AI-driven recommendation systems, errors and inaccuracies can introduce biases into these algorithms, which build upon themselves over time. For example, X has shadowbanned accounts based solely on their association with other shadowbanned accounts, not because of their individual activity (Oremus 2019).

We focus on the economic foundations of a platform's incentive to moderate content with different strategies and their public policy implications. We develop a model that allows us to compare content removal and shadowbanning to determine their impacts on platform profit, consumer surplus, and social welfare. In our model, Internet users have the option to participate on an ad-based platform to post and read content. Users are heterogeneous in the degree of extremeness of their content and they gain either positive or negative utility from reading content which depends on the content's extremeness relative to their own. A platform decides whether to moderate content (shadowban or remove content) to maximize its profit. Users know whether the platform uses content removal; however, they are uncertain about the platform's use of shadowbanning. Users decide to participate on the platform based on their beliefs regarding the probability that shadowbanning is implemented by the platform. Our model allows us to examine the effects of various factors (e.g., users' posting and reading utilities, user beliefs, and imperfect technology) on the platform's moderation decisions, user participation, and welfare.

Our work has produced multiple findings. First, we demonstrate that a platform does not always have incentive to implement shadowbanning. If common user beliefs that the platform implements shadowbanning are high, then the platform prefers to shadowban only when the posting utility is relatively greater than the reading utility; otherwise, it always prefers shadowbanning over no moderation. Thus, similar to content removal, the platform's moderation decision is influenced by the relative size of the posting utility versus the reading utility. Second, similar to content removal, shadowbanning expands the platform's user base. However, the mechanisms through which this expansion occurs differ between the two moderation strategies. Content removal achieves this expansion by excluding extreme users, whereas shadowbanning reduces the disutility of reading extreme content for all users. Although content removal primarily attracts users with a low to moderate degree of extremeness, shadowbanning allows the platform to attract users across the spectrum of extremeness. Indeed, more-extreme users can be the majority of users participating on a platform implementing shadowbanning. Third, with shadowbanning, the platform chooses stricter moderation, resulting in the moderation of a larger volume of content compared to content removal, except in cases where posting or reading utilities are very small. Fourth, shadowbanning increases consumer surplus compared to content removal if users have moderate beliefs, and users care more about reading others' content than posting their own. However, shadowbanning never leads to a higher consumer surplus compared to no moderation. Lastly, shadowbanning increases social welfare compared to content removal and no moderation if users have moderate beliefs.

In an extension discussed in Section 6, we examine heterogeneous user beliefs that increase in extremeness and find the main results hold with some differences. With common beliefs, extreme users always participate, but with heterogeneous beliefs, highly extreme users may opt out. This can reduce platform profit from shadowbanning, leading it to prefer content removal or no moderation. This finding aligns with our observation that platforms have no motivation to disclose their use of shadowbanning to users. Additionally, shadowbanning with heterogeneous beliefs always results in lower consumer surplus and social welfare than other strategies, unlike with common beliefs where it may sometimes outperform them.

We also investigate the effects of imperfect technology for detecting extreme content when the platform implements shadowbanning, as discussed in Section 7. We find that a platform's decision to implement shadowbanning with imperfect technology depends on the accuracy of the technology, and that user beliefs no longer influence this decision. Specifically, if the technology is sufficiently accurate, then the platform may adopt shadowbanning regardless of user beliefs. This result highlights the importance of measures beyond transparency—such as accuracy, fairness, and bias mitigation—and points to the need for mechanisms that enforce and monitor accuracy standards for moderation technologies.

In addition, we investigate when a platform may implement shadowbanning under mandated content removal policies (Section 8). We find that strict content removal policies deter shadowbanning when users prioritize posting over reading; otherwise, the platform has an incentive to shadowban. When common user beliefs about shadowbanning are low, stricter (more lax) content removal policies lead to stricter (more lax) shadowbanning, but this response reverses with high common user beliefs. This counterintuitive finding suggests that blanket content removal policies, if implemented without considering user beliefs and other moderation strategies, could produce unintended consequences such as over- or under-moderation.

Finally, we extend our model to a two-period dynamic setting in which shadowbanned users update their beliefs about being shadowbanned again. As shown in Section 9, the results closely align with those of the static model, with some differences. Although the static model shows the platform refrains from shadowbanning only when common user beliefs are high, the dynamic model suggests the platform may avoid shadowbanning even with moderate or low initial common beliefs. Thus, a significant increase in the beliefs of shadowbanned users can reduce the platform's incentive to implement shadowbanning in the long term.

These findings highlight the complex trade-offs platforms face and underscore the importance of aligning moderation strategies with broader social and economic outcomes. The policy implications of our analysis depend on which metric—market coverage, volume of moderated content, platform profit, consumer surplus, social welfare, or a combination thereof—is prioritized. Table 4 summarizes key results across our base and extended models.

The remainder of the paper is organized as follows. Section 2 reviews the literature, and Section 3 outlines the model's notation and assumptions. Section 4 summarizes prior results on no moderation and content removal. Section 5 presents our analysis of shadowbanning with common user beliefs and compares outcomes across moderation strategies. Section 6 extends the model to heterogeneous user beliefs. Sections 7 and 8 examine imperfect detection technology and shadowbanning with mandated content removal, respectively. Section 9 explores a dynamic model where users update their beliefs. Section 10 concludes.

## 2. Literature Review

Our study relates to literature about content moderation strategies used by social media platforms, an area with limited published theoretical research. Liu et al. (2022) are among the first to model a platform's decisions regarding content removal moderation and to investigate how different revenue models (advertising and subscription fee) affect the incentives of the platform for moderation. They also investigate the effect of imperfect technology for detecting extreme content on a platform's decisions. Some of our set-up is similar to that in Liu et al. (2022), but we focus on shadowbanning.

Madio and Quinn (2024) examine incentives for advertising-supported platforms to invest in content moderation to attract advertisers and control advertising pricing. They show that mandating moderation of unsafe content can benefit advertisers. However, users may be worse off due to being exposed to a greater number of ads. Unlike their work, our research incorporates content creators, excludes advertiser decision making, and focuses on shadowbanning rather than content removal. Beknazar-Yuzbashev et al. (2022) show that a platform moderates content at the expense of content consumption. They further provide a theoretical argument that a platform's incentives for moderating harmful content are not necessarily aligned with those of users (Beknazar-Yuzbashev et al. 2024). Similarly, we investigate the platform's incentives for content moderation.

Kominers and Shapiro (2024) examine the possibilities and limitations of content moderation, differentiating between transparent moderation, where users are informed about and trust the moderator's policy, and opaque moderation, where users lack knowledge of the policy and may distrust the moderator's intentions or information. They find that without robust reputation systems or other trust-building mechanisms, content moderation is effective in blocking only a specific type of harmful content. Similarly, we explore the effects of shadowbanning as an opaque moderation strategy. However, shadowbanning differs significantly from the strategies they examine. Although they examine two types of harmful content, we consider a continuum of content differing in its extremeness. Furthermore, they focus on the effectiveness of the moderation strategies and the role of trust in determining their success, whereas we focus on comparing the impacts of content removal and shadowbanning on user participation, platform profit, and welfare.

We investigate the platform's incentives for content moderation, whereas other literature explores content moderation from different perspectives. Mostagir and Siderius (2023) and Dwork et al. (2024) study the impacts of content moderation policies on online communities and content diversity. Fagan (2020) investigates content moderation from the policymakers' perspective, in which lawmakers decide whether platforms should be liable for the content posted on their platforms. Wu (2023) examines the effects of a social planner's decision to enforce stricter policies aimed at limiting the consumption of sensitive content. Their findings reveal that content creators shift from open channels to secret ones, ultimately increasing social harm. Nan et al. (2023) examine the interplay between government regulation and platforms' self-regulation and show that competition influences platforms' regulation intensity and impacts consumer surplus. Our research, like these studies, provides insights for policymakers, specifically regarding the impacts of various content moderation strategies on platform profit, user participation, and moderation threshold.

Our economic modeling approach complements Information Systems research on platform governance and algorithmic transparency. Tiwana et al. (2010) establish that platform governance involves balancing openness with control—a fundamental tension evident in content moderation decisions where platforms must coordinate user activities while maintaining ecosystem value. Building on this foundation, He et al. (2024) provide empirical evidence that algorithmic content moderation creates distinct economic tradeoffs compared to human moderation, particularly regarding scalability and resource allocation. Huber et al. (2017) demonstrate how platforms navigate tensions between standardized governance mechanisms (like automated rules) and contextual sensitivity, explaining why platforms might choose opaque moderation strategies to minimize governance costs while preserving flexibility. Additionally, Möhlmann et al. (2021), Bauer and Gill (2024) show that algorithmic opacity in platform management triggers diverse user behavioral responses, providing insights into how users might adapt to shadowbanning strategies. This literature on platform governance mechanisms and algorithmic transparency complements our economic analysis by highlighting the institutional and behavioral dimensions of content moderation decisions.

Several empirical studies investigate the effects of content moderation on user engagement (Beknazar-Yuzbashev et al. 2022), harmful or toxic content (Müller and Schwarz 2023, Andres and Slivko 2021), content characteristics and user participation (Gopal et al. 2025, Dwork et al. 2024), offline behaviors (Müller and Schwarz 2023), and welfare (Jiménez-Durán 2023). Other studies examine the impact of automated moderation tools, such as bots (He et al. 2024), the role of providing explanations following moderation actions (Jhaver et al. 2019), and the influence of regulations like Germany's Network Enforcement Act (NetzDG) on user behavior and digital communities (Schwarz et al. 2022). The above research studies moderation strategies that are transparent or have potential to be transparent. In contrast, a small body of empirical research explores the effects of shadowbanning as inherently opaque. Chen and Zaman (2024) introduce an optimization-based approach and simulate a shadowbanning strategy designed to shape opinions toward a specific distribution. Their findings reveal that shadowbanning can manipulate opinions and amplify polarization. Other empirical studies on shadowbanning rely on interviews, surveys, and experiments (Delmonaco et al. 2024, Zhao 2024, Jiménez-Durán 2023), due to the lack of publicly available data on shadowbanned content. Although the EU's Digital Services Act (DSA) recently introduced a unique new data source (European Commission 2023), significant challenges remain in comprehensively investigating platform moderation actions, particularly concerning accuracy, fairness, and bias, because of privacy concerns. This highlights the importance of theoretical studies like ours, which offer insights into the broader implications of content moderation.

## 3. Notation and Assumptions

In this section, we outline the assumptions underlying our theoretical model, which is in part built upon the work of Liu et al. (2022). Given our use of specific forms, when stating and deriving functions, we suppress arguments relating to parameters in order to focus on variables.

### 3.1. Users

We consider a social media platform with a mass of users that post their own content and read content posted by others. The degree of extremity or offensiveness of the content is assessed on a scale ranging from zero to one, where one is the most extreme. We define a user by their content's degree of extremeness (how extreme their content typically is). That is, a user located at $x$ is the one that posts content with a degree of extremeness $x \in [0,1]$. Therefore, we take users to be differentiated based on the degree of extremeness of their content, leading to the following assumption.

**Assumption 1.** ***(User Heterogeneity)*** *Users differ in their content's degree of extremeness, which follows a uniform distribution denoted as* $x \sim U[0,1]$.

Users derive two forms of utility from their interactions with the platform: one from posting (creating and sharing) content and another from reading (consuming) content. According to research in consumer psychology, people that have more extreme ideas tend to be more vocal about them (Miller and Morrison 2009, Yildirim et al. 2013, Mathew et al. 2019). Liu et al. (2022) assume that the utility of posting content on social media differs among users and increases in their degree of extremeness, $x$. That is, users with a greater degree of extremeness, a higher $x$, derive greater utility from posting. We use $\alpha > 0$ to scale posting utility from extremeness. A higher value of $\alpha$ indicates a greater difference in the posting utilities of any two users.

**Assumption 2.** ***(Posting Utility)*** *Each user posts one content and a user with the degree of extremeness $x$ gains utility $\alpha x$ from posting a content, where* $\alpha \geq 0$.

Liu et al. (2022) argues that users typically prefer content that aligns with their preferences. In their model, this translates to users experiencing the highest reading utility, denoted by $v \in \mathbb{R}^+$, when they encounter content with a degree of extremeness similar to their own. Recognizing that a user is exposed to all content on the platform, they further assume $v < 1/2$ to ensure that the least extreme user located at $x = 0$ has a negative utility if content moderation is not employed. We make the same assumption.

**Assumption 3.** ***(Reading Utility)*** *Users gain utility from reading content on social media, which is constant and denoted as $v$, where* $v < 1/2$.

There are few studies on users' preferences towards social media content and how they respond to content that varies in extremeness compared to their own. Liu et al. (2022) rely on a plausible assumption known as "extremeness aversion": a user tends to feel uncomfortable when encountering content that surpasses their degree of extremeness, but may be more accepting of content that is less extreme than their own.

**Assumption 4.** ***(Reading Disutility)*** *A user with a degree of extremeness $x$ faces disutility $z$ from reading content with a degree of extremeness* $z \in [x,1]$.

Given that users are exposed to and consume all content available, the total reading utility for a user at $x$ reading posts with a degree of extremeness in the range of $[0,1]$ is given by $v - \int_x^1 z dz$. The utility function of a user includes both posting and reading utilities. When the platform does not conduct any moderation, the user utility function, denoted as $U(x)$, is given by Liu et al. (2022):

$$U(x) = \alpha x + v - \int_x^1 z dz. \tag{1}$$

### 3.2. Platform Moderation Strategy

The platform decides whether to conduct content moderation and, if so, chooses the optimal strategy and the optimal moderation threshold. The platform determines an optimal moderation threshold based on extremeness, which is indicated by $y \in [0,1]$. Content from users with a degree of extremeness exceeding this threshold ($x > y$) undergoes moderation and content from users with a degree of extremeness $x \leq y$ is not moderated. Notably, $y = 1$ implies no content moderation and $y = 0$ indicates that all content is subject to moderation. Platforms typically depend on human moderators, AI systems, or a combination of both to identify and moderate extreme content. We assume that the platform can accurately identify and moderate extreme content, and later relax this assumption in Section 7.

**Assumption 5.** *The platform accurately detects and can moderate all extreme content.*

We define three content moderation strategies: 1) *No Moderation (N)*, where the platform chooses not to moderate any content; 2) *Content Removal (R)*, where the platform moderates extreme content and removes it from the platform; and 3) *Shadowbanning (S)*, in which the platform moderates extreme content by employing shadowbanning (see Section 3.2.2). When the platform does not implement any moderation ($N$), the user's utility function remains as in (1). However, when the platform adopts a moderation strategy ($R$ or $S$), it impacts users' utility functions.

**3.2.1. Content Removal** If the platform chooses a content removal strategy, then any content with a degree of extremeness greater than the platform's moderation threshold, $y$, is removed. Removing extreme content from the platform has an opportunity cost, $\alpha x$, as well as a psychological cost, $c \in \mathbb{R}^+$, for users, caused by the limitations imposed on their ability to post such content. To discourage users subject to content moderation from participating in the platform, we take the cost of being moderated for users to be higher than the reading utility.

**Assumption 6.** *The cost of being moderated for users, c, is a) uniform across both content removal and shadowbanning strategies, and b) exceeds their reading utility, $c > v$.*

The anticipated utility that a user gains from participating in a social media platform with content removal is contingent upon whether the user's degree of extremeness surpasses or falls short of the platform's moderation threshold, $y$. Users with $x \leq y$ experience disutility only from content falling within the range of $[x, y]$, and they do not experience disutility from content within $[y, 1]$. This is due to the removal of posts in the range $[y, 1]$. Users with $x > y$ experience a moderation cost $c$ which is higher than the reading utility and these users do not participate on the platform. The user utility function facing content removal is given by

$$U^R(x) = \begin{cases} \alpha x + v - \int_x^y z\,dz & if x \leq y \\ -c + v & if x > y. \end{cases} \tag{2}$$

**3.2.2. Shadowbanning** If the platform implements shadowbanning, then any content with a degree of extremeness greater than the platform's moderation threshold is moderated. Shadowbanned content remains on the platform but is hidden from all users. Therefore, there is no reading disutility for users because they do not read the shadowbanned content.

As discussed earlier, although users are uncertain whether the platform actually implements shadowbanning, their knowledge that the platform has the capability implies that their posts may be shadowbanned. Also, the platform does not notify the user when the content is shadowbanned. Therefore, users do not know the platform's moderation strategy. We take two approaches to user beliefs about shadowbanning: common beliefs and heterogeneous beliefs. With common beliefs, all users believe that the probability that the platform implements shadowbanning is $\theta$, where $\theta \in (0,1)$, and the probability that the platform does not implement shadowbanning is $1-\theta$.

**Assumption 7a**. ***(Common Beliefs)*** *Users do not know whether the platform shadowbans content. They share the same prior belief that the platform implements shadowbanning.*

Similar to content removal, users incur a moderation cost when their content is shadowbanned. The expected utility that users derive depends on their belief that the platform implements shadowbanning. With common beliefs, the expected utility for users can be represented as

$$E(U^S(x)) = \begin{cases} \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + \left[\alpha x + v - \int_x^y z\,dz\right]\theta & if x \leq y \\ \left[\alpha x + v - \int_x^1 z\,dz\right][1-\theta] + [-c+v]\,\theta & if x > y. \end{cases} \tag{3}$$

As shown in (3), when $\theta = 1$, the user's expected utility is equivalent to that under content removal. When $\theta = 0$, it corresponds to the case of no moderation. Thus, both content removal and no moderation arise as special cases of the model we propose. Although an important difference exists between content removal and shadowbanning (that users are unaware that the platform employs shadowbanning, as discussed in Section 1) that is captured through a static incomplete information (Bayesian) game, our model can be interpreted as a generalization of content removal.

Although users are unaware of whether the platform implements shadowbanning and are not notified when their content is shadowbanned, they may not share a common belief about its implementation. Users that frequently engage with or publish extreme content are often more aware of potential moderation consequences. This awareness stems from their experiences with content moderation on other platforms or in different contexts, community discussions about different moderation approaches, as well as from observing general patterns of how similar content is treated across the platform. Consequently, they may perceive their own content as being at higher risk of shadowbanning. Therefore, users form beliefs about the likelihood that shadowbanning occurs without being able to definitively determine whether their specific posts are shadowbanned. Conversely, users that regularly interact with or produce less-extreme content may be less aware of the platform's policies regarding extreme content. To address this, we consider heterogeneous beliefs, where user belief that the platform implements shadowbanning increases with the degree of extremeness of their content. We take user belief that the platform implements shadowbanning as $\beta x$, where $\beta \in (0,1)$, and the belief that shadowbanning is not implemented is $1-\beta x$.

**Assumption 7b**. ***(Heterogeneous Beliefs)*** *Users do not know whether the platform shadowbans content. They have heterogeneous beliefs that the platform implements shadowbanning, denoted by* $\beta(x) = \beta x$.

With heterogeneous user beliefs, the expected utility for a user is expressed as follows, with the analysis discussed in Section 6.

$$E(U^S(x)) = \begin{cases} \left[\alpha x + v - \int_x^1 z\,dz\right][1-\beta x] + \left[\alpha x + v - \int_x^y z\,dz\right]\beta x & if x \leq y \\ \left[\alpha x + v - \int_x^1 z\,dz\right][1-\beta x] + [-c+v]\beta x & if x > y. \end{cases} \tag{4}$$

### 3.3. Platform Profit

We focus on a monopolist platform with an advertising-based revenue model to investigate the platform's content moderation strategy. We take the platform's moderation cost as constant, and because it has no effect on the outcome we consider it zero. The platform charges advertisers for each user (or piece of content). As a result, the overall advertising revenue scales with the platform's user base, making it directly proportional to the number of users on the platform. The platform profit function is given by

$$\max_y \pi(y) = \max_y \zeta X(y), \tag{5}$$

where $\zeta \in \mathbb{R}^+$ is the per-user advertising fee that is determined by a competitive market and is exogenous to our model. $X(y)$ indicates the platform's user base, that is, the proportion of users that participate on the platform because their utility is positive. We define $X(y)$ later.

### 3.4. Timing

We model the problem in two stages. In the first stage, the platform decides whether to conduct moderation and, if so, the optimal moderation threshold given its strategy. In the second stage, users decide whether to participate on the platform. If the strategy is content removal, then users observe the platform's moderation threshold. If the strategy is shadowbanning, then they do not.

## 4. No Moderation and Content Removal (Liu et al. 2022)

We briefly restate the no moderation and content removal analyses presented in Liu et al. (2022).

### 4.1. No Moderation Strategy ($N$)

The platform does not moderate content, thus its moderation threshold is $y = 1$. The solution for no moderation can be characterized by the least extreme user that joins the platform, denoted by $\hat{x}^N(1)$. Because the users' utility is increasing in $x$ according to (1), the indifferent user is given by $\hat{x}^N(1) = -\alpha + \sqrt{\alpha^2 + 1 - 2v}$ as derived from solving $U^N(x^N(1)) = 0$ for $x$. Consequently, users in range $[\hat{x}^N(1), 1]$ participate on the platform and form the platform's user base. For a platform with no moderation, the user base is given by $X^N(1) = \int_{\hat{x}^N(1)}^{1} dx$ and is illustrated in Figure 1a by the bold line. From (5), the platform profit is given by $\pi^N(1) = \zeta X^N(1) = \zeta[1 + \alpha - \sqrt{\alpha^2 + 1 - 2v}]$.

### 4.2. Content Removal Strategy ($R$)

For content removal, the user that is indifferent between using the platform and not is denoted by $\hat{x}^R(y)$, and any users in the range $[\hat{x}^R(y), y]$ participate on the platform. Thus, the user base is $X^R(y) = [\hat{x}^R(y), y]$ which is illustrated in Figure 1b using the bold line. Comparing the user bases of these two strategies, no moderation and removal, in Figure 1, extreme users with $x > y$ do not participate with content removal, leading to a reduction in the platform's user base; however, users with some degree of extremeness, at least $\hat{x}^R(y)$, participate and increase the user base.

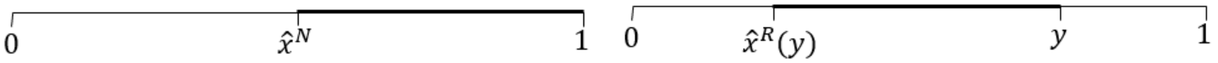
From (2) we can solve for $\hat{x}^R(y)$ by setting $U(x^R(y))$ to zero for any given moderation threshold $y$. The solution is given as follows:

$$\hat{x}^R(y) = \begin{cases} -\alpha + \sqrt{\alpha^2 + y^2 - 2v} & if \quad y \geq \sqrt{2v} \\ 0 & if \quad y < \sqrt{2v}. \end{cases} \tag{6}$$

If $y \geq \sqrt{2v}$, then the platform has more lax content moderation, such that more content moderation, or decreasing $y$, decreases $\hat{x}^R(y)$. As a result, more users with less extreme views participate on the platform. This is how moderating extreme content can help the platform expand its user base. When $y < \sqrt{2v}$, the platform employs stricter content moderation, and all users with less extreme views are already participating. Further content removal reduces the platform's user base.

#### 4.2.1. The Platform's Moderation Decision

The platform chooses the optimal threshold for moderating content, $y^{R*}$, to maximize its revenue. We restate the proposition from Liu et al. (2022) that summarizes the optimal moderation threshold for content removal for a platform with an advertising revenue model.



(a) No Moderation Strategy (N)

(b) Content Removal Strategy (R)

**Figure 1 Illustration of the Platform's User Base**

**Proposition 1.** ***(Liu et al. 2022, p. 6)*** *"A platform with advertising as its revenue model does not always have incentives to conduct content moderation. It conducts content moderation $y^{A*} = \sqrt{2v}$ if and only if the posting utility in the market is sufficiently small relative to the maximum reading utility, or $\alpha < \alpha^A \equiv \sqrt{2v}$ . The optimal revenue is given by $\pi^{A*} = \zeta\sqrt{2v}$ . Otherwise, the platform does not moderate content ($y^{A*} = 1$), and its optimal revenue is given by $\pi^{A*} = \zeta[1 + \alpha - \sqrt{\alpha^2 + 1 + 2v}]$".*

To increase its profit, the platform has to increase its user base by determining the optimal $y$. From (5) the platform profit is $\pi^R(y) = \zeta X^R(y) = \zeta[y - \hat{x}_1^R(y)] = \zeta[y + \alpha - \sqrt{\alpha^2 - 2v + y^2}]$. The first-order condition of the platform profit with respect to $y$ is

$$\frac{\partial \pi^R(y)}{\partial y} = \zeta \left[1 - \frac{y}{\sqrt{\alpha^2 - 2v + y^2}}\right] = 0. \tag{7}$$

If $\alpha \geq \sqrt{2v}$, then $\partial\pi^R(y)/\partial y \geq 0$ and if $\alpha < \sqrt{2v}$, then $\partial\pi^R(y)/\partial y < 0$. Thus, if users care sufficiently more about posting utility than reading utility, $\alpha \geq \sqrt{2v}$, then the platform profit is increasing in $y$ and the platform sets the highest value for the optimal moderation threshold, $y^{R*} = 1$. This implies that the platform does not remove content. However, if $\alpha < \sqrt{2v}$, then the platform profit is decreasing in $y$ and the platform sets the minimum value for the optimal moderation threshold, $y^{R*} = \sqrt{2v}$. This implies that the platform removes content with the degree of extremeness greater than the optimal moderation threshold, $x > y^{R*} = \sqrt{2v}$. In other words, the platform removes content to expand its user base and, as a result, increases its profit if and only if $\alpha < \sqrt{2v}$.

## 5. Shadowbanning Strategy ($S$): Common User Beliefs

Following Assumption 7a, we begin by characterizing the equilibrium configuration for users with common beliefs that the platform implements shadowbanning for any given moderation threshold $y$. With shadowbanning we have two indifferent users for two groups of users. The first group consists of users with a degree of extremeness greater than the moderation threshold, $x \geq y$, with the indifferent user denoted by $\hat{x}_2^S(\theta)$. The second group consists of users with a degree of extremeness smaller than the moderation threshold, $x < y$, with the indifferent user denoted by $\hat{x}_1^S(\theta, y)$. We refer to the first group of users as more-extreme and the second as less-extreme. The following lemma summarizes our analysis. Proofs of all lemmas and propositions are given in the Appendix.

**Lemma 1.** *For any moderation threshold $y \in [0,1]$, there exists $\hat{x}_1^S(\theta, y) \in [0, y]$ and $\hat{x}_2^S(\theta) \in [0,1]$ such that, in equilibrium, all users in range $[\hat{x}_1^S(\theta, y), y]$ and $[\hat{x}_2^S(\theta), 1]$ participate on the platform. Therefore, the user base of the platform is $X^S(y) = [\hat{x}_1^S(\theta, y), y] \cup [\hat{x}_2^S(\theta), 1]$. Furthermore, $\hat{x}_2^S(\theta)$ is increasing in $\theta$ and $\hat{x}_1^S(\theta, y)$ is weakly decreasing in $\theta$.*

The indifferent users $\hat{x}_2^S(\theta)$ and $\hat{x}_1^S(\theta,y)$ are given as:

$$\hat{x}_2^S(\theta) = \begin{cases} -\alpha + \sqrt{\alpha^2+1-2v+\frac{2\theta[c-v]}{1-\theta}} & if \quad \theta < \frac{v+\alpha}{c+\alpha} = \bar{\theta} \\ 1 & if \quad \theta \geq \frac{v+\alpha}{c+\alpha} = \bar{\theta}, \end{cases} \tag{8}$$

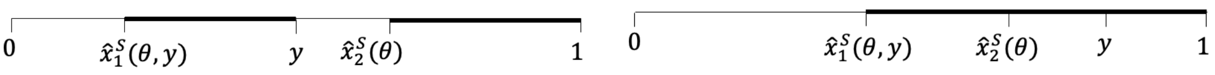
and

$$\hat{x}_1^S(\theta,y) = \begin{cases} max\{0, -\alpha + \sqrt{\alpha^2+1-2v-\hat{x}_2^S(\theta)^2+\theta\,[\hat{x}_2^S(\theta)^2-1]+y^2}\} & if \quad y < \hat{x}_2^S(\theta) \\ min\{y, -\alpha + \sqrt{\alpha^2+1-2v+\theta\,[y^2-1]}\} & if \quad y \geq \hat{x}_2^S(\theta). \end{cases} \tag{9}$$

The value of $\bar{\theta}$ is determined by solving for $\theta$ from $E(U^S(x)) = 0$ when $x = 1$. Because $E(U^S(x))$ increases in $x$, a negative value for $E(U^S(1))$ implies 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 $\hat{x}_2^S(\theta) = 1$.

Lemma 1 outlines the scenario where less-extreme users in the range $[\hat{x}_1^S(\theta,y), y]$ and more-extreme users in the range $[\hat{x}_2^S(\theta), 1]$ participate on the platform. From (9), depending on whether the moderation threshold $y$ is greater than $\hat{x}_2^S(\theta)$, shadowbanning results in two distinct user configurations as illustrated in Figure 2. Among users subject to shadowbanning, moderately extreme users experience the greatest disutility from both reading extreme content and the possibility of their content being shadowbanned. If the platform implements strict shadowbanning (smaller $y$), then the significant decrease in utility of moderate users may prompt them to not participate on the platform, leading to two disjoint user segments, as shown in Figure 2a. In contrast, if the platform implements lax shadowbanning (larger $y$), then moderate users participate, resulting in a contiguous user segment, as shown in Figure 2b.

The location of the indifferent more-extreme user, $\hat{x}_2^S(\theta)$, changes with $\theta$ and does not change with $y$. That is, the most extreme users in the range $[\hat{x}_2^S(\theta), 1]$ participate on the platform based on their beliefs that a platform may shadowban their content, $\theta$, and not the platform's moderation threshold $y$. If $\theta$ increases, then $\hat{x}_2^S(\theta)$ increases. That is, with higher user beliefs that the platform implements shadowbanning, fewer more-extreme users participate. However, the location of the indifferent less-extreme user, $\hat{x}_1^S(\theta,y)$, changes with both $\theta$ and $y$. Stricter shadowbanning (a smaller value of $y$) or higher user beliefs (a larger value of $\theta$) decrease $\hat{x}_1^S(\theta,y)$, attracting more users with less extreme content to participate, thereby increasing the platform's user base. As a result, shadowbanning can help the platform retain users and expand its user base to include both less extreme and more extreme users. In contrast, content removal and no moderation led to the platform losing more extreme and less extreme users, respectively.



(a) (b)

**Figure 2 Illustration of the Platform's User Base - Shadowbanning with Common User Beliefs**

### 5.1. The Platform's Moderation Decision

Using Lemma 1 and (5), the profit of the platform implementing shadowbanning is given by

$$\pi^S(y) = \zeta\left[1 - \hat{x}_2^S(\theta) + y - \hat{x}_1^S(\theta, y)\right]. \tag{10}$$

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

**Lemma 2.** *The optimal moderation threshold for a platform implementing shadowbanning can be characterized as follows:*

*Case I. If $\theta \geq \bar{\theta}$, then $\hat{x}_2^S(\theta) = 1$ and a) if $\alpha \geq \sqrt{2v}$, then $y^{S*}(\theta) = 1$; b) if $\alpha < \sqrt{2v}$, then $y^{S*}(\theta) = \sqrt{2v}$.*

*Case II. If $\theta < \bar{\theta}$, $\hat{y}^S < \hat{x}_2^S(\theta)$, and a) if $\theta < \theta_0$, then $y^{S*}(\theta) = \hat{x}_2^S(\theta)$; b) if $\theta \geq \theta_0$, then $y^{S*}(\theta) = \hat{y}^S(\theta)$.*

*Case III. If $\theta < \bar{\theta}$ and $\hat{y}^S \geq \hat{x}_2^S(\theta)$, then $y^{S*} \in [\hat{x}_2^S(\theta), \hat{y}^S(\theta)]$.*

*Where $\hat{y}^S(\theta) = \sqrt{2v + [1-\theta]\left[\hat{x}_2^S(\theta)^2 - 1\right]}$ is determined by solving for $y$ from $\hat{x}_1^S(\theta) = 0$, and $\theta_0$ is given by solving for $\theta$ from $\partial \pi^S(y)/\partial y = 0$. From (8), $\hat{x}_2^S(\theta) = -\alpha + \sqrt{\alpha^2 + 1 - 2v + \frac{2\theta[c-v]}{1-\theta}}$ and $\bar{\theta} = \frac{\alpha+v}{\alpha+c}$.*

In *Case I*, if user beliefs that the platform implements shadowbanning are high, $\theta \geq \bar{\theta}$, then the platform does not shadowban content when $\alpha \geq \sqrt{2v}$. Conversely, if $\alpha < \sqrt{2v}$, then the platform shadowbans content using the same moderation threshold as it would for content removal, leading to equivalent profit. The rationale for this similarity lies in the fact that when user beliefs that the platform implements shadowbanning are sufficiently high, $\theta \geq \bar{\theta}$, all more-extreme users, $x > y$, opt not to participate, thus, $\hat{x}_2^S(\theta) = 1$. Consequently, for the rest of our analysis, we focus mainly on the scenario where $\theta < \bar{\theta}$, where a platform implementing shadowbanning behaves differently than a platform implementing content removal.

Now consider *Case II*, when $\theta < \bar{\theta}$ and $\hat{y}^S(\theta) < \hat{x}_2^S(\theta)$. The platform maximizes its profit to determine the optimal moderation threshold. From (10), the first-order partial derivative of the platform profit with respect to $y$ is given by

$$\frac{\partial \pi^S(y)}{\partial y} = \zeta\left[1 - \frac{y}{\sqrt{\alpha^2 - \theta + 1 - 2v - [1-\theta]\,\hat{x}_2^S(\theta)^2 + y^2}}\right] = 0. \tag{11}$$

There is no solution for $y$ when solving $\partial \pi^S(y)/\partial y = 0$. Consequently, the optimal shadowbanning threshold is one of the corner solutions. Solving for $\theta$ from $\partial \pi^S(y)/\partial y = 0$, $\theta_0$ is given. If $\theta \leq \theta_0$, then $\partial \pi^S(y)/\partial y \geq 0$ and the platform chooses the maximum value for the optimal moderation threshold, $y^{S*}(\theta) = \hat{x}_2^S(\theta)$. If $\theta > \theta_0$, then $\partial \pi^S(y)/\partial y < 0$ and the platform chooses the minimum value for the optimal moderation threshold, $y^{S*}(\theta) = \hat{y}^S(\theta)$.

In *Case III*, when $\theta < \bar{\theta}$ and $\hat{y}^S(\theta) \geq \hat{x}_2^S(\theta)$, all users participate, thus the platform's user base equals one, $X^S = 1$. The platform profit reaches its maximum value, specifically $\pi^S(y^{S*}(\theta)) = \zeta$ with our specific forms, for all optimal moderation thresholds, where $y^{S*}(\theta) \in [\hat{x}_2^S(\theta), \hat{y}^S(\theta)]$. As a tie-breaking rule, we assume that the platform selects the lowest optimal moderation threshold, i.e., $y^{S*}(\theta) = \hat{x}_2^S(\theta)$ (representing the most strict moderation policy), to make the platform as moderated as possible.

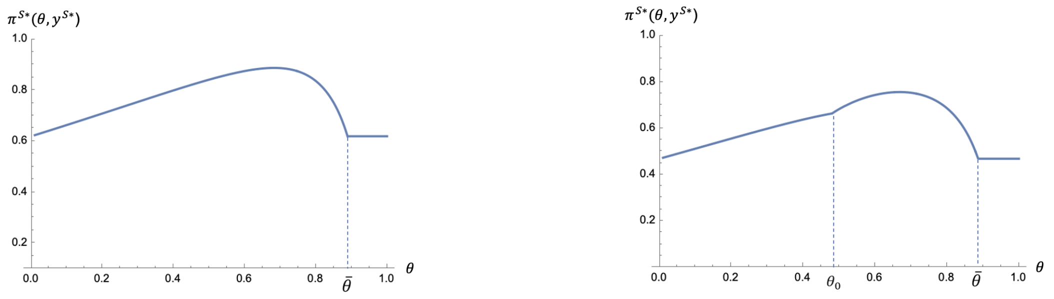
Using Lemma 2 and (10), we can determine the maximum profit of the platform. The following proposition summarizes the platform's decision for shadowbanning.

**Proposition 2.** *A platform does not shadowban content when the common user beliefs that the platform implements shadowbanning are high, $\theta \geq \bar{\theta}$, and the posting utility is sufficiently high relative to the reading utility, $\alpha \geq \sqrt{2v}$; otherwise, it implements shadowbanning.*

When the user beliefs that the platform implements shadowbanning are not high, $\theta < \bar{\theta}$, regardless of whether users gain more posting utility vs. reading utility, the platform prefers to shadowban over no moderation. Furthermore, when such user beliefs that the platform implements shadowbanning are high, $\theta \geq \bar{\theta}$, the platform shadowbans content if and only if the posting utility is sufficiently small, $\alpha < \sqrt{2v}$. Otherwise, if $\theta \geq \bar{\theta}$ and $\alpha \geq \sqrt{2v}$, then the platform does not shadowban, extreme users participate, and less-extreme users do not participate on the platform.

Proposition 2 indicates that under certain conditions shadowbanning allows the platform to have a larger user base, resulting in higher profit compared to no moderation. This is because with shadowbanning, the platform reduces the users' reading disutility, which is a significant factor in the participation of less-extreme users. Therefore, the platform implementing shadowbanning can not only retain more-extreme users, similar to no moderation, but also attract less-extreme users, resulting in higher profit compared to no moderation.

Focusing on the case where $\theta < \bar{\theta}$ and using (10), the platform's optimal profit is either $\pi^{S*}(y) = 1 - \hat{x}_2^{S*}(\theta) + \hat{y}^S(\theta)$, or $\pi^{S*}(y) = 1 - \hat{x}_1^{S*}(\theta, y)$ depending on whether $\alpha \geq \sqrt{2v}$ and whether $\theta < \theta_0$.



(a) $\alpha \geq \sqrt{2v}$

(b) $\alpha < \sqrt{2v}$

**Figure 3 Optimal Profit of the Platform Implementing Shadowbanning with Common User Beliefs**

($v = 0.2$, $c = 0.3$, $\alpha = 0.3$, or $\alpha = 0.6$)

Despite differences in the expressions of the platform's optimal profit, they are both concave. There is an initial positive relationship between profit and user belief up to a certain threshold, beyond which profit decreases as user beliefs increase. Figure 3 illustrates this.

**Lemma 3.** *The optimal moderation threshold for a platform implementing shadowbanning, $y^{S*}(\theta)$, lies within the set $\{\hat{x}_2^S(\theta), \hat{y}^S(\theta), \sqrt{2v}\}$.*

The conditions under which each threshold applies are explained in detail in Appendix **??**, and a summary of these conditions is provided in Table 1. As the optimal shadowbanning threshold $y^{S*}(\theta) = \sqrt{2v}$ is equal to the optimal content removal threshold, we focus our discussion on the characteristics of $\hat{x}_2^S(\theta)$ and $\hat{y}^S(\theta)$.

An increase in the posting utility, $\alpha$, results in a lower optimal shadowbanning threshold according to the negative comparative statics: $\partial\hat{x}_2^S(\theta)/\partial\alpha < 0$ and $\partial\hat{y}^S(\theta)/\partial\alpha < 0$. Essentially, as posting utility increases, the platform opts for stricter shadowbanning, resulting in more volume of content being shadowbanned. However, the impact of reading utility on the optimal shadowbanning threshold is not uniform. When the user beliefs that the platform implements shadowbanning are moderate, $\theta_0 < \theta$, a higher reading utility leads to more lax shadowbanning as $\partial\hat{y}^S(\theta)/\partial v > 0$. Conversely, when the user beliefs are small, $\theta < \theta_0$, an increase in reading utility results in stricter shadowbanning as $\partial\hat{x}_2^S(\theta)/\partial v < 0$. Additionally, an increase in the user's cost of being moderated prompts the platform to adopt more lax shadowbanning as $\partial\hat{x}_2^S(\theta)/\partial c > 0$ and $\partial\hat{y}^S(\theta)/\partial c > 0$.

### 5.2. Comparing Strategies

We now investigate the differences between no moderation, content removal, and shadowbanning. Firstly, we compare platform profit under each strategy and explore market coverage. We then examine the stringency of the platform's moderation strategy. We compare shadowbanning with no moderation when $\alpha \geq \sqrt{2v}$ because it is not optimal for a platform using content removal to remove content when $\alpha \geq \sqrt{2v}$. If $\alpha < \sqrt{2v}$, then we compare shadowbanning with content removal.

**Proposition 3.** *The platform profit with shadowbanning is Pareto superior to the platform profit with no moderation or content removal.*

We summarize our results in Table 1. If $\theta \geq \bar{\theta}$, then a platform with shadowbanning behaves like a platform with content removal, and makes equal profit. However, when $\theta < \bar{\theta}$, the platform opts for shadowbanning irrespective of the relative values of posting and reading utilities. From Proposition 2 shadowbanning aids a platform in expanding its user base compared to no moderation, thereby increasing profits. Complementing this, Proposition 3 establishes that the user base and profit of a platform implementing shadowbanning also exceeds that of a platform with content removal.

**Lemma 4.** *If a platform implements shadowbanning and the moderation cost for users is small enough, $c \leq \bar{c}$, then it can cover the market.*

From *Case III* of Lemma 2, the platform covers the market when $\hat{y}^S(\theta) \geq \hat{x}_2^S(\theta)$. Solving for for $c$ from $\hat{y}^S(\theta) = \hat{x}_2^S(\theta)$, $\bar{c}$ is given as $\bar{c} = [2v + \theta - 1 + 2\alpha[1-\theta]\sqrt{2v\theta - \theta + \theta^2}]/2\theta^2$. This Lemma underscores how a covered market depends on our exogenous parameters, especially on the cost of moderation for users, $c$. If the cost is small enough, $c \leq \bar{c}$, then the platform covers the market, otherwise, the platform cannot cover the market. An increase in the reading utility $v$ or the posting utility $\alpha$ broadens the range of $c$ that can lead to a covered market. In other words, a higher value of $c$ requires a corresponding increase in either $\alpha$ or $v$ in order for the platform to cover the market. This Lemma leads us to the following proposition:

**Proposition 4.** *a) If user beliefs that the platform implements shadowbanning are small enough, $\theta < \bar{\theta}$, and the cost of moderation for users is also small enough, $c \leq \bar{c}$, then the platform implementing shadowbanning can cover the market. b) The platform implementing content removal never covers the market.*

The primary objective of the platform is to expand its user base, as this directly impacts profitability. Shadowbanning allows the platform to attract a larger user base by increasing users' expected utility. However, the platform cannot cover the market if the moderation cost for users is high. Although shadowbanning creates uncertainty about content moderation, encouraging participation from extreme users, these users choose not to participate when the moderation cost is high. In contrast, content removal reduces expected disutility for moderate and less-extreme users by excluding extreme users, but this prevents the platform from ever covering the market.

**Proposition 5.** *Shadowbanning results in a larger volume of moderated content compared to content removal, as indicated by $y^{S*}(\theta) \leq y^{R*}$, except when user beliefs that the platform implements shadowbanning are low and they derive little utility from reading or posting content.*

**Table 1 Summary of Results for Shadowbanning with Common User Beliefs**

| | $\theta < \bar{\theta}$ | | $\theta \geq \bar{\theta}$ |
|---|---|---|---|
| | $\theta < \theta_0$ | $\theta \geq \theta_0$ | |
| $\alpha < \sqrt{2v}$ | $X^{S*} = [\hat{x}_1^{S*}(\theta, y), 1]$[i]<br>$y^{S*}(\theta) = \hat{x}_2^S(\theta)$[ii] | $X^{S*} = [0, \hat{y}^S(\theta)] \cup [\hat{x}_2^{S*}(\theta), 1]$<br>$y^{S*}(\theta) = \hat{y}^S(\theta) < y^{R*}$ | $X^{S*} = [0, \sqrt{2v}]$<br>$y^{S*}(\theta) = \sqrt{2v} = y^{R*}$ |
| $\alpha \geq \sqrt{2v}$ | $X^{S*} = [\hat{x}_1^{S*}(\theta, y), 1]$[i]<br>$y^{S*}(\theta) = \hat{x}_2^S(\theta)$[ii] | | $X^{S*} = [\hat{x}_1^{S*}, 1]$<br>$y^{S*}(\theta) = 1 = y^{N*}$ |

[i] If moderation cost for users is small, $c < \bar{c}$, then the platform covers the market with moderation threshold $y^{S*}(\theta) = \hat{y}^S(\theta) < y^{R*}$. As a result, the user base is $X^{S*} = [0, 1]$. For brevity, these conditions are omitted from the table.

[ii] For any values of $\alpha$ and $c$, if $v$ is small, then $\hat{x}_2^S(\theta) > y^{R*}$; conversely, if $v$ is large, then $\hat{x}_2^S(\theta) < y^{R*}$. Similarly, for any values of $v$ and $c$, if $\alpha$ is small, then $\hat{x}_2^S(\theta) > y^{R*}$; whereas if $\alpha$ is large, then $\hat{x}_2^S(\theta) < y^{R*}$. For brevity, we do not include this result in the table.

[−] We explain the process of deriving $\bar{\theta}$ and $\theta_0$ in Lemma 2.

A platform implementing shadowbanning often has stricter moderation than a platform using content removal as the shadowbanning moderation threshold is lower than that of content removal, $y^{S*}(\theta) \leq y^{R*}$. This is because a platform shadowbans content to decrease the disutility users face from reading extreme content, thereby motivating less-extreme users to participate. Thus, stricter shadowbanning helps the platform have a larger user base; however, it also leads to a larger volume of moderated content compared to content removal.

Nonetheless, a platform implementing shadowbanning can have more lax moderation compared to a platform using content removal. This happens when user beliefs that the platform implements shadowbanning are low, $\theta < \theta_0$, and users also have a sufficiently low reading or posting utility. This is because the least extreme users are already excluded because of their beliefs that the platform implements shadowbanning and their low utility from reading or posting content. Thus, there is no need for the platform to have stricter moderation, as there is no need to further mitigate expected reading disutility for expansion of the user base among the less-extreme users.

### 5.3. Welfare Analysis

Consumer surplus, denoted as $CS^K(y^*)$, is the total utility of all participating users, with user utility considered to be measurable in monetary terms, where $K \in N, R, S$ represents different moderation strategies. Social welfare, denoted as $W^K(y^*)$, is defined as the sum of consumer surplus and platform profit. These are expressed as

$$CS^K(y^*) = \int_{X^{K*}} E(U^K(x))\,dx \quad \text{and} \quad W^K(y^*) = CS^K(y^*) + \pi^K(y^*). \tag{12}$$

We compare consumer surplus and social welfare of shadowbanning with that of no moderation when $\alpha \geq \sqrt{2v}$ and that of content removal when $\alpha < \sqrt{2v}$.

**Table 2 Summary of Welfare Analysis Results for Shadowbanning with Common User Beliefs**

| | $\theta < \bar{\theta}$ | | $\theta \geq \bar{\theta}$ |
|---|---|---|---|
| | $\theta < \theta_1$ | $\theta \geq \theta_1$ | |
| $\alpha < \sqrt{2v}$ | $\pi^{S*}(y) > \pi^{R*}(y)$<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ | $\pi^{S*}(y) > \pi^{R*}(y)$<br>$CS^{S*}(y) > CS^{R*}(y)$<br>$W^{S*}(y) > W^{R*}(y)$ | $\pi^{S*}(y) = \pi^{R*}(y)$<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ |
| $\alpha \geq \sqrt{2v}$ | $\pi^{S*}(y) > \pi^{N*}(y)$<br>$CS^{S*}(y) < CS^{N*}(y)$<br>$W^{S*}(y) < W^{N*}(y)$ | $\pi^{S*}(y) > \pi^{N*}(y)$<br>$CS^{S*}(y) < CS^{N*}(y)$<br>$W^{S*}(y) > W^{N*}(y)$ | $\pi^{S*}(y) = \pi^{N*}(y)$<br>$CS^{S*}(y) = CS^{N*}(y)$<br>$W^{S*}(y) = W^{N*}(y)$ |

– In Lemma 2, we explain the process of deriving $\bar{\theta}$.
– We determine $\theta_1$ by solving for $\theta$ from the equation $CS^{S*}(y) = CS^{R*}(y)$. Similarly, solving for $\theta$ from $W^{S*}(y) = W^{N*}(y)$ and $W^{S*}(y) = W^{R*}(y)$ yields different values of $\theta$. For brevity and readability, we assume these values are identical and denote them collectively as $\theta_1$.
– It's important to note that $\theta_1$ serves a different purpose than $\theta_0$ in the Table 1.

**Proposition 6.** *a) If users have moderate beliefs that the platform implements shadowbanning, then shadowbanning yields a higher consumer surplus compared to content removal. Otherwise, shadowbanning results in a lower consumer surplus than content removal. b) Shadowbanning reduces consumer surplus compared to no moderation.*

The results of our welfare analysis are presented in Table 2. To understand Proposition 6, we first compare the user base of a platform implementing shadowbanning, $[\hat{x}_1^{S*}(\theta, y), 1]$, with that of a platform with no moderation, $[\hat{x}^{N*}(1), 1]$. Although $\hat{x}^{N*}(1) > \hat{x}_1^{S*}(\theta, y)$, indicating that the shadowbanning user base is larger than that of no moderation, consumer surplus does not follow the same relation. This is because, with shadowbanning, more-extreme users gain lower expected utility due to uncertainty, compared to the utility they receive on a platform with no moderation. Furthermore, the additional consumer surplus generated by less-extreme users encouraged to participate with shadowbanning does not offset the loss in consumer surplus for more-extreme users, especially when $\alpha \geq \sqrt{2v}$. A similar dynamic occurs when the platform chooses shadowbanning over content removal. The uncertainty imposed on user utility leads to a lower consumer surplus unless user belief is moderate.

**Proposition 7.** *If user beliefs that the platform implements shadowbanning are moderate, then social welfare with shadowbanning is greater than with content removal or no moderation.*

Proposition 7 implies that when user beliefs are either low or high, no moderation and content removal result in higher social welfare compared to shadowbanning. However, when user beliefs are moderate, shadowbanning increases social welfare, even though it may reduce consumer surplus. Because the increase in platform profit from shadowbanning is sufficient to offset the loss in consumer surplus, it results in higher social welfare compared to no moderation or content removal.

## 6. Shadowbanning Strategy ($S$): Heterogeneous Beliefs

We now consider the impact of varying user beliefs that the platform implements shadowbanning by using $\beta(x)$ for heterogeneous user beliefs instead of $\theta$ for common user beliefs while keeping other assumptions and model settings unchanged. Specifically, we follow Assumption 7b instead of Assumption 7a. Detailed analysis is provided in Appendix B. We briefly discuss the similarities and differences between the results of our shadowbanning models with common and heterogeneous beliefs. In subsequent sections, the superscript $S$ continues to denote shadowbanning. The arguments $\theta$ and $\beta$ distinguish between shadowbanning with common and heterogeneous user beliefs.

### 6.1. User Segmentation

User segmentation in the two models—common and heterogeneous beliefs—is similar, except when the user belief is high. With common user beliefs, there is one indifferent user for the more-extreme users, denoted by $\hat{x}_2^S(\theta)$, and consequently users within the range $[\hat{x}_2^S(\theta), 1]$ participate on

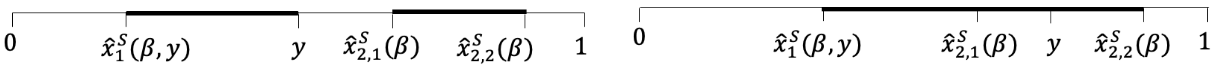


**Figure 4** Illustration of the Platform's User Base - Shadowbanning with Heterogeneous User Beliefs

the platform. Conversely, when users have heterogeneous beliefs, there are two indifferent users, $\hat{x}_{2,1}^S(\beta)$ and $\hat{x}_{2,2}^S(\beta)$ where users in the range $[\hat{x}_{2,1}^S(\beta), \hat{x}_{2,2}^S(\beta)]$ may participate on the platform. If $\hat{x}_{2,2}^S(\beta) \geq 1$, then users in range $[\hat{x}_{2,1}^S(\beta), 1]$ participate on the platform which is similar to user participation with common user beliefs. But if $\hat{x}_{2,2}^S(\beta) < 1$, then user participation between these two models of user beliefs differs. Figure 4 clarifies this contrast when compared with Figure 2. When $\beta$ is sufficiently large, heterogeneous user beliefs substantially increase as $x$ increases, thus, extreme users with a high degree of extremeness approaching 1 do not participate on the platform. For these users the potential disutility they may incur if their content is shadowbanned becomes substantial enough to deter their participation.

Table 3 summarizes our results with heterogeneous user beliefs in a manner similar to Table 1 for common user beliefs. A comparison of the two reveals that Table 3 adds one more column and a slight shift in the user beliefs ranges. The rightmost column in Table 3 represents user participation conditions unique to the analysis with heterogeneous user beliefs.

## 6.2. The Platform's Moderation Decisions

Our results concerning the optimal moderation threshold with heterogeneous user beliefs mirror those with common user beliefs. Specifically, the optimal shadowbanning threshold belongs to the

**Table 3** Summary of Results for Shadowbanning Strategy with Heterogeneous User Beliefs

| | $\beta < min\{\hat{\beta}, \bar{\beta}\}$[i] | $min\{\hat{\beta}, \bar{\beta}\} < \beta < max\{\hat{\beta}, \bar{\beta}\}$ | | $max\{\hat{\beta}, \bar{\beta}\} < \beta$ |
|---|---|---|---|---|
| | | Small c | Large c | |
| $\alpha < \sqrt{2v}$ | $X^{S*} = [\hat{x}_1^{S*}(\beta, y), 1]$[ii]<br>$\pi^{S*}(y) > \pi^{R*}(y)$[iii]<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ | $X^{S*} = [0, \hat{y}^S(\beta)] \cup [\hat{x}_{2,2}^{S*}(\beta), 1]$<br>$\pi^{S*}(y) > \pi^{R*}(y)$<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ | $X^{S*} = [0, \sqrt{2v}]$<br>$\pi^{S*}(y) = \pi^{R*}(y)$<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ | $X^{S*} = [0, \hat{y}^S(\beta)] \cup [\hat{x}_{2,1}^{S*}(\beta), \hat{x}_{2,2}^{S*}(\beta)]$<br>$\pi^{S*}(y) < \pi^{R*}(y)$<br>$CS^{S*}(y) < CS^{R*}(y)$<br>$W^{S*}(y) < W^{R*}(y)$ |
| $\alpha \geq \sqrt{2v}$ | $X^{S*} = [\hat{x}_1^{S*}(\beta, y), 1]$[ii]<br>$\pi^{S*}(y) > \pi^{N*}(y)$<br>$CS^{S*}(y) < CS^{N*}(y)$<br>$W^{S*}(y) < W^{N*}(y)$ | $X^{S*} = [\hat{x}_1^{S*}(\beta, y), 1]$[ii]<br>$\pi^{S*}(y) > \pi^{N*}(y)$<br>$CS^{S*}(y) < CS^{N*}(y)$<br>$W^{S*}(y) < W^{N*}(y)$ | $X^{S*} = [\hat{x}_1^{S*}(\beta, y), 1]$<br>$\pi^{S*}(y) = \pi^{N*}(y)$<br>$CS^{S*}(y) = CS^{N*}(y)$<br>$W^{S*}(y) = W^{N*}(y)$ | $X^{S*} = [0, \hat{x}_{2,2}^{S*}(\beta)]$<br>$\pi^{S*}(y) < \pi^{N*}(y)$<br>$CS^{S*}(y) < CS^{N*}(y)$<br>$W^{S*}(y) < W^{N*}(y)$ |

[i] In the proof of Lemma **??** in the appendix, we explain the derivation of $\hat{\beta}$ and $\bar{\beta}$, which are used to determine the conditions under which some or none of the more-extreme users participate on the platform.

[ii] The platform can cover the market under some conditions. That is $\hat{x}_1^{S*}(\beta, y) = 0$. However, the results in the table remain unchanged.

[iii] If $\beta < min\{\hat{\beta}, \bar{\beta}\}$ and both reading and posting utility are very small, then the profit of a platform implementing shadowbanning can be lower than that of content removal, $\pi^{S*}(y) < \pi^{R*}(y)$. For brevity, we do not include this result in the table.

set $\{\hat{y}^S(\beta), \hat{x}^S_{2,1}(\beta), \sqrt{2v}\}$ which is similar to the one with common user beliefs. However, there are slight variations in the outcomes relating to the platform profit. Comparing Tables 1 and 3 demonstrate that if heterogeneous user beliefs are sufficiently high, $\beta > max\{\hat{\beta}, \bar{\beta}\}$, then users within the range $[\hat{x}^S_{2,1}(\beta), 1]$ abstain from participation and the profit of a platform implementing shadowbanning is lower than that of the platform implementing content removal or no moderation, and the platform chooses not to implement shadowbanning. In contrast, with sufficiently high common user beliefs, the profit of the platform implementing shadowbanning equals that of no moderation and content removal, and the platform is indifferent between content moderation strategies. With heterogeneous user beliefs, the platform can also be indifferent between content moderation strategies. This occurs when users hold moderate beliefs that the platform implements shadowbanning, $min\{\hat{\beta}, \bar{\beta}\} < \beta < max\{\hat{\beta}, \bar{\beta}\}$, and the cost of shadowbanning for users, $c$, is sufficiently high. These results demonstrate that extreme user participation and the absence of high user beliefs are crucial for the profitability of platforms employing shadowbanning.

### 6.3. Welfare Analysis

Results regarding social welfare and consumer surplus differ between common and heterogeneous user beliefs. With heterogeneous user beliefs, consumer surplus and social welfare with a platform that implements shadowbanning never exceed that of no moderation and content removal (see Table 3). As we saw earlier, this is not the case for common user beliefs. If common user beliefs about shadowbanning implementation are moderate, then social welfare is higher than that of content removal and no moderation. Additionally, if user posting utility is sufficiently lower than reading utility and common beliefs are moderate, then the consumer surplus from shadowbanning with common user beliefs is higher than that of content removal (see Table 2).

In sum, the comparison between shadowbanning with common user beliefs and with heterogeneous user beliefs reveals key findings. User segmentation is largely similar in both models, except for users with a high degree of extremeness approaching 1. With common user beliefs, these extreme users always participate, whereas with heterogeneous beliefs, they may opt out. When these extreme users do not participate under heterogeneous beliefs, the platform earns lower profits from shadowbanning than from content removal or no moderation. In contrast, with common user beliefs, shadowbanning consistently yields equal or higher profits compared to content removal or no moderation. Additionally, with heterogeneous beliefs, both consumer surplus and social welfare are consistently lower with shadowbanning than with content removal or no moderation. However, with common user beliefs, shadowbanning can achieve higher consumer surplus and social welfare than the other strategies under specific conditions.

# 7. Shadowbanning with Imperfect Technology

## 7.1. The Platform's Moderation Decisions

We now relax Assumption 5 and expand our examination to explore the implications of imperfect technology. Following the approach outlined by Liu et al. (2022), we consider two types of errors in the 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. All other assumptions and settings from our main model (shadowbanning with common user beliefs) remain unchanged. For brevity, we focus on the main results; details of the analysis are provided in Appendix C.

**Proposition 8.** *Regardless of user beliefs, a necessary condition for the platform to implement shadowbanning is that the technology is sufficiently accurate.*

From Proposition 8, a platform's decision to implement shadowbanning with imperfect technology depends on the accuracy of the technology, and user beliefs no longer influence this decision. Specifically, if the technology is sufficiently accurate, then the platform may implement shadowbanning regardless of user beliefs. This is because shadowbanning with sufficient accuracy yields greater benefits by retaining less extreme and moderate users, outweighing the loss of extreme users and expanding the overall user base—independent of user beliefs. However, user beliefs still shape the composition of the user base: when beliefs about shadowbanning are high, less extreme users are more likely to participate, whereas low beliefs tend to attract more extreme users. Our finding aligns with Liu et al. (2022), which emphasizes the necessity of sufficiently accurate technology for moderation via content removal. Nonetheless, this result contrasts with our earlier analysis (Section 5.1), where user beliefs affect the platform's shadowbanning decision.

## 7.2. Better Technology and Less Content Moderation.

Our findings show that the platform implements less strict shadowbanning as accuracy increases. This result aligns with that of Liu et al. (2022), which demonstrates that when accuracy is sufficiently high, platforms adopt a more relaxed content removal. In addition, we investigate the impact of an increase in user beliefs on the moderation threshold when the technology is imperfect. Our results show that when the technology is imperfect, platforms implement less strict shadowbanning as user beliefs increase. This aligns with the results of our main model, where a larger $\theta$ leads to a higher optimal shadowbanning threshold, as indicated by the positive comparative statics: $\partial \hat{y}^{S*}(\theta)/\partial\theta > 0$. Details of the analysis are provided in Appendix C.

## 7.3. Incentive for Imperfect Content Moderation Technology.

A platform using content removal "*may choose imperfect technology even if there is no cost to improving it when the cost to users subject to moderation (c) is small*" (Liu et al. 2022, p. 12).

This is because, with content removal, a less accurate technology benefits extreme users more than it harms moderate ones. Therefore, if $c$ is small, then a less accurate technology increases the number of extreme users more than it reduces moderate users. However, our findings regarding shadowbanning differ. Specifically, we find that if the cost of moderation for users, $c$, is small, then the platform always chooses perfect technology. In contrast, if the cost of moderation for users is large, then the platform has an incentive to maintain imperfect technology. This is because, with a large cost of moderation, extreme users have already lost their incentive to participate. Higher accuracy decreases the number of extreme users more than it increases the number of moderate users. Therefore, the platform's most effective strategy for expanding its user base is not to increase moderate user participation, but rather to retain and expand extreme users by maintaining lower accuracy. Details of the analysis are provided in Appendix C.

## 8. Shadowbanning with Mandated Content Removal Policy

### 8.1. The Platform's Moderation Decisions

Here we assume that a policymaker defines a moderation threshold policy (see the discussion of the assumption in Appendix D.3), denoted by $y_r \in [0,1]$, which requires the platform to remove content with extremeness degree greater than $y_r$. The platform may also implement shadowbanning by choosing a threshold $y_s \in \mathbb{R}^+$ where $y_s \leq y_r$. The platform shadowbans content when its extremeness falls within the range $y_s \leq x < y_r$. Due to policy constraints on content removal, the platform's only decision is the shadowbanning threshold. Users are aware of the content removal threshold, $y_r$, but are unaware of the shadowbanning threshold, $y_s$. All other assumptions and settings from our main model (shadowbanning with common user beliefs) remain unchanged. The detailed analysis is provided in the Appendix D, and here we focus on the main results.

**Proposition 9.** *If the policymaker enforces a strict content removal policy (small $y_r$) and user posting utility is relatively greater than their reading utility ($\alpha > \sqrt{2v}$), then a platform does not implement shadowbanning; otherwise, it does.*

This result suggests that when users prioritize posting over reading, a strict content removal policy deters the platform from implementing shadowbanning, as shadowbanning would not expand the user base. The strict removal policy already substantially reduces user participation, and shadowbanning would further exacerbate this reduction. However, under a lax removal policy, the platform implements shadowbanning to attract more users.

### 8.2. User Segmentation and Shadowbanning Threshold

The user segmentation (illustrated in Figure D.2) is similar to the user segmentation when the platform implements shadowbanning with heterogeneous beliefs (Figure 4) as heterogeneous beliefs discourage extreme users from participating, akin to the effects of the removal policy. Furthermore, the impacts of changes in $y_r$ on the platform's optimal shadowbanning threshold and user segmentation depends on user beliefs. If user beliefs are low (refer to Appendix D.2), then the user segmentation is continuous. An increase in $y_r$ shifts the user segmentation rightward, while a decrease shifts it leftward. Thus, if policymakers adopt a more lax (stricter) content removal policy, then the platform responds with more lax (stricter) shadowbanning.

If user beliefs are high, then the user segmentation is not continuous and some moderate users do not participate on the platform (Figure D2a). If policymakers adopt a more lax content removal policy (decrease in $y_r$), then fewer moderate users participate. In other words, under this lax content removal policy, the platform opts for stricter shadowbanning, which reduces participation of moderate users. Specifically, the platform shadowbans both moderate and extreme users (whose content is not 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 the strict content removal policy.

We also find that, given a content removal threshold mandated by policymakers, if user beliefs increase (larger $\theta$), then the platform chooses a higher shadowbanning threshold (implementing more lax shadowbanning), which results in reduced shadowbanned content. This result is in line with the findings of our main model. We also find that the joint impact of higher user beliefs (larger $\theta$) and stricter content removal policy (smaller $y_r$) on the shadowbanning threshold depends on user beliefs. When user beliefs are high, they reinforce each other, leading the platform to choose more lax shadowbanning. However, when user beliefs are low, these factors have opposing effects, and the net impact on the shadowbanning threshold is unclear.

## 9. Shadowbanning with Updated Beliefs

We now consider a two-period game where users shadowbanned in the first period increase their common beliefs that they will be shadowbanned again, but not to certainty. Specifically, shadowbanned users increase their beliefs by a small amount, $\epsilon \in [0,1]$, such that their beliefs in the second period becomes $\theta + \epsilon < 1$. For users that were not shadowbanned in the first period, their beliefs remain unchanged at $\theta$. All other assumptions and settings from the main model (shadowbanning with common user beliefs) remain unchanged. The detailed analysis is provided in Appendix E; here, we focus on the main results.

We find that in this dynamic model where some users update their beliefs our results are similar to our static model where users do not update their beliefs. First, as in our static model, the platform implements shadowbanning when user beliefs are not high. Regardless of the increase in the beliefs of shadowbanned users ($\epsilon$), there are conditions under which our dynamic and static models yield identical outcomes in terms of profitability, shadowbanning thresholds, and user participation. Second, user segmentation in both models is similar, as illustrated in Figure 2. Finally, under certain conditions, the platform covers the market in both periods of our dynamic model, similar to our static model.

There are minor differences between the results of our dynamic and static models. First, although in our static model the platform does not implement shadowbanning when user beliefs are high, in our dynamic model it may also refrain from shadowbanning even with moderate or low initial beliefs. This is because of the increase in shadowbanned users' beliefs, $\epsilon$. When $\epsilon$ increases, it reduces the threshold belief above which the platform may choose not to implement shadowbanning.

In addition, in our dynamic model the platform always selects a shadowbanning threshold greater than or equal to that 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 number of users that update their beliefs in the second period. b) By encouraging participation from some moderate users in the first period. These users retain their initial beliefs and are less likely to withdraw in the second period. But, in our static model these moderate users do not participate due to the lower shadowbanning threshold (stricter shadowbanning). These results are intuitive because they reflect the platform's need to balance content moderation with user participation. In our dynamic model, a higher shadowbanning threshold encourages moderate users to join and prevents moderation of extreme users early, reducing the risk of a decline in participation over time.

The platform's long-term profit in our dynamic model may not always exceed that of our static model. This is because, although a higher shadowbanning threshold encourages participation from moderate users, it may 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 belief, $\theta$, and the increase in their belief, $\epsilon$.

Table 4 summarizes the comparison of equilibrium outcomes across five models examined in our study.

## 10. Conclusion

We develop an economic foundation that allows us to compare content removal and shadowbanning to determine their relative impacts on platform profit, consumer surplus, and social welfare. Using a stylized formulation taken in part from the literature, we find that a platform implements shadowbanning when common user beliefs that the platform shadowbans extreme content are not high. We observe that, similar to content removal, shadowbanning contributes to the platform's user

**Table 4 Summary of Comparisons Across the Five Models in the Study**

| | Shadowbanning with common user beliefs | Shadowbanning with heterogeneous user beliefs | Shadowbanning with imperfect technology | Shadowbanning with mandated content removal policy | Shadowbanning with updated beliefs |
|---|---|---|---|---|---|
| Two types of user segmentation (contiguous or disjoint) | Yes | Yes | Yes | Yes | Yes |
| Can users with any degree of extremeness participate? | Yes | No | Yes | No | Yes |
| Can the platform cover the market? | Yes | Yes | Yes | No | Yes |
| Does the platform always shadowban when the user beliefs are small enough? | Yes | Yes | No | No | No |
| Can the platform profit be greater or equal to that of content removal and no moderation? | Yes | Yes | Yes | Yes | Yes |
| If user beliefs increase, then does the platform implement more lax shadowbanning? | Yes | Yes | Yes | Yes | Yes |
| Can shadowbanning threshold be lower than content removal threshold? | Yes | Yes | Yes | Yes | Yes |

base expansion as without content moderation, less-extreme users are deterred from participating due to the disutility associated with reading extreme content.

The key difference between content removal and shadowbanning is user segmentation. Content removal facilitates the participation of less-extreme users by excluding the extreme users. In contrast, shadowbanning can foster the participation of users with any degree of extremeness. When implementing shadowbanning, the platform mostly opts for a lower moderation threshold, resulting in stricter moderation compared to content removal. We also show that shadowbanning does not necessarily enhance consumer surplus or social welfare. We further demonstrate that the main results remain consistent even when accounting for users with heterogeneous beliefs about the platform implementing shadowbanning. Finally, the results from our extended models—including shadowbanning with imperfect technology, shadowbanning with mandated content removal policy, and shadowbanning with updated beliefs—offer further insights about shadowbanning.

Our findings have several policy implications. First, the platform has higher profits when user beliefs that shadowbanning is implemented are low or moderate. Thus, policies that require the platform to disclose its use of shadowbanning can lead users to form higher beliefs that shadowbanning is being implemented. This may not be in the platform's best interest. Second, although shadowbanning can increase platform profit, market coverage, and even social welfare, there are some settings where it may not enhance consumer surplus compared to content removal or no moderation. This highlights a misalignment between platform incentives and user welfare.

Third, recent regulatory frameworks have established transparency requirements, mandating platforms to publish moderation rules, notify users of actions, provide explanations, and enable appeals. Although transparency requirements in content moderation can be beneficial, these requirements fundamentally conflict with shadowbanning, which operates through opacity and is what distinguishes it from other content moderation strategies. Thus, shadowbanning is incompatible with transparency regulations. As Leerssen (2023) highlights, regulatory frameworks, such as those under the Digital Services Act (DSA), prohibit platforms from implementing shadowbanning. However, there is insufficient robust evidence to fully understand the consequences of banning shadowbanning. We find that shadowbanning can enhance both consumer surplus and platform profit. Therefore, rather than mandating transparency, policymakers could focus on defining appropriate use cases (e.g., addressing extreme content from bots and spam resources) and developing effective monitoring frameworks that preserve the benefits of shadowbanning without undermining its efficacy.

Finally, we find that a platform's decision to implement shadowbanning with imperfect technology is not influenced by user beliefs. This result highlights the need for policymakers to consider measures beyond transparency—such as accuracy, fairness, and bias mitigation. Policymakers could introduce mechanisms that enforce minimum accuracy standards for moderation technologies, including periodic third-party audits, while recognizing that privacy concerns may limit how much platforms can disclose about their moderation decisions. For example, sharing original user-generated content for external review could infringe on user privacy rights, creating additional challenges.

From a managerial perspective, shadowbanning allows platforms to retain extreme users by allowing their continued participation while limiting the visibility of their potentially harmful content. Though beneficial short-term, its effectiveness may decline as user beliefs evolve. This implies that shadowbanning might be most effective when used strategically, such as for managing temporary surges in harmful content or targeting automated behaviors. However, potential misuse—including suppressing marginalized voices or enabling censorship and polarization—requires platforms to develop guidelines and regularly reassess their moderation strategies.

Future research can extend our analysis by considering the platform's cost of moderation and when users actively exert effort to determine whether a platform employs shadowbanning. A limitation of our work is that we assume the same users' cost of being moderated, $c$, for both content removal and shadowbanning strategies. In practice, user costs may differ between content removal and shadowbanning, with platforms potentially preferring shadowbanning when user costs are lower. Another limitation of our work is that we employ a specific form for our user utility function. Thus, we show certain results can occur, but have not shown that they can occur across the range of possible utility functions. Future research could explore how a lack of user awareness regarding shadowbanning may distort feedback signals about content attractiveness. This may result in emotional harm, such as feelings of isolation, distrust, and frustration (Nicholas 2022). In addition, it would be valuable to examine how competition among platforms or user multihoming affects our results such as the platform's moderation decision and user participation.

## Acknowledgments

We express our gratitude to Dr. Cristobal Cheyre for insightful feedback and invaluable contributions as the discussant of our research at Workshop on Information Systems and Economics (WISE) 2023. We also thank our colleagues at the University of Calgary for helpful suggestions, Jeanette Burman for excellent editing assistance, and Ashkan Eshghi for his generous and invaluable help.

## References

Andres R, Slivko O (2021) Combating online hate speech: The impact of legislation on twitter. Technical report, ZEW Discussion Papers.

Apodaca T, Uzcátegui-Liggett N (2024) How we investigated shadowbanning on instagram. *The Markup* URL `https://themarkup.org/automated-censorship/2024/02/25/how-we-investigated-shadowbanning-on-instagram`, accessed: 2025-05-02.

Bauer K, Gill A (2024) Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. *Information Systems Research* 35(1):226–248.

Beknazar-Yuzbashev G, Jiménez Durán R, McCrosky J, Stalinski M (2022) Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346* .

Beknazar-Yuzbashev G, Jiménez Durán R, Stalinski M (2024) A model of harmful yet engaging content on social media. *Available at SSRN* .

Brown S (2021) Social media is broken. a new report offers 25 ways to fix it. URL `https://mitsloan.mit.edu/ideas-made-to-matter/social-media-broken-a-new-report-offers-25-ways-to-fix-it`.

Chen YS, Zaman T (2024) Shaping opinions in social networks with shadow banning. *Plos one* 19(3):e0299977.

CIGI-IPSOS (2019) Cigi-ipsos global survey, internet security  trust. URL `https://www.cigionline.org/sites/default/files/documents/2019CIGI-IpsosGlobalSurveyPart3-SocialMedia,FakeNews,Algorithms.pdf`, retrieved July 23, 2023.

Constine J (2019) Instagram now demotes vaguely "inappropriate" content. URL `https://techcrunch.com/2019/04/10/instagram-borderline/`.

Cook J (2020) Instagram's ceo says shadow banning "is not a thing." that's not true. URL `https://www.huffpost.com/entry/instagram-shadow-banning-is-real_n_5e555175c5b63b9c9ce434b0`.

Cotter K (2023) "Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society* 26(6):1226–1243.

Delmonaco D, Mayworm S, Thach H, Guberman J, Augusta A, Haimson OL (2024) “What are you doing, TikTok?”: How marginalized social media users perceive, theorize, and “prove” shadowbanning. *Proceedings of the ACM on Human-Computer Interaction* 8(CSCW1):1–39.

Dixon S (2023) Daily time spent on social networking by users worldwide from 2012 to 2023. URL `https://www.statista.com/statistics/433871/daily-social-media-usage/`, retrieved July, 2023.

Dstribute (2022) How to prevent ghost posting on craigslist? URL `https://dstribute.io/craigslist/how-to-prevent-ghost-posting-on-craigslist/`, accessed: 2025-05-21.

Dwork C, Hays C, Kleinberg J, Raghavan M (2024) Content moderation and the formation of online communities: A theoretical framework. *Proceedings of the ACM Web Conference 2024*, 1307–1317.

European Commission (2023) Digital services act transparency database. `https://transparency.dsa.ec.europa.eu/`, accessed: 2024-12-18.

Fagan F (2020) Optimal social media content moderation and platform immunities. *European Journal of Law and Economics* 50(3):437–449.

Fowler GA (2022) Shadowbanning is real: Here’s how you end up muted by social media. URL `https://www.washingtonpost.com/technology/2022/12/27/shadowban/`, accessed: 2025-04-12.

Gillespie T (2022) Reduction / Borderline Content / Shadowbanning. *Yale Journal of Law  Technology* .

Gopal RD, Hojati A, Patterson RA (2025) A little bit goes a long way: Indirect effects of content moderation on online social media. *International Journal of Electronic Commerce* 29(1):39–64.

He Q, Hong Y, Raghu T (2024) Platform governance with algorithm-based content moderation: An empirical study on reddit. *Information Systems Research* .

Huber TL, Kude T, Dibbern J (2017) Governance practices in platform ecosystems: Navigating tensions between cocreated value and governance costs. *Information Systems Research* 28(3):563–584.

Jhaver S, Bruckman A, Gilbert E (2019) Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–27.

Jiménez-Durán R (2023) The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *George J. Stigler Center for the Study of the Economy & the State Working Paper* (324).

Kominers SD, Shapiro JM (2024) Content moderation with opaque policies. Technical report, National Bureau of Economic Research.

Leerssen P (2023) An end to shadow banning? transparency rights in the digital services act between content moderation and curation. *Computer Law & Security Review* 48:105790.

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

Madio L, Quinn M (2024) Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy* .

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

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

Möhlmann M, Zalmanson L, Henfridsson O, Gregory RW (2021) Algorithmic management of work on online labor platforms: When matching meets control. *MIS quarterly* 45(4).

Mostagir M, Siderius J (2023) When should platforms break echo chambers? Technical report, University of Michigan and Massachusetts Institute of Technology, Working paper.

Müller K, Schwarz C (2023) The effects of online content moderation: Evidence from president Trump's account deletion. *Available at SSRN 4296306* .

Nan G, Ding N, Li G, Li Z, Li D (2023) Two-tier regulation models for the user-generated content platform: A game theoretic analysis. *Decision Support Systems* 175:114034.

Nicholas G (2022) Shedding Light on Shadowbanning URL `http://dx.doi.org/10.31219/osf.io/xcz2t`.

Oremus W (2019) Twitter admits it was hiding some people's tweets by mistake — again. URL `https://onezero.medium.com/twitter-admits-it-was-hiding-some-peoples-tweets-by-mistake-again`, retrieved December 01, 2023.

Pasquale F (2015) *The black box society: The secret algorithms that control money and information* (Harvard University Press).

Schwarz C, Jiménez Durán R, Müller K (2022) The effect of content moderation on online and offline hate: Evidence from germany's netzdg. Technical report, CEPR Discussion Papers.

Tiwana A, Konsynski B, Bush AA (2010) Research commentary—platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information systems research* 21(4):675–687.

Walsh R (2006) Micro-isv: From vision to reality. 183 (Apress), ISBN 9781590596012, URL `https://www.oreilly.com/library/view/micro-isv-from-vision/9781590596012`.

Wu Y (2023) Creation, consumption, and control of sensitive content. *Marketing Science* .

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

Zhao L (2024) Algorithmic camouflage: Exploring the shadowbans imposed by algorithms to moderate the content of chinese gay men. *Big Data & Society* 11(4):20539517241296037.