Decentralizing Content Moderation *

Dongkyu Chang†
City University of Hong Kong

Adrian Segura ‡
Universitat Pompeu Fabra

Pengfei Zhang§
University of Texas at Dallas

January 30, 2024

Abstract

In this paper, we apply bargaining game and mechanism design to the study of platform content moderation. We start by analyzing and comparing four popular moderation procedures: centralization, flagging, counter-notice, and voting. Among the four, counter-notice minimizes the over-removal error, voting minimizes the under-removal error, and centralization minimizes the procedural cost. We generalize the welfare comparison to a mechanism design setup where the platform specifies the assignment and cost rules. The error-free mechanism that implements the efficient assignment rule is a repeated counter-notice procedure similar to Wikipedia's Talk Page. The consumer optimal mechanism that trades off errors with costs is instead centralized platform adjudication, which is common in spam filters.

Keywords: Content Moderation, Digital Platforms, Dispute resolution, Mechanism Design.

*This paper has greatly benefited from the comments received from XXX. The usual disclaimers apply.

†Department of Economics and Finance, 9-200, 9/F, Lau Ming Wai Academic Building (LAU) Tat Chee Avenue, Kowloon Tong, Kowloon, Hong Kong. Email: donchang@cityu.edu.hk.

‡Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Barcelona, Spain. Email: adrian.segura@upf.edu.

§*Corresponding author.* School of Economic, Political and Policy Sciences, 800 W. Campbell Road, Mailstop: GR31, Richardson, Texas, US. Email: pengfei.zhang@utdallas.edu.

JEL Codes: D74, K13, K24, L51.

1 Introduction

The Internet undeniably impacts our everyday lives, offering an unprecedented amount and variety of online content that transforms the way of doing business, educating, or communicating. Massive amounts of text, images, and videos are posted on digital platforms on a daily basis. While the proliferation of internet content has brought a myriad of benefits, it also carries the potential for significant harm. One of the most concerning aspects is the ease with which false information and misinformation can spread, leading to the distortion of facts and the manipulation of public perception. The internet also provides a venue for hate speech and harassment, creating an environment where individuals and communities can be targeted and emotionally influenced. The Internet's vastness further makes it a breeding ground for illegal activities such as cybercrime, identity theft, and the distribution of harmful content. As such, how to address and mitigate these potential damages to ensure a safer and more responsible online environment becomes a primary policy concern across the world.

Content moderation - the governance mechanism that removes inappropriate and controversial content - is increasingly adopted by digital platforms to control the flow of information online. For example, just in the third quarter of 2021, Facebook reported taking action on 22.1 million pieces of hate speech content, 4.4 million pieces of violent and graphic content, and 2.2 million pieces of terrorist propaganda. Likewise, YouTube reported removing 15.8 million videos for violating its community guidelines, and over 34 million comments were removed for violating policies in the first half of 2021. During the 2016 US presidential campaign, Twitter's selective restriction on political speech puts content moderation once again under the stage light. [1]

[1]The Washington Post, September 19, 2022. Trump may wish he hadn't pushed to block online content moderation. (https://www.washingtonpost.com/politics/2022/09/19/

Figure 1: Tweet exchange between the European Commissioner for the Internal Market and Elon Musk the day after Twitter's acquisition deal was closed.

A central question in content moderation is how to design the procedures for moderating content. There are several options in the moderation design space. Some platforms recruit human moderators to do manual reviews.[2] Some platforms opt for automated moderation that relies on algorithms to match a pattern of abusive content. Conventionally, moderation is centralized and only carried out by one powerful moderator. Moderation has become increasingly decentralized today such that it is carried out by many dispersed moderators making local decisions. In fact, some platforms, including Wikipedia, completely delegate the task to their user community and try to reach a consensus with the community. Underlying all the different designs is the question of how to reduce errors while keeping costs low. In other words, which procedure should the government and the platform recommend?

This paper studies the optimal procedures used by online platforms to moderate content. We analyze content moderation from a game theoretic perspective. We start by comparing four popular moderation procedures: centralization, flagging, notice and counter-notice, and voting. Under centralization, the platform makes the moderation decision. The other three procedures all delegate the moderation decision to the users to some extent, with different degrees and forms of user participation. The flagging procedure gives the plaintiff an opportunity to send a removal notice, and it is the simplest among the three. The counter-notice procedure allows the plaintiff and the defendant to

trump-social-media-republicans-texas-florida/).

[2]TikTok, for instance, hired more than 10,000 moderators per year.

counter each other's notices. The voting procedure invites all the users to enact a moderation decision collectively. We model the procedures as games between players who are harmed by the content (which we call the plaintiffs) and players who are causing harm (which we call the defendants). If the content remains, the defendant gets the value of the content while the plaintiff pays for the harm. If the content gets removed, both parties' payoffs are normalized to zero. Thus, the plaintiff gains from moderation whereas the defendant loses from it. In each game, the content is removed with some probability at certain costs, given different players' profiles. The content moderation outcome is then described by a pair of functions specifying the removal probability and the expected procedural cost. Using this model, we compare the social efficiency of the three procedures.

Each procedure has its own advantage in either reducing error or reducing cost. If the primary concern centers around mitigating the welfare loss from over-removal bias, the counter-notice procedure is the most efficient option of the three. If the primary concern is instead the welfare loss from under-removal bias, the voting procedure is the most efficient option. Centralization, however, proves to be the cost-minimizing procedure of the three.

Importantly, decentralization in the form of user participation increases efficiency considerably. The rationale lies in the ability of user participation to resolve disputes preemptively before the disputes escalate to the platform ruling. Both counter-notice and voting increase users' participation by extending the bargaining periods or involving more users in the bargaining. Users' participation aligns the moderation outcome with the direction of efficiency. For the subset of users (h, v) whose outcome is determined by the platform decision, decentralization improves the probability that they are going to get a rightful moderation outcome. For the mass of users (h, v) who are marginal in notice decisions, decentralization flips their moderation outcome from either a type I error or a type II error to an efficient one. Decentralization, however, might not maximize consumers' welfare as it imposes substantial procedural costs on users.

The comparison of the four procedures appears to be generalizable to more sophisticated procedures. We thus turn to the mechanism design approach to find the optimal procedure among all possible ones. In our mechanism design setup, harm and value of

the content are private information of the plaintiff and the defendant respectively. The platform commits to an assignment rule that specifies the probability function of removal and a cost rule that specifies the plaintiff's and the defendant's procedural costs. A moderation mechanism is feasible if it satisfies both incentive compatibility for truth-telling and individual rationality. We define two notions of social optimality. The first one is an error-free mechanism that implements the efficient allocation rule (i.e., zero over-removal error and zero under-removal error). The second one is a consumer-optimal mechanism that maximizes the sum of the expected payoffs of the plaintiff and the defendant. We solve for the direct-relevation mechanisms and show the extensive form game that decentralizes them.

The mechanism design approach turns out to be fruitful. The error-free procedure that implements the efficient allocation rule is a repeated counter-notice procedure, where the two parties take turns to respond until one player gives up. This mechanism is similar to Wikipedia's Talk Page and YouTube's Content ID. Intuitively, a repeated counter-notice procedure allows the two parties to credibly reveal their private information. For example, a defendant with a high value for the content in question is willing to hold out longer than a plaintiff with relatively less harm from the content. Thus, a carefully designed counter-notice procedure could elicit private information from both parties and minimize both over- and under-removal biases.

However, a repeated counter-notice procedure may result in excessive procedural costs and thus a loss in consumer welfare. Centralization turns out to be optimal for consumer welfare. The consumer-optimal procedure (which maximizes total payoffs, including costs) is a zero-cost flagging procedure that selects a winner immediately based solely on the comparison between the ex ante average value and harm. This procedure, common in email servers' spam filters, aims to minimize the loss in consumer welfare caused by prolonged argumentation.

In essence, the optimal choice of a moderation mechanism depends on the welfare criterion of the platform. If the platform's primary concern is to minimize type-I and type-II errors in its moderation decision, a repeated counter-notice procedure would be optimal. However, if the platform also takes consumers' costs associated with the moder-

ation process into account, it would be optimal to employ a zero-cost flagging procedure.

The rest of the paper proceeds as follows: Section 2 describes the virtues and flaws of the most common content moderation procedures and provides a short literature review. Section 3 contains a welfare analysis of three procedures. Section 4 looks into the efficient procedure from a mechanism design approach. Section 5 briefly concludes.

2 Background and Related literature

This section has a twofold purpose. First, we introduce three procedures commonly used by platforms to regulate the content. Second, we discuss the related literature and our contributions.

2.1 Content Moderation Procedures

Moderation can be carried out by a single moderator, usually being the platform where the content is to be displayed. Centralized moderation forces users to report the unwanted content to the platform, which then decides whether or not to remove it. One of the virtues of these systems is that moderation decisions can be enforced quickly. However, giving all the power to the platform can be to the detriment of users' preferences which may not be in line with those of the platform. A spam filter would be an example of a centralized mechanism.

The first decentralized moderation procedure is flagging, which is commonly used within social media platforms and other user-generated content platforms as a mechanism to report offensive content (Crawford and Gillespie, 2016). Flagging involves clicking or flagging specific content (such as videos, posts, etc.) that users believe should be removed or modified by the platform. Flagging is a standard feature found in almost all platforms where social interactions are prevalent, such as Facebook, Twitter, YouTube, Instagram, among others. The specific functioning of this system varies depending on the level of sophistication that the platform has decided to implement. Some platforms

simply allow users to report content as offensive or harmful, while others enable users to provide detailed reasons for why the content requires modification or deletion. The motivations for using this flagging procedure are diverse and can range from addressing issues like harassment, violence, identity theft, and copyright violations, among others.

Next, we discuss the notice and counter-notice procedure. This process is implemented by certain platforms in response to allegations that specific content is in violation of either legal mandates or platform rules. It is frequently used in cases of copyright infringement or defamatory content (Urban et al., 2017). The Digital Millennium Copyright Act in the US notably has included the notice and counter-notice procedure as the default mechanism for resolving disputes over online copyrighted content. Under this regime, copyright owners can send takedown notices to online service providers to remove the content they believe to be infringing. [3] If the receiving user believes their content was wrongly removed, they have the option to file a counter-notification to request the restitution of the content. If a copyright owner wishes to keep the content disabled after receiving a counter notice, he will need to initiate a legal action seeking a court order. [4] Some platforms have extended this procedure beyond what is legally required. YouTube Content ID, for instance, implements a three-round counter-notice process unless one party gives up. [5]

The third procedure we will cover is voting. Some prominent Q&A platforms, including Reddit[6] and Stack Overflow[7], have embraced this alternative system to handle disputes that arise between users concerning the deletion or modification of uploaded content. In this system, platforms delegate the decision-making power to the mass of users or editors themselves. The process typically involves a multi-step approach. First, users can report content they find inappropriate, offensive, or in violation of platform

[3] Section 512(c) outlines the statutory requirements that must be met for all notices. These requirements include identifying the allegedly plagiarized work, providing some reference to the alleged plagiarizer, expressing good faith on the part of the complaining party, and confirming the accuracy of the information provided. Once these criteria are met, the platform is obligated to remove the content promptly.

[4] Online platforms have a strong incentive to comply with the process otherwise they risk losing their safe harbor protection (which shields them from lawsuits). See Grimmelmann and Zhang (2023).

[5] See https://support.google.com/youtube/answer/2797454.

[6] You can find its content moderation policies at https://www.redditinc.com/policies/content-policy.

[7] See the conduct agreement that every moderator has to sign at Stack Overflow: https://stackoverflow.com/help/mod-agreement-policies

rules. Once a report is submitted by users or directly supplied by the platform, trained human moderators or automated systems review the reported content to assess its compliance with the platform conduct guidelines. The moderators may follow a set of pre-defined criteria to evaluate the content impartially. Depending on the platform's policies, multiple moderators may review the same content to reduce bias and ensure fair judgment. After examining the reported content, moderators may vote on whether to keep, remove, or take other appropriate actions on it. The decision is often determined by a majority vote or adherence to a pre-defined threshold of agreement among the moderators.

Some platforms may use a combination of the aforementioned procedures. Wikipedia, for example, develops a moderation system that involves all three techniques and a centralized arbitration as a last resort. [8] All users can flag articles for possible removal. [9] Each article is associated with a Talk Page where editors can discuss the content of the article repeatedly. [10] The Talk Page can thus be viewed as a repeated version of the counter-notice procedure. If discussion on the Talk Page fails, editors may seek outside volunteers to help resolve the dispute. This includes asking for a "Third opinion", opening a "Request for comment", and making a request at an appropriate noticeboard. [11] Wikipedia emphasizes the consensus of the community in which every registered user can participate and vote. Voting and community-building takes place either online such as in Skype, chat rooms, e-mail, or in a real local pub or other venue at Wikipedia face-to-face meetings. [12] In case all other avenues have been exhausted, the editors may take the dispute to the Arbitration Committee, which will issue and enforce a binding decision. [13]

2.2 Related Literature

Content moderation has been a subject of extensive interest in the legal and social science literature. A pioneering work in that literature is Grimmelmann (2015), arguing that con-

[8] See Jemielniak (2014) and Clark et al. (2019) for detailed documentation of Wikipedia's content policy.
[9] See https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_forum.
[10] See https://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines.
[11] For a more detailed explanation of the process, see https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution.
[12] See https://en.wikipedia.org/wiki/Wikipedia:Meetup.
[13] See https://en.wikipedia.org/wiki/Wikipedia:Arbitration/Policy.

tent moderation is the reason some online communities thrive while others become ghost towns. Grimmelmann (2015) provides a taxonomy of the moderation design space.[14] The principal techniques of moderation are the rules that decide which content to allow or remove and how to enforce these decisions. The removal decision can be performed manually or through automated algorithms, transparently or confidentially, centralized with one powerful moderator or decentralized with dispersed moderators. It can also be ex-ante deterrence (i.e., preventing unwanted content beforehand) or ex-post punishment (i.e., addressing issues after they arise). Our analysis compares a wide range of moderation designs commonly used in practice, including the centralized ex-ante mechanism, and the decentralized ex-post removal mechanism. [15] This literature also recognizes that platform users play a pivotal role in moderation, and their active engagement leads to a shift in the platform's governance model, from a centralized authority to a variety of decentralized and community-driven approaches. (Crawford and Gillespie, 2016; Jemielniak, 2014; West, 2018; see Gillespie et al. (2023) for a survey)

The economic literature on content moderation is in more limited supply. A number of empirical papers study certain moderation procedures we discussed in this paper. Jiménez Durán (2022) studies flagging in Twitter and finds that flagging increases the likelihood of content deletion, and content moderation can marginally increase advertising revenues of the platform. Zhang (2021) studies the notice-and-takedown procedure in GitHub and finds a persistent drop in original contributions following the takedown of a repository. Kwan et al. (2023) examines the voting-type mechanism that involves the voluntary participation of buyers and sellers as jurors to handle disputes on the platform, and finds potential issue of in-group bias that may arise through the process. Underlying these empirical findings is the question of whether the moderation procedure achieves socially desirable outcome, together with the trade-off of over-removal versus under-removal. [16]

The theoretical literature in economics on content moderation is still in its infancy.

[14] See Klonick (2017) for a discussion of the definition of content moderation.

[15] See Douek (2022) for a discussion on the ex-ante approach. See Goldman (2021) for a variety of moderation remedies beyond content removal.

[16] Other papers investigate moderation techniques that are beyond the scope of this paper. For example, Bolton et al. (2018) and Bolton et al. (2023) studies the feedback revision rule used by some platforms.

A few recent works begin to develop models for content moderation. Madio and Quinn (2021) studies the incentive of an ad-funded platform to filter unsafe content, and demonstrates that platforms are more incentivized to moderate content when users' preference align with advertisers' preference. Liu et al. (2022) study the implication of revenue models for the platform's decision to moderate extreme content, and find that moderation is more likely to occur under an advertising model rather than a subscription model. Grimmelmann and Zhang (2023) develop a model of moderation based on externalities, imperfect information, and investigation costs, and use it to show the tradeoff of over-moderation versus under-moderation. All of the above papers focus on what kind of content to moderate from the platform's perspective, and moderation is entirely the platform's decision in these papers. [17] Our paper instead focus on *how* to moderate. We model moderation as an outcome of the joint decision of the platform and the participating users, thereby showing the decentralization of content moderation.

The papers closest in form to ours are Papanastasiou et al. (2023) and Lee and Cui (2023). Papanastasiou et al. (2023) compare the centralized dispute resolution mechanism with a semi-decentralized mechanism (where the seller is allowed to remove reviews without consulting the platform) in the context of malicious reviews and blackmail. Their game-theoretical analysis shows that decentralization, when implemented properly, can improve efficiency in dispute resolution. Similarly, Lee and Cui (2023) analytically compare centralized platform-led adjudication and crowd-sourced voting mechanism for online labor platforms. They find that decentralization can result in a fairer outcome and eliminate the bias of the centralized mechanism if there is sufficient heterogeneity in users' private signals. Our paper is different in four aspects. First, we expand the scope and consider three most common content moderation procedures. We compare their respective moderation outcome and relative efficiency and advantages. Second, our model is institution-free and all of our analysis is based on the distribution of harm and value of the content, which is applicable in many settings. Third, we break down the inefficiency into three components: type I error, type II error, and procedural costs. Lastly and importantly, we are not aware of any paper that explores and generalizes the economic implications of moderation procedures in a mechanism design framework. To date, our

[17] There is a related literature on whether platforms should be held legally liable for harmful content. See Fagan (2020), De Chiara et al. (2021), Hua (2021), and Grimmelmann and Zhang (2023).

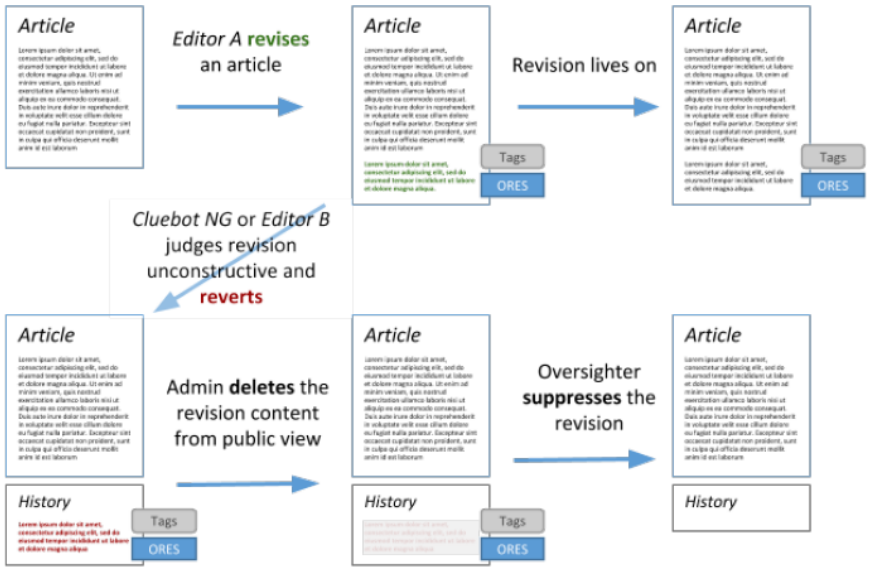


Figure 2: Article revision and content removal on Wikipedia. Source: Clark et al. (2019)

paper appears the first one to apply mechanism design to content moderation problem. Additional insights have been uncovered by the mechanism design approach. For instance, centralization can be optimal if procedural costs is weighted equally as wrongful removal.

3 A Model of Content Moderation

In what follows, we present a game theoretic analysis of the four popular moderation procedures discussed in section 2. The first procedure is centralized platform adjudication. The other three procedures all delegate the moderation decision to the users in some degree, but the ways the users can participate are different. The flagging procedure gives the plaintiff an opportunity to send a removal notice, and it is the simplest among the four. The counter-notice procedure allows the plaintiff and the defendant to counter each other's notices. The voting procedure invites all the users to vote before enacting a moderation decision. Our purpose is to (i) make positive predictions on the removal outcome for each procedure, and (ii) identify the normative implications of the source of inefficiency for each procedure.

We start with bilateral disputes on an item of content. There is an individual who is harmed by the content, we call him the plaintiff (denoted by subscript p). There is another individual whose content harms, we call him the defendant (denoted by subscript d). The harm can be copyright infringement, defamation, misinformation, etc.

Payoffs. If the content remains, the defendant gets a payoff of v and the plaintiff gets a payoff of $-h$. If the content is removed, both the plaintiff and the defendant get 0. Therefore, the plaintiff gains from removal and has an incentive to take down the content; the defendant loses from removal and has an incentive to keep the content. h and v are private information: only the plaintiff observes h and only the defendant observes v. The distribution $F_h(h)$ and $F_v(v)$ are common knowledge.

Content Moderation. Content moderation is a set of procedures that results in the removal of the content with some probability at certain costs. Let $P(h, v)$ be the probability function on whether the content is removed.

Social Efficiency. A content moderation outcome is socially efficient if it maximizes the total allocation payoffs of the plaintiff and the defendant, i.e., $P(h, v)h + (1 - P(h, v))v$. Thus, $P^e(h, v) = 1$ if $h > v$ and $P^e(h, v) = 0$ otherwise. That is, removing the content is socially efficient if the harm exceeds the value. [18] A moderation procedure is error-free if it induces a socially efficient moderation outcome.

3.1 Centralization

We begin with a simple model of centralization as a benchmark. The platform decides whether to remove the content. There is a probability p that the platform will remove. The probability p is announced ex-ante and it is common knowledge for all users. p depends on the substance of the content in question and the strength of the related legal enforcement. For instance, child pornography is strictly prohibited in all the jurisdictions, so $p = 1$. If the content is an infringing work of copyright, p would be large under the DMCA, but bounded away from 1 because of the fair use doctrine.

[18] Note that procedure costs are not included.

The resulting allocation is $P^{\text{cen}}(h, v) = p$ for any pair of (h, v), which diverges from the efficient allocation rule $P^e(h, v)$. We can quantify the welfare loss. We start by defining the two types of errors a moderation procedure can cause. Let α be the type I error that the content is removed in equilibrium, but social efficiency requires the content to remain. Thus α denotes the welfare loss from over-removal. Type I error occurs if $P^e(h, v) = 0$ and $P^{\text{cen}}(h, v) > 0$. Under centralization, this means $h \leqslant v$ and the removal happens erroneously with probability p.

$$\begin{aligned} \alpha^{\text{cen}} &= \int_0^1 \int_0^v (P^e - P^{\text{cen}})(h - v) dF_h(h) dF_v(v) \\ &= p \int_0^1 \frac{v^2}{2} dv \\ &= p. \end{aligned} \tag{1}$$

Let β be the type II error that the content is not removed in equilibrium, but social efficiency requires the content to remove. Thus β denotes the welfare loss from under-removal. Type II error occurs if $P^e(h, v) = 1$ and $P^{\text{cen}}(h, v) < 1$. Under centralization, this means $h > v$ and non-removal happens erroneously with probability $1 - p$.

$$\begin{aligned} \beta^{\text{cen}} &= \int_0^1 \int_v^1 (P^e - P^{\text{cen}})(h - v) dF_h(h) dF_v(v) \\ &= (1 - p) \int_0^1 \frac{(1 - v)^2}{2} dv \\ &= \frac{1 - p}{6}. \end{aligned} \tag{2}$$

Let W^{cen} be the welfare under centralization and can be decomposed as

$$W^{\text{cen}} = W^e - \alpha^{\text{cen}} - \beta^{\text{cen}} \tag{3}$$

The welfare loss is decomposed into four parts: type I error, type II error, and procedural costs (which is zero for centralization). Over-removal weighs more in the welfare loss. This simple form of centralization is inefficient because the simple randomization cannot align the moderation outcome with the disputants' preferences.

3.2 Flagging Procedure

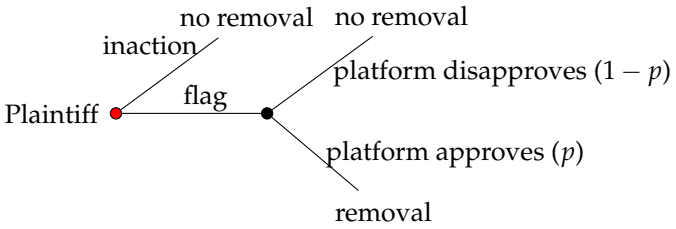


Figure 3: The Timing of the Game Under the Flagging Procedure

The simplest form of decentralized content moderation would be the flagging procedure. The procedure allows a single user to file a complaint against a disputed content.

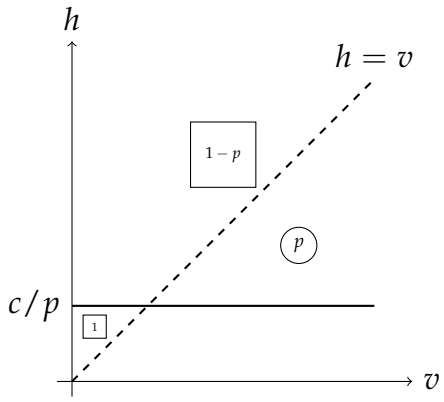
Timing. The flagging procedure is depicted in figure 3 where a red node is a decision node of the plaintiff, a black node is a nature's move, an edge is an action, and an end node is an outcome. The plaintiff chooses whether to flag the content or not. There is a cost c to send a notice. If he chooses inaction, the content remains. If he chooses to flag, the platform will decide whether to honor his request. As before, there is a probability p that the platform will approve. The platform's approval will lead to the removal of the content.

A pure strategy of the plaintiff is the binary choice of flagging. This is a decision problem. We make the tie-breaking assumption that if either the plaintiff or the defendant is indifferent between posting and inaction, she will choose inaction. The plaintiff will send a notice if the expected value of platform decision is greater than the cost of sending a notice:

$$ph > c. \tag{4}$$

The probability of a platform decision is $1 - F(c/p)$. The probability of a removal is $p \Pr(h > c/p) = p[1 - F_h(c/p)]$. The decision to report a flag is similar to buying a lottery, though this lottery has significant external effects on social welfare.

Not all removals are efficient. The flagging procedure is not error-free. The resulting



Note: Circle denotes over-removal, and Square denotes under-removal.

Figure 4: Equilibrium vs. Efficiency under the flagging procedure

allocation $P^{\text{flag}}(h, v)$ is

$$P^{\text{flag}}(h, v) = \begin{cases} p & \text{if } h > c/p, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

which diverges from the efficient allocation rule $P^e(h, v)$. Figure 4 depicts a graphical comparison of the equilibrium and the efficiency under the flagging procedure in the (h, v) space. We pick $p = 1$ for illustration purposes. The horizontal solid line $h = c$ is the indifference condition of the plaintiff regarding sending a notice. Above the solid line, the plaintiff sends the notice; below the line, the plaintiff will not. The dashed line $h = v$ further divides the quadrant into four parts. Above the line, $h > v$ so removal is socially efficient; below the line, $h \leqslant v$ so carrying the content is socially efficient. The area with neither a circle nor a square is where the equilibrium outcome is socially efficient. The area above the solid line and below the dashed line is denoted as circle. The circle area corresponds to Type I error and this is where the procedure leads to the error of over-removal. If (h, v) falls into this region, the content should never be removed but it will be removed under the flagging procedure. Notice that the pair (h, v) in the circle area tends to be large, implying large efficiency loss of over-removal. The area below the solid line and above the dashed line is denoted as square. The square area corresponds to Type II error and this is where the procedure leads to the error of under-removal. If (h, v) falls into this region, the content should be removed but it will not be removed under the flagging procedure. Notice that the pair (h, v) in the square area tends to be small,

implying limited efficiency loss of under-removal.

A standard comparative statics exercise is to see whether the equilibrium outcome will be improved by varying some of the parameters. As the cost c becomes lower, the solid line shifts downward resulting in a smaller square area but a larger circle area. A reduction in under-removal error by changing c will increase the over-removal error. Increasing p will reduce over-removal error at the cost of increasing under-removal error. The above observations illustrate a trade-off in fine-tuning the parameters: fixing the procedure, reducing one type of error will lead to an increase in the other type of error.

Type I error occurs if $P^e(h,v) = 0$ and $P^{\text{flag}}(h,v) > 0$. Under flagging procedure, this means $h \leqslant v$, $h > c/p$, and the removal happens erroneously with probability p.

$$\begin{aligned}\alpha^{\text{flag}} &= \int_0^1 \int_{c/p}^{v} (P^e - P^{\text{flag}})(h-v) dF_h(h) dF_v(v) \\ &= p \int_0^1 \frac{(v-c/p)^2}{2} dv \\ &= p \left[\frac{(1-c/p)^3}{6} - \frac{(-c/p)^3}{6} \right].\end{aligned} \tag{6}$$

As c increases, α^{flag} decreases. Type II error occurs if $P^e(h,v) = 1$ and $P^{\text{flag}}(h,v) < 1$. Under flagging procedure, this means either (i) $h > v$, $h \leqslant c/p$, and no removal ever happens, or (ii) $h > v$, $h > c/p$, and non-removal happens erroneously with probability $1-p$.

$$\begin{aligned}\beta^{\text{flag}} &= \int_0^1 \left[\int_v^{c/p} (P^e - P^{\text{flag}})(h-v) dF_h(h) + \int_{\max\{v,c/p\}}^1 (P^e - P^{\text{flag}})(h-v) dF_h(h) \right] dF_v(v) \\ &= \int_0^1 \frac{(c/p-v)^2}{2} dv + (1-p) \int_{c/p}^1 \frac{h^2}{2} dh \\ &= \left[\frac{(c/p)^3}{6} - \frac{(c/p-1)^3}{6} \right] + (1-p) \left[\frac{1}{6} - \frac{(c/p)^3}{6} \right]\end{aligned} \tag{7}$$

As c increases, β^{flag} increases as well. Let W^{flag} be the welfare under the flagging proce-

dure and can be decomposed as

$$W^{\text{flag}} = W^e - \alpha^{\text{flag}} - \beta^{\text{flag}} - c[1 - \frac{c}{p}] \tag{8}$$

where $c[1 - \frac{c}{p}]$ is the expected cost of the flagging procedure. The welfare loss is decomposed into four parts: type I error, type II error, and procedural costs.

The equilibrium is inefficient because it lead to either over-removal or under-removal. It leads to over-removal because the plaintiff does not take into account the negative externality of content removal to the users who value the content. It leads to under-removal because the transaction cost of the procedure prevents low types to initiate a claim.

3.3 The Counter-notice Procedure

The plaintiff files a complaint against a disputed content. The creator of the content, the defendant, has the right to make a counter-notice to rebut the complaint. Settlement negotiations are then conducted against the outside option of a platform decision.

Timing. Figure 5 describes the notice and counter-notice procedure where a red node is a decision node of plaintiff, a blue node is a decision node of defendant, a black node is a nature's move, an edge is an action, and an end node is an outcome. The plaintiff chooses whether to send a notice or not. The notice can be viewed as a settlement proposal specifying that the plaintiff will drop the lawsuit if the defendant agrees to terminate the disputed work. The defendant decides whether to accept the settlement. If the defendant accepts, the disputed content will be removed. If the defendant rejects, she will send a counter-notice, in which case the plaintiff will have to choose whether to proceed to platform decision or drop the complaint. If the plaintiff chooses not to litigate, the content remains on the platform. If a platform decision does take place, there is a probability $p \in [0, 1]$ that the plaintiff will prevail and the content will be removed. With probability $1 - p$, the defendant prevails, the court dismisses the claim, and the content remains.

For welfare comparison, we will start with the assumption that both the plaintiff and the defendant know h and v. We will then discuss how incomplete information distorts

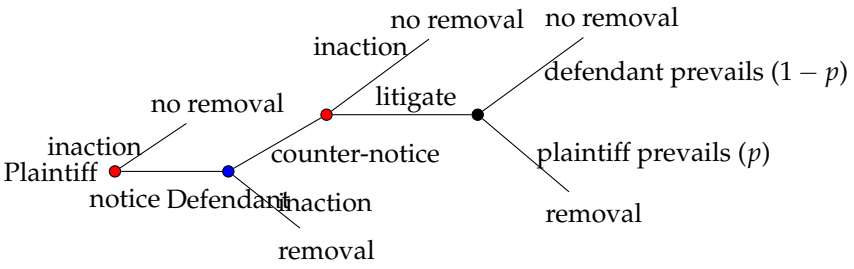


Figure 5: The Timing of the Game Under the Notice and Counter-notice Procedure

the incentive for truth-telling.

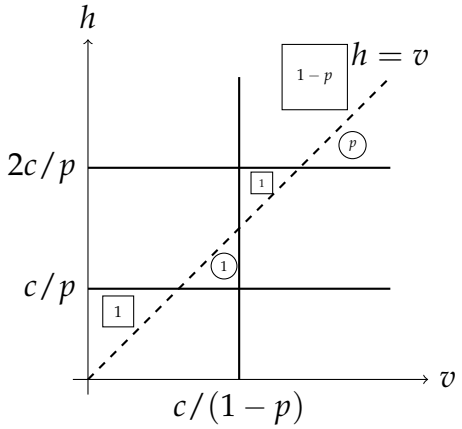
Under the assumption of complete information, a pure strategy of the plaintiff is a complete contingent plan that picks the notice choices for every realization of the previous history. A pure strategy of the defendant is a contingent plan that picks the counter-notice choice for every realization of the previous history. The solution concept is subgame perfect Nash equilibrium.

Under the notice and counter-notice procedure, there are two ways a content can be removed: either because the defendant complies with the takedown notice, or because the plaintiff prevails in the platform decision. Likewise, there are two ways this procedure might keep the content: either because the plaintiff retracts his notice after receiving a counter-notice, or because the defendant prevails in the platform decision. Platform decision will happen if the defendant chooses to send a counter-notice and the plaintiff does not retract his notice such that the two parties cannot reach an agreement via settlement talks.

We solve the game by backward induction. The defendant will respond to the notice only if the plaintiff's threat to litigate is credible. The litigation threat is credible if and only if $ph > c$, i.e., the expected gain from the platform decision is greater than the notice cost. If the above condition holds, the defendant will issue a counter-notice if and only if

$$(1-p)vI\{ph - c > 0\} - c > 0,$$

the expected gain from the platform decision (conditional on the credibility of the threat) is greater than the payoff from inaction which leads directly to removal. Anticipating this,



Note: Circle denotes over-removal, and Square denotes under-removal.

Figure 6: Equilibrium vs. Efficiency under the Counter-notice procedure

the plaintiff will send the notice in the first place if

$$(ph - c)I\{(1-p)v - c > 0\} + hI\{(1-p)v - c < 0\} > c.$$

Suppose $(1-p)v > c$ such that the notice will be countered, the plaintiff will send a notice if $ph > 2c$. In this case, the equilibrium outcome is one where the plaintiff sends a notice, the defendant sends a counter-notice, and the two parties proceed to platform decision. Suppose $(1-p)v \leqslant c$ such that the notice will not be countered, the plaintiff will send a notice if $h > c$, which is implied by $ph > c$. In this case, the equilibrium outcome is one where the plaintiff sends a notice, the defendant chooses inaction, and thus content is removed.

The counter-notice procedure is not error-free either. The resulting allocation $P^{\text{notice}}(h, v)$ is

$$P^{\text{notice}}(h,v) = \begin{cases} 1 & \text{if } h > \frac{c}{p} \text{ and } v \leqslant \frac{c}{1-p}, \\ p & \text{if } h > \frac{2c}{p} \text{ and } v > \frac{c}{1-p}, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

which diverges from the efficient allocation rule $P^e(h, v)$. Notice that $P^{\text{notice}}(h, v) = P^{\text{flag}}(h, v)$ if $p = 1$ or $p = 0$ but for generic value of p, $P^{\text{notice}}(h, v)$ improves upon $P^{\text{flag}}(h, v)$. Figure 6 extends figure 4 to graphically compare the equilibrium and the ef-

ficiency under the notice and counter-notice procedure in the (h, v) space. The dashed line $h = v$ remains the same. The four solid lines divide the quadrant into six parts. The blocks above the line c and left to the line c/ε are the equilibrium outcomes where the content is removed without a counter-notice. The block above the line $2c$ and right to the line c/ε is the equilibrium outcome where the content is removed by platform decision. The other four blocks in the southeast corner are the inaction equilibrium. The area with neither a circle nor a square is where the equilibrium outcome is socially efficient. The square area corresponds to Type II error under the counter-notice procedure, and this is where the procedure leads to the error of under-removal. Notice that in the graph, it is the same area as under the flagging procedure. The circle areas correspond to Type I error under the counter-notice procedure, and this is where the procedure leads to the error of over-removal. It is important to notice that the circle area is smaller than that under the flagging procedure. The counter-notice procedure might be able to reduce the one type of error *without* increasing the other type of error. This welfare improvement stems from the fact that the counter-notice procedure divides the preferences into finer sets and result in better moderation outcome accordingly.

Type I error occurs if $P^e(h, v) = 0$ and $P^{\text{notice}}(h, v) > 0$. Under the counter-notice procedure, this means either (i) initial notice not being countered ($h \leqslant v$, $h > c/p$, and $v < c/(1-p)$), or (ii) plaintiff prevails in platform decision ($h \leqslant v$, $h > 2c/p$, $v > c/(1-p)$), and the removal happens erroneously with probability p.

$$\begin{aligned}
\alpha^{\text{notice}} &= \int_0^{c/(1-p)} \int_{c/p}^{v} (P^e - P^{\text{notice}})(h-v) dF_h(h) dF_v(v) + \int_{c/(1-p)}^{1} \int_{2c/p}^{v} (P^e - P^{\text{notice}})(h-v) dF_h(h) dF_v(v) \\
&= \int_0^{c/(1-p)} \frac{(v - c/p)^2}{2} dv + p \int_{c/(1-p)}^{1} \frac{(v - 2c/p)^2}{2} dv \\
&= \left[\frac{(c/(1-p) - c/p)^3}{6} - \frac{(-c/p)^3}{6} \right] + p \left[\frac{(1 - 2c/p)^3}{6} - \frac{(c/(1-p) - 2c/p)^3}{6} \right]
\end{aligned} \tag{10}$$

As c increases, α^{notice} increases. Type II error occurs if $P^e(h, v) = 1$ and $P^{\text{notice}}(h, v) < 1$. Under the counter-notice procedure, this means (i) no initial notice ($h > v$, $h \leqslant c/p$), or (ii) initial notice being countered and no second notice ($h > v$, $h \in [c/p, 2c/p)$, and $v > c/(1-p)$), or (iii) defendant prevails in platform decision ($h > v$, $h > 2c/p$, $v >$

$c/(1-p)$), and non-removal happens erroneously with probability $1-p$.

$$\begin{aligned}
\beta^{\text{notice}} &= \int_0^1 \int_v^{c/p} (P^e - P^{\text{notice}})(h-v) dF_h(h) dF_v(v) + \int_{c/(1-p)}^1 \int_v^{2c/p} (P^e - P^{\text{notice}})(h-v) dF_h(h) dF_v(v) \\
&\quad + \int_{2c/p}^1 \int_{c/(1-p)}^h (P^e - P^{\text{notice}})(h-v) dF_v(v) dF_h(h) \\
&= \int_0^1 \frac{(c/p-v)^2}{2} dv + \int_{c/(1-p)}^1 \frac{(2c/p-v)^2}{2} dv + (1-p) \int_{2c/p}^1 \frac{(h-c/(1-p))^2}{2} dh \\
&= \left[\frac{(c/p)^3}{6} - \frac{(c/p-1)^3}{6} \right] + \left[\frac{(2c/p-c/(1-p))^3}{6} - \frac{(2c/p-1)^3}{6} \right] \\
&\quad + (1-p) \left[\frac{(1-c/(1-p))^3}{6} - \frac{(2c/p-c/(1-p))^3}{6} \right].
\end{aligned} \tag{11}$$

As c increases, β^{notice} increases as well. Let W^{notice} be the welfare under the counter-notice procedure and can be decomposed as

$$W^{\text{notice}} = W^e - \alpha^{\text{notice}} - \beta^{\text{notice}} - c[1-\frac{c}{p}] - c[1-\frac{c}{p}][1-\frac{c}{1-p}] - c[1-\frac{c}{1-p}][1-\frac{2c}{p}] \tag{12}$$

where $c[1-\frac{c}{p}]$ is the expected cost of the first plaintiff notice, $c[1-\frac{c}{p}][1-\frac{c}{1-p}]$ is the expected cost of the counter-notice, and $c[1-\frac{c}{1-p}][1-\frac{2c}{p}]$ is the expected cost of the second plaintiff notice. The welfare loss is decomposed into four parts: type I error, type II error, and procedural costs.

Incomplete information may give users an incentive to over-state their types. When the plaintiff decides whether to send a notice, he may want to send the notice even though his type h is below the cost c. Likewise, a defendant may want to send a counter-notice even though his type v is below the cost c. The reason is that the plaintiff (defendant) has an incentive to build a reputation of a higher-threat type in the hope that the defendant (plaintiff) would be deterred to send a counter-notice. In a Perfect Bayesian equilibrium, a plaintiff will send the initial notice if $h > \hat{h}_1$, will send the second notice if $h > \hat{h}_2$, and the defendant will send the counter-notice if $v > \hat{v}$. The cutoffs $\hat{h}_1$, $\hat{h}_2$ and $\hat{v}$ can be found by substituting the posterior belief into the users' best response functions (which leads to four fixed point equations). It can be shown that $\hat{h}_1 < c/p$, $\hat{h}_2 < 2c/p$, and $\hat{v} < c/(1-p)$. So both the plaintiff and the defendant are more likely to send notice and counter-notice.

In figure 6, the two horizontal lines would shift downward and the vertical line would shift inward. The area of platform decision grows, but the characterization of the welfare loss remains the same.

3.4 The Voting Procedure

Content disputes may go beyond two individuals and thus involve multilateral bargaining. There is a group of individuals n_h who are harmed by the content, and there is another group of individuals n_v whose content harms. If the content remains, the defendant i gets a payoff of v_i and the plaintiff i gets a payoff of $-h_i$. If the content gets removed, all users get a payoff of 0. Every h_i and v_i are private information. A moderation outcome is socially efficient if it maximizes $P(\boldsymbol{h}, \boldsymbol{v}) \sum_i h_i + (1 - P(\boldsymbol{h}, \boldsymbol{v})) \sum_i v_i$. Removing the content is socially efficient if the *total* harm exceeds the *total* value.

The most popular moderation procedure for multilateral procedure is the voting procedure. Moderation decisions are made by the community of all users $n = n_h + n_v$. The affirmative votes of fraction $k \leqslant n$ are required to enact a moderation decision. If an agreement is not reached for the disputed content, there will be a platform decision.

This model shows a conflict resolution method in which decision-making is not handled by a single individual, resulting in judgments that are similar to those made in a court of law by a panel of judges. In this scenario, though, decision-makers have interests at stake since they must select what content should remain on the platform, and they reveal their preferences about it through voting. As a result, the incentives to resolve the argument in one direction or another are biased in favor of the user in charge of making the suggestion.

Timing. Figure 7 describes the voting procedure where a red node is a decision node of plaintiff, a blue node is a decision node of defendant, a yellow node is a decision node for a platform, a square black node is a nature's move, an edge is an action, and an end node is an outcome. A plaintiff is randomly chosen among n_h users of the harmed group. The plaintiff chooses whether to send a notice or not. If a notice has been sent, all the n

members in the community choose to vote yes or no. If the number of affirmative votes is greater than or equal to the threshold k, the notice is enacted and the content will be removed. Otherwise, platform decision will take place, which may result in a removal of the content with probability p.

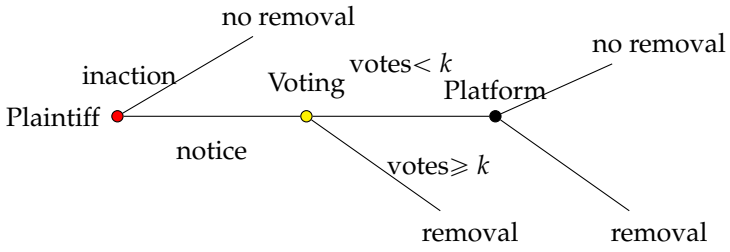


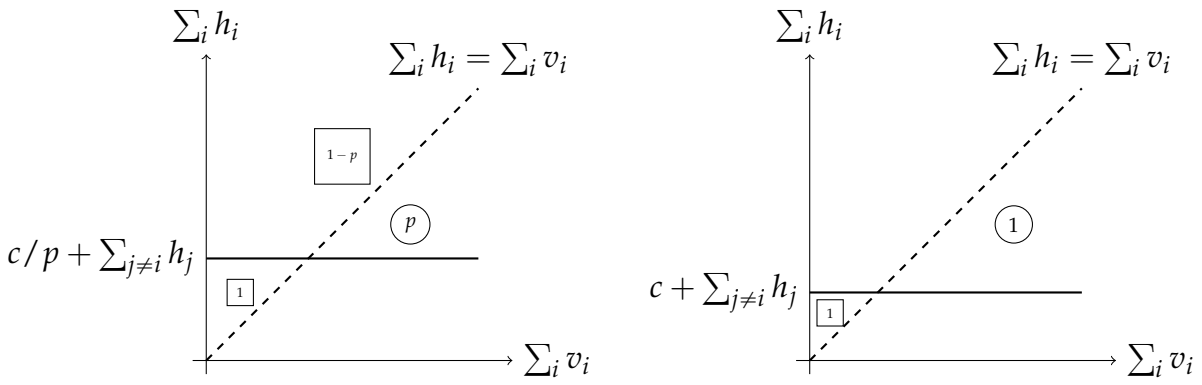
Figure 7: The Timing of the Game Under the Voting Procedure

At the voting stage, because there are only two candidates, users will vote sincerely. Sincere voting will be the equilibrium outcome of the subgame. Under sincere voting, type h users will vote Yes, and type v users will vote No. If the threshold is $k/n = 1/2$, it will be the median user that determines the moderation outcome. If the threshold is $k/n = 3/4$, it will be the top quartile user that determines the moderation outcome. The *number* of harmed users n_h would play a key role here. If $n_h \geqslant k$, then the removal request will be enacted, and consequently, the plaintiff finds it optimal to send a notice if $h_i > c$. If $n_h < k$, then the removal request is rejected, the removal decision is left to the platform, and the plaintiff is willing to send a notice only if $ph_i > c$. Also observe that the plaintiff's preference h_i weighs more than the harm suffered by the rest of the group.

The voting procedure is not error-free. The resulting allocation $P^{\text{vote}}(\boldsymbol{h}, \boldsymbol{v})$ is

$$P^{\text{vote}}(\boldsymbol{h}, \boldsymbol{v}) = \begin{cases} 1 & \text{if } h_i > c \text{ and } n_h \geqslant k, \\ p & \text{if } h_i > \frac{c}{p} \text{ and } n_h < k, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

which diverges from the efficient allocation rule $P^e(\boldsymbol{h}, \boldsymbol{v})$. Figure 8 extends figure 4 to graphically compare the equilibrium and the efficiency under the voting procedure in the total harm and the total value space $(\sum_i h_i, \sum_i v_i)$. The comparison now depends on whether there would be an agreement on removal. Figure 8(a) shows the case where



(a) $n_h < k$ (b) $n_h \geqslant k$

Note: Circle denotes over-removal, and Square denotes under-removal.

Figure 8: Equilibrium vs. Efficiency under the Voting procedure

$n_h < k$ such that the number of harmed users is not sufficient to support the removal. The dashed line $\sum_i h_i = \sum_i v_i$ is the efficient outcome. The square area corresponds to β under the voting procedure, and this is where the procedure leads to the error of under-removal. The circle areas correspond to α under the voting procedure, and this is where the procedure leads to the error of over-removal. The only case the content might get removed is through platform decision. In this case, both the square and the circle area are exactly the same as that under the flagging procedure. When $n_h < k$, the voting procedure extends what would happen under the flagging procedure to multilateral disputes. Figure 8(b) shows the case where $n_h \geqslant k$ such that the number of harmed users is sufficient to support the removal. The voting will remove the content for sure if a notice has been sent. Compared with $n_h < k$, the over-removal bias grows substantially because the area of the circle is larger and the probability is higher. But at the same time the under-removal bias decreases considerably.

If $n_h < k$, we have $\alpha^{\text{vote}} = \alpha^{\text{flag}}$, $\beta^{\text{vote}} = \beta^{\text{flag}}$, $W^{\text{vote}} = W^{\text{flag}}$. We therefore focus on $n_h \geqslant k$. Type I error occurs if $P^e(h,v) = 0$ and $P^{\text{vote}}(h,v) > 0$. Under the voting procedure, this means $h \leqslant v$, $h > c$.

$$\begin{aligned}\alpha^{\text{vote}} &= \int_{[0,1]^{n_h+n_v-1}} \int_c^{\sum_i v_i + \sum_{j\neq i} h_j} (P^e - P^{\text{vote}})(\sum_i h_i - \sum_i v_i) dF_h(h_i) dF_h(\boldsymbol{h}_{-i}) dF_v(\boldsymbol{v}) \\ &= \int_{[0,1]^{n_h+n_v-1}} \frac{(\sum_i v_i + \sum_{j\neq i} h_j - c)^2}{2} d\boldsymbol{h}_{-i} d\boldsymbol{v}.\end{aligned} \tag{14}$$

where $\sum_i v_i + \sum_{j \neq i} h_j$ follows an Irwin-Hall distribution. As c increases, α^{vote} decreases. Type II error occurs if $P^e(h, v) = 1$ and $P^{\text{vote}}(h, v) < 1$. Under the voting procedure, this means either (i) $h > v, h \leqslant c$.

$$\begin{aligned} \beta^{\text{vote}} &= \int_{[0,1]^{n_h+n_v-1}} \int_{\sum_i v_i + \sum_{j \neq i} h_j}^{c} (P^e - P^{\text{vote}})(\sum_i h_i - \sum_i v_i) dF_h(h_i) dF_h(\boldsymbol{h}_{-i}) dF_v(\boldsymbol{v}) \\ &= \int_{[0,1]^{n_h+n_v-1}} \frac{(c - \sum_i v_i - \sum_{j \neq i} h_j)^2}{2} d\boldsymbol{h}_{-i} d\boldsymbol{v}. \end{aligned} \tag{15}$$

As c increases, β^{vote} increases as well. Let W^{vote} be the welfare under the flagging procedure and can be decomposed as

$$W^{\text{vote}} = W^e - \alpha^{\text{vote}} - \beta^{\text{vote}} - c[1 - c] \tag{16}$$

where $c[1 - c]$ is the expected cost of the voting procedure (assuming voting is costless). The welfare loss is decomposed into four parts: type I error, type II error, and procedural costs. The inefficiency of the voting procedure comes from the fact that it ignores the intensity of the preference distribution.

3.5 Comparison of Four Moderation Procedures

In this subsection, we compare the four moderation procedures. Each procedure can change the over-removal loss, the under-removal loss, the cost of the procedure. We will reach the conclusion that if the primary concern is welfare loss from over-removal bias, the counter-notice procedure is the most efficient option of the four. If the primary concern is welfare loss from under-removal bias, the voting procedure is the most efficient option. Centralization, however, is the cost-minimizing procedure of the four.

Figure 9 plots the over-removal error resulting from the four procedures as a function of the probability p. The blue line plots the type I error of the flagging procedure. The green line plots the type I error of the counter-notice procedure. The red line plots the type I error of the voting procedure. For all four procedures, over-removal bias decreases as p grows. There is a clear ranking of the four procedures in terms of over-removal error:

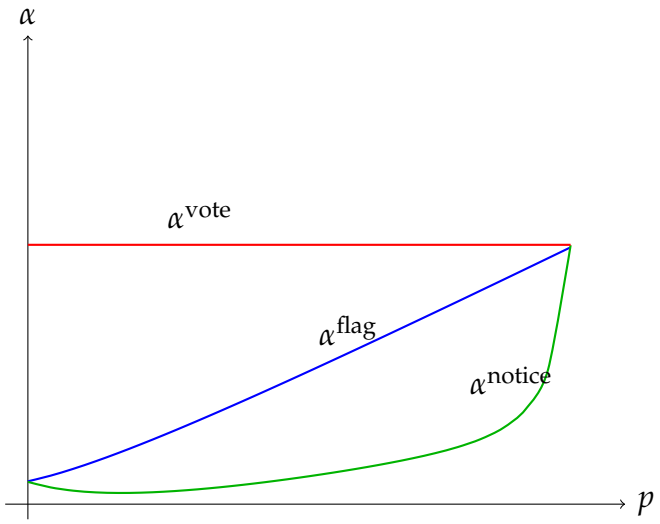


Figure 9: Comparing Over-removal Error of Four Procedures as a Function of Platform Decision

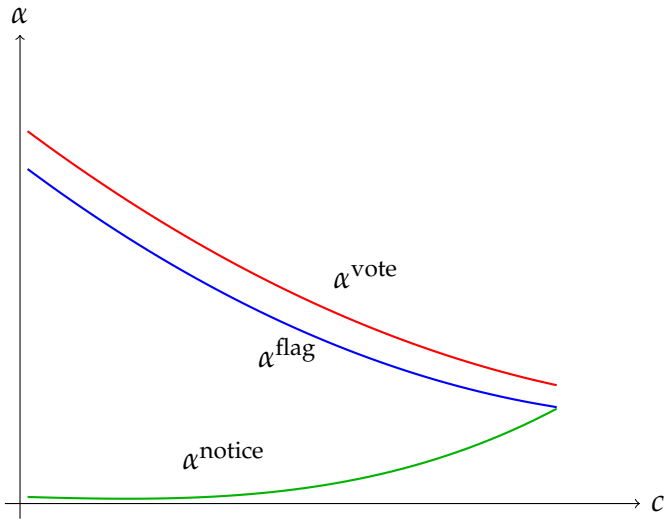


Figure 10: Comparing Over-removal Error of Four Procedures as a Function of Procedural Cost

for all p, the green line is below the blue line, which is below the red line. The counter-notice procedure has the smallest over-removal error among the four procedures. Figure 10 plots the over-removal error resulting from the four procedures as a function of the notice cost c. The same observation holds: for all c, the green line is below the blue line, which is below the red line.

If the inefficiency concern for content moderation is over-removal, the counter-notice

procedure is the best option.

Proposition 1. *For all* (c, p), $\alpha^{notice} \leqslant \alpha^{flag} \leqslant \alpha^{vote} \leqslant \alpha^{cen}$. *The counter-notice procedure minimizes the over-removal error* α *among the four procedures.*

The counter-notice procedure is superior because of *preference revelation*. Counter-notice allows the defendant to reveal her preference so that the settlement can solve the dispute for more (h, v) pairs, and settlement is more efficient than platform decision. This explains why copyright-based platform such as YouTube prefers the counter-notice procedure because DMCA abuse and over-removal are the primary concern and content providers should have an opportunity to respond.

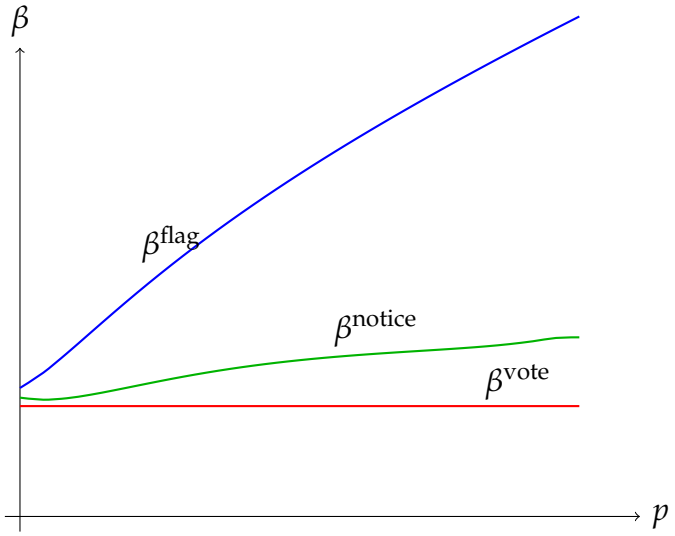


Figure 11: Comparing Under-removal Error of Four Procedures as a Function of Platform Decision

Figure 11 plots the under-removal error resulting from the four procedures as a function of the probability p. The blue line plots the type II error of the flagging procedure. The green line plots the type II error of the counter-notice procedure. The red line plots the type II error of the voting procedure. For all four procedures, under-removal bias increases as p grows. There is a clear ranking of the four procedures in terms of under-removal error: for all p, the red line is below the green line, which is below the blue line. The voting procedure has the smallest under-removal error among the four procedures. Figure 10 plots the under-removal error resulting from the four procedures as a function

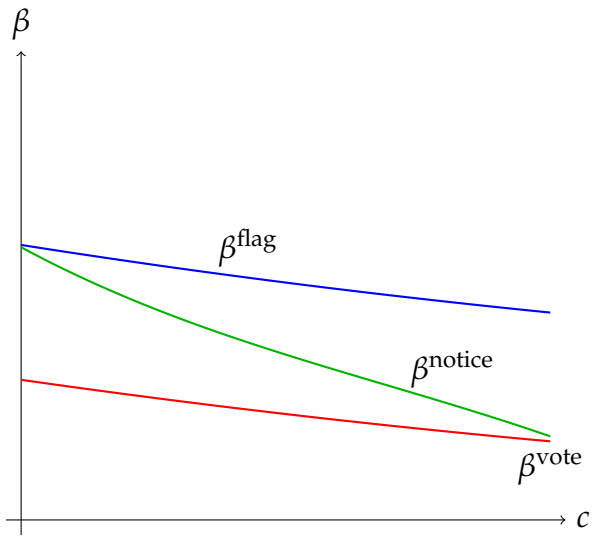


Figure 12: Comparing Under-removal Error of Four Procedures as a Function of Procedural Cost

of the notice cost c. The same observation holds: for all c, the red line is below the green line, which is below the blue line.

If the inefficiency concern for content moderation is under-removal, the voting procedure is the best option.

Proposition 2. *For all (c, p), $\beta^{vote} \leqslant \beta^{notice} \leqslant \beta^{flag} \leqslant \beta^{cen}$. The voting procedure minimizes the under-removal error β among the four procedures.*

The voting procedure is superior because of *preference aggregation*. Voting aggregates preferences of all community members so that if $n_h > k$ the harmed group is the majority, decentralization solves the dispute completely for the (h, v) pairs, and decentralization is more efficient than the platform decision. This explains why Reddit and Quora prefers voting because under-removal is the primary concern for misleading knowledge and consensus is what they emphasize.

Centralization has both a larger type I error and a larger type II error, but it does have one advantage for efficiency: it saves the costs of the users. It is straightforward to show that centralization has a smaller expected cost than all the decentralized procedures.

Proposition 3. *For all (c, p), centralization is the least cost procedure among the four procedures.*

The comparison of the four procedures appears to be generalizable to more sophisticated procedures. Both counter-notice and voting increases users' participation. The counter-notice procedure prolongs the periods the disputed users can negotiate. The voting procedure involves more users in the moderation process. And increase in users' participation helps. Decentralization in the form of user participation ends up increasing efficiency and consumers' welfare considerably. They reduce either over-removal bias (type I error) or under-removal bias (type II error), and sometimes reduces both errors. User participation can resolve disputes before they reach the platform for a random decision, and the change in moderation outcome usually goes in the right direction of efficiency. For the mass of users (h, v) whose outcome is determined by the platform decision, decentralization improves the probability that they are going to get a rightful moderation outcome. For the mass of users (h, v) who are marginal in notice decisions, decentralization flips their moderation outcome from either a type I error or a type II error to an efficient one.

There is a range of possible procedures. User participation can be measured by various dimensions. It can be the number of periods a single user is involved in the moderation negotiation. It may also be the number of users involved in the moderation negotiation. A general procedure of decentralized content moderation can be described by $< c, p, k, T >$ where c is the cost of a notice, p is the probability of a platform removal, k is the number of users involved, T is the number of rounds. For bilateral disputes, $k = 2$ and the procedure is described by a triple $< c, p, T >$. For example, YouTube Content ID can be described as $k = 2, T = 3$. Wikipedia's Talk Page would be described as $k = 2, T = \infty$. An example of multi-lateral procedure would be Wikipedia dispute resolution noticeboard, which can be described as $k = n, T > 2$. Solving and comparing them one by one would be implausible. However, the method of mechanism design can help us determine what the optimal procedure is among all the possible procedures, which is the focus of the next section.

4 Mechanism Design

In this section, we employ the mechanism design approach to investigate the general properties that error-free and consumer-optimal moderation procedures necessarily have. This approach is useful in two aspects: First, by identifying efficient (error-free) and consumer-optimal mechanisms, the mechanism design approach will provide various policy implications on how to improve the existing moderation procedures. Second, this approach also allows us to analyze the incentives of plaintiffs and defendants without any reference to a particular moderation procedure, and hence provides more general and robust insights.

Here is the road map of this section. First, we define the environment we investigate with the mechanism design approach in Section 4.1. The environment is essentially identical to one in the last section; our primary goal would be to introduce some useful definitions and notations. Next, in Section 4.2, we investigate whether and how a mechanism induces the error-free procedure in the presence of asymmetric, or equivalently, the efficient moderation rule with agents' cost (disutility) incurred throughout the moderation process being disregarded. Next, we explore how a mechanism induces the consumer welfare allocation that maximizes the surplus from moderation process less the sum of total cost incurred by agents throughout the process.

4.1 Environment

The environment is essentially identical to the one discussed at the beginning of Section 3, and its main ingredients will be summarized briefly. For simplicity, our focus will be on the case with two agents (users): one plaintiff and one defendant. The key insights in this section easily extend to the case with multiple defendants and plaintiffs. We will analyze the general case with multiple agents on both sides in Section B of the appendix.

Agents and Payoffs. Consistent with the previous sections, we define a content moderation procedure as a process to determine whether to retain or remove disputed content

from the platform. We will represent the decision to keep the content in question as $a = 0$, and the decision to remove the content as $a = 1$.

There are two users, a plaintiff (denoted by subscript p) and a defendant (denoted by subscript d). Each user's *net payoff* from each moderation outcome $a = 0$ or $a = 1$ is given by

$$u_p(a, c_p; h) = ah - c_p \quad \text{and} \quad u_d(a, c_d; v) = (1 - a)v - c_d \tag{17}$$

where $c_i \geqslant 0$ is user i's any cost (disutility) incurred throughout the moderation process, h and v are respectively each user's benefit from own favorable outcome. As in the previous sections, the true values of h and v are assumed private information (i.e., private types) of the plaintiff and defendant, respectively. For each $i = d$ and p, θ_i generally refers to the private type of each agent i. That is, θ_p represents the plaintiff's private type h, and θ_d represents the defendant's private type v. In accordance with the standard convention in the literature, θ_{-i} denotes the private type of the opponent of agent i.

While the realization of h and v is private information of each agent, the common prior beliefs regarding these values are given by the distribution functions F_h and F_d respectively. For each agent $i = p$ and d, F_i admits differentiable density functions f_i, where the hazard rates $f_i(\cdot)/[1 - F_i(\cdot)]$ are strictly increasing. We assume that both F_d and F_p have a common support denoted as $\Theta \equiv [\underline{\theta}, \overline{\theta}]$, where $\underline{\theta} = 0$. In other words, the lowest types, $h = \underline{\theta}$ and $v = \underline{\theta}$, are indifferent between keeping and removing the content.

Moderation Mechanism. Suppose that the platform has the ability to design every detail of the moderation process. The revelation principle allows us to reformulate any given moderation process as a direct-revelation (moderation) mechanism as follows: First, the plaintiff and defendant learn their private types (private information), denoted as h and v respectively. Then, the plaintiff and defendant confidentially report their private information directly to the mechanism; let $\tilde{h}$ and $\tilde{v}$ denote those reports, where h and v may not coincide with $\tilde{h}$ and $\tilde{v}$ in principle. Finally, based on these reports, the platform decides to remove the content under consideration with probability $P(\tilde{h}, \tilde{v}) \in [0, 1]$, and the plaintiff and defendant incur costs $(c_p, c_d) = (C_p(\tilde{h}, \tilde{v}), C_d(\tilde{h}, \tilde{v})) \equiv C(\tilde{h}, \tilde{v})$. We generically denote

a (direct-revelation) moderation mechanism as $\mu = (P, C)$, where P and C are referred to as the assignment rule and cost rule, respectively.

The platform's task is to design the allocation rule P and the cost rule $C = (C_p, C_d)$. The design of the allocation rule P amounts to the platform's setting and committing to the moderation rule (i.e., whether to keep or remove the content based on the private information reported by the plaintiff and defendant). Additionally, we assume that the platform has the ability to design the cost rule C by adjusting the specific details of the moderation procedure (e.g., expected duration of the moderation process, required amount of evidence for each user to submit, etc.) to control the level of costs incurred by each user during the process.

Again by the revelation principle, we can assume without loss of generality that the platform focuses on direct-revelation moderation mechanisms that satisfy two conditions: (i) the plaintiff and defendant find it optimal to truthfully report their types (incentive-compatibility), and (ii) their expected payoffs in the mechanism are non-negative (individual-rationality). A moderation mechanism $\mu = (P, C)$ will be referred to as *feasible* if it satisfies both incentive-compatibility and individual rationality. Any given allocation rule P is called *implementable* if there exists a cost rule C that makes (P, C) feasible. In such cases, we also say that P is *implementable with the cost rule* C. Let $\mathcal{M}$ represent the set of all feasible moderation mechanisms.

Again by the revelation principle, we may assume without loss that the platform restricts attention to direct-revelation moderation mechanisms such that (i) the plaintiff and defendant find it optimal to report their types truthfully (incentive-compatibility), and (ii) their expected payoffs in the mechanism is non-negative (individual-rationality). We will call a moderation mechanism $\mu = (P, C)$ *feasible* if it is incentive-compatible and individual rational. In addition, an allocation rule P will be called *implementable* if there is a cost rule C that renders (P, C) feasible; in this case, we will also say that P is *implementable with cost rule* C. Let $\mathcal{M}$ denote the set of all feasible moderation mechanisms.

The three components of a moderation mechanism, $P(h, v)$, $C_p(h, v)$, and $C_d(h, v)$ describe *ex post* assignment and cost realized after both v and h are reported to the mechanism. However, the plaintiff and defendant decides whether they reveal their type truth-

fully in the *interim* stage where the other user's report is still uncertain. Thus, to discuss each user's incentive to truthfully report own type to the mechanism, it is useful to define *interim* allocation and cost as follows:

$$\overline{P}_i(\theta_i) := \int_0^{\overline{\theta}} P(h,v)\mathrm{d}F_{-i}(\theta_{-i}), \qquad \overline{C}_i(\theta_i) := \int_0^{\overline{\theta}} C_i(\theta_i,\theta_{-i})\mathrm{d}F_{-i}(\theta_{-i}) \quad \forall i = d,p. \tag{18}$$

Additionally, it is also convenient to denote the probability of the mechanism keeping the content by $Q(h,v) := 1 - P(h,v)$, and furthermore, $\overline{Q}_i(\theta_i) := 1 - \overline{P}_i(\theta_i)$ for all θ_i and i.

Finally, one can easily show that a moderation mechanism $\mu = (P,C)$ is individual-rational and incentive-compatible if and only if the following four conditions jointly hold:

$$\overline{P}_p(h) \text{ and } \overline{Q}_d(v) \text{ are non-decreasing in } h \text{ and } v, \text{ respectively,} \tag{F$_1$}$$

$$\overline{C}_p(h) = h\overline{P}_p(h) - \int_0^h \overline{P}_p(x)\mathrm{d}x - \overline{C}_p(0) \quad \forall h \in [0,\overline{h}], \tag{F$_2$}$$

$$\overline{C}_d(v) = v\overline{Q}_d(h) - \int_0^v \overline{Q}_d(x)\mathrm{d}x - \overline{C}_d(0) \quad \forall v \in [0,\overline{v}], \tag{F$_3$}$$

$$\overline{C}_i(0) \leqslant 0 \quad \text{for both } i = p \text{ and } d. \tag{F$_4$}$$

The proof follows the standard lines, and hence, it is omitted here. See, for example, Börgers (2015, Chapter 3).

4.2 Error-free Mechanism

In this section, we investigate whether the *error-free* allocation rule (also called *efficient* allocation rule) such that

$$P^e(h,v) = \begin{cases} 1 & \text{if } h > v, \\ 0 & \text{if } h < v. \end{cases}$$

is implementable. $P^e(h,v)$ indeed dictates the "correct" decision in the sense that the mechanism removes the content if and only if the harm h from the content exceeds the value v of keeping it, thereby eliminating any possibility of both type-I and type-II errors.

Substituting $P(h,v) = P^e(h,v)$ in the definition of interim allocation $\overline{P}$, we obtain

$$\overline{P}^e_p(h) := \int_0^{\overline{v}} I\{h > v\} \mathrm{d}F_d(v) = F_d(h),$$
$$\overline{Q}^e_d(v) = 1 - \overline{P}^e_d(v) := 1 - \int_0^{\overline{h}} I\{h > v\} \mathrm{d}F_p(h) = F_p(v),$$

where I stands for the indicator function. Note that both $\overline{P}^e_p(h)$ and $\overline{Q}^e_d(v)$ are increasing in h and v, respectively, and hence, (F_1) holds with $\overline{P}_p = \overline{P}^e_p$ and $\overline{Q}_d = \overline{Q}^e_d$. Furthermore, substituting $\overline{P}_p = \overline{P}^e_p = F_d$ and $\overline{Q}_d = \overline{Q}^e_d = F_p$, we can rewrite ($F_2$)–($F_4$) as

$$\overline{C}_p(h) = hF_d(h) - \int_0^h F_d(x)\mathrm{d}x - \overline{C}_p(0) \quad \forall h \in [0,\overline{h}], \tag{19}$$
$$\overline{C}_d(v) = vF_p(v) - \int_0^v F_p(x)\mathrm{d}x - \overline{C}_d(0), \quad \forall v \in [0,\overline{v}], \tag{20}$$
$$\overline{C}_p(0) \leqslant 0 \quad \text{and} \quad \overline{C}_d(0) \leqslant 0. \tag{21}$$

Hence, the error-free allocation rule $P^e(h,v)$ is implementable by a (direct-revelation) moderation mechanism if and only if we can find the cost rule $C = (C_p, C_d)$ that satisfies (19)–(21). In fact, it is straightforward that all these conditions hold with $(C_p, C_d) = (C^e_p, C^e_d)$ where

$$C^e_i(\theta_i, \theta_{-i}) := \theta_i F_{-i}(\theta_i) - \int_0^{\theta_i} F_{-i}(x)\mathrm{d}x \quad \forall \theta_i \in \Theta, \quad \forall i \in \{p,d\}, \tag{22}$$

so that each agent's ex post cost is independent of the other player's type (i.e., independent of the other agent's report to the mechanism).

Proposition 4. *The error-free allocation rule P^e is implementable with the cost rule $C^e = (C^e_p, C^e_d)$ as stated in (22).*

A direct-revelation mechanism, such that all agents need to report their private information directly to the mechanism, can be challenging to operate in practice. Among various practical issues, one common obstacle is that even agents themselves often cannot explicitly quantify their harm h or benefit v from the content under consideration, making it difficult for them to submit a precise report to the mechanism. Thus, it is a practically

important question whether we can achieve the same outcome without asking the agents explicitly submit their private information to the mechanism.

In fact, there is a simple method to achieve the error-free allocation rule without directly referencing h and v. This procedure is similar to the counter-notice procedure discussed in Section 3.3 but with infinite horizon. Let's consider a "war of attrition" game between the plaintiff and defendant. Time is continuous and represented by $t \in [0, \infty)$. The moderation process begins at $t = 0$, and at any point in time $t \geqslant 0$, both agents simultaneously decide whether to continue arguing with the other agent or drop the case. The game ends when one user decides to drop the case, allowing the other agent to claim the right to keep or remove the content under consideration. At each time t before dropping the case, the plaintiff and defendant respectively incur the instantaneous cost

$$\gamma_i(t) := \rho(t)\frac{f_{-i}(\rho(t))\rho'(s)}{1 - F_{-i}(\rho(t))} \quad \forall i \in \{p, d\} \text{ and } \forall t \geqslant 0, \quad \text{where} \quad \rho(t) := \overline{\theta}[1 - e^{-t}]. \tag{23}$$

Thus, if an agent $i \in \{p, d\}$ decides to drop the case at time t, the agent's total realized cost is

$$c_i = \int_0^t \gamma_i(\rho(s))\mathrm{d}s = \int_0^t \rho(s)\frac{f_{-i}(\rho(s))}{1 - F_{-i}(\rho(s))}\mathrm{d}s.$$

If the plaintiff drops first, then the game ends with the outcome $a = 1$ (removal of the content) being implemented. On the other hand, if the defendant drops first, the outcome $a = 0$ is implemented. In either case, the plaintiff and defendant obtain their final payoffs as specified in (17).

Now, we show that the following strategy profile constitutes a subgame perfect equilibrium:conditional on the other agent not having dropped the case yet, each player $i \in \{p, d\}$ of type θ_i drops the case at time

$$T(\theta_i) = \log \frac{1}{1 - (\theta_i/\overline{\theta})}.$$

Note that $T(\cdot)$ is independent of i, and thus, the plaintiff and defendant with the same value of θ_i employ the same stopping time. Furthermore, T is increasing in θ_i, where its inverse is given by $\rho(\cdot)$ as defined in (23).

To prove that the stopping time T is indeed optimal for both agents, suppose that both agents have not dropped the case yet up to time $t \geqslant 0$. To compute the agent i's marginal benefit from continuing for another infinitesimal time Δ, note that the probability that the other agent will drop the case over the time interval $(t, t+\Delta)$ is

$$\Pr\left\{t < T(\theta_{-i}) < t+\Delta | T(\theta_{-i}) > t\right\} = \frac{\Pr\{t < T(\theta_{-i}) < t+\Delta\}}{\Pr\{T(\theta_{-i}) > t\}} = \int_t^{t+\Delta} \frac{f_{-i}\left(\rho(s)\right)\rho'(s)}{1-F_{-i}(\rho(t))}\mathrm{d}s$$

Thus, the agent i's marginal net benefit from continuing by an infinitesimal time Δ is

$$V_t(\Delta|\theta_i) := \theta_i \int_t^{t+\Delta} \frac{f_{-i}\left(\rho(s)\right)\rho'(s)}{1-F_{-i}(\rho(t))}\mathrm{d}s - \int_t^{t+\Delta} \gamma_i(s)\mathrm{d}s = \int_t^{t+\Delta} (\theta_i - \rho(s)) \frac{f_{-i}\left(\rho(s)\right)\rho'(s)}{1-F_{-i}(\rho(t))}\mathrm{d}s.$$

Differentiating this net benefit and then evaluating it at $\Delta = 0$,

$$V_t'(0|\theta_i) = [\theta_i - \rho(t)] \underbrace{\frac{f_{-i}(\rho(t))}{1-F_{-i}(\rho(t))}\rho'(t)}_{>0} \geqslant 0 \quad \Longleftrightarrow \quad \theta_i \geqslant \rho(t) \quad \Longleftrightarrow \quad t \leqslant T(\theta_i).$$

Thus, it is indeed optimal for the agent of type θ_i to stop at $t = T(\theta_i)$.

In the equilibrium identified above, the defendant drops before the plaintiff if and only if $T(h) < T(v)$, or equivalently, if and only if $h < v$.[19] Thus, the error-free allocation rule P^e is achieved in equilibrium. One can also show that each agent's cost realized in equilibrium coincides with (22); thus, the equilibrium outcome is exactly equivalent to the outcome of the error-free mechanism discussed in Proposition 4.[20]

Finally, note that the war of attrition game bears resemblance to the counter-notice procedure discussed in Section 3.3. In fact, we can interpret the war of attrition game as a version of the counter-notice procedure with an infinite number of rounds, where each instantaneous time t represents one round, and both agents have the option to make a notice or a counter-notice at each round. The intuition behind why this game implements the efficient allocation rule is straightforward. As the number of rounds increases, the platform can induce a more precise outcome, leading to the elimination of both type-I

[19] The event $h = v$ occurs with zero probability, and thus, we may disregard this case without any loss.

[20] By the revenue equivalence theorem, any mechanism that implements P^e necessarily induces the same level of cost for each agent.

and type-II errors.

The war of attrition game is very similar to the content moderation system used by YouTube. In essence, this platform has a content moderation policy based on the fact that users can flag the content -an algorithm can also do it based on the likelihood of the content being offensive- and every time the content is flagged, human reviewers are in charge of deciding whether to remove it or not based on its compliance with YouTube's guideline policies. If the content is offensive, it will not only be removed from the platform, but the users who have uploaded the content will receive a warning and, if they repeat their behavior, a strike. In this process there is the possibility of filing an appeal by the user whose content has been flagged. In short, it exemplifies how plaintiff and defendant would interact by sending notices and counter-notices of the content that affects them, like the war of attrition game.

Another example is Wikipedia with its so-called 'Talk Pages'. These discussion pages are portals where editors (users with editing rights) can discuss about improving the content of a given entry on the platform. For each interaction it is possible to submit a response from another authorized editor (counter-notice). The final result of those interactions leads to keeping the content as it was or incorporating the changes suggested by the different users.

4.3 Consumer Optimal Mechanism

Now let's shift our focus to the case that the platform aims to maximize the *(weighted) consumer surplus*, which is defined as the weighted sum of the expected payoffs for both the plaintiff and the defendant. As before, we can utilize the revelation principle and narrow our attention to feasible mechanisms in $\mathcal{M}$. In any feasible mechanism $\mu = (P, C)$, the plaintiff's and defendant's ex ante expected payoffs are respectively given by

$$\begin{aligned}
&E^{h\sim F_p, v\sim F_d}[u_p(P(h,v), C_p(h,v); h)] + E^{h\sim F_p, v\sim F_d}[u_d(P(h,v), C_p(h,v); v)] \\
&= \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} [hP(h,v) - C_p(h,v)] \mathrm{d}F_p(h) \mathrm{d}F_d(v) + \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} [vQ(h,v) - C_d(h,v)] \mathrm{d}F_p(h) \mathrm{d}F_d(v)
\end{aligned}$$

where the expectation operators are taken with respect to the distributions $h \sim F_p$ and $v \sim F_d$. Thus, we can write the platform's problem to identify the consumer optimal (incentive-compatible) mechanism as follows:

$$\max_{(P,C)} \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \begin{bmatrix} \lambda_p(hP(h,v) - C_p(h,v)) \\ +\lambda_d(vQ(h,v) - C_d(h,v)) \end{bmatrix} \mathrm{d}F_p(h)\mathrm{d}F_d(v) \tag{CS}$$

$$\text{subject to} \quad (\mathrm{F}_1)\text{–}(\mathrm{F}_4),$$

where $\lambda_p \in [0,1]$ and $\lambda_d \equiv 1 - \lambda_p \in [0,1]$ respectively stands for the weight to the plaintiff's and the defendant's expected payoffs in the consumer surplus. We will denote the solution of (CS) by $\mu^c = (P^c, C^c)$, where the superscript "c" is a shorthand for "consumer optimal," and call it the consumer optimal (moderation) mechanism.

Substituting $\overline{C}_p(h)$ in the objective function with (F_2) and then integrating by parts, we obtain

$$\begin{aligned}
\int_0^{\overline{\theta}} \int_0^{\overline{\theta}} [hP(h,v) - C_p(h)]\mathrm{d}F_p(h)\mathrm{d}F_d(v) &= \int_0^{\overline{\theta}} \left[\overline{P}_p(h)h - \overline{C}_p(h,v)\right] \mathrm{d}F_p(h) \\
&= \int_0^{\overline{\theta}} \left[\int_0^h \overline{P}_p(x)\mathrm{d}x\right] \mathrm{d}F_p(h) - \overline{C}_p(0) \\
&= \int_0^{\overline{\theta}} \left[\frac{1 - F_p(h)}{f_p(h)} \overline{P}_p(h)\right] \mathrm{d}F_p(h) - \overline{C}_p(0) \\
&= \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \left[\frac{1 - F_p(h)}{f_p(h)} P(h,v)\right] \mathrm{d}F_v(v)\mathrm{d}F_p(h) - \overline{C}_p(0).
\end{aligned}$$

Similarly,

$$\int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \left[vQ_d(v) - C_d(h,v)\right] \mathrm{d}F_d(v)\mathrm{d}F_p(h) = \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \left[\frac{1 - F_d(v)}{f_d(v)} Q(h,v)\right] \mathrm{d}F_d(v)\mathrm{d}F_h(v) - \overline{C}_d(0).$$

Thus, after relabeling $P(h,v)$ and $Q(h,v)$ in (CS) with $x_p(h,v)$ and $x_d(h,v)$ respectively,

we can reformulate the platform's problem as follows.

$$\max_{x_d, x_p, \overline{C}_d(0), \overline{C}_p(0)} \sum_{i=p,d} \left\{ \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \left[\lambda_i \frac{1 - F_i(\theta_i)}{f_i(\theta_i)} x_i(\theta_i, \theta_{-i}) \right] \mathrm{d}F_i(\theta_i) \mathrm{d}F_{-i}(\theta_{-i}) - \overline{C}_i(0) \right\} \tag{CS'}$$

$$\text{subject to} \quad \overline{C}_d(0) \geqslant 0, \quad \overline{C}_p(0) \geqslant 0, \quad 0 \leqslant x_p(h,v) = 1 - x_d(h,v) \leqslant 1 \quad \forall (h,v) \in \Theta^2,$$

$$\int_0^{\overline{\theta}} x_i(\theta_i, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i}) \text{ is non-decreasing in } \theta_i \quad \forall i = p, d.$$

The reformulated problem (CS') of the platform is formally equivalent to the problem of designing the revenue-maximizing auction (Myerson, 1981) with one exception: the virtual valuation in the auction design problem is here replaced by the weighted inverse hazard rate $\lambda_i(1 - F_i(\theta_i))/f_i(\theta)$.[21] Thus, in what follows, we can apply Myerson's approach to solve the platform's problem (CS').

Note that weighted inverse hazard rate $\lambda_i(1 - F_i(\theta_i))/f_i(\theta)$ in the objective function (CS') is not increasing in θ_i. In fact, as we assume that the hazard rate is increasing in θ_i, the weighted inverse hazard rate is decreasing over the entire domain $[0, \overline{\theta}]$. Thus, we need to take the standard "ironing" procedure (Myerson, 1981) to replace the weighted inverse hazard rate with its average value over $[0, \overline{\theta}]$:

$$\lambda_i \int_0^{\overline{\theta}} \frac{1 - F_i(\theta_i)}{f_i(\theta_i)} \mathrm{d}F_i(\theta_i) = \lambda_i \int_0^{\overline{\theta}} \theta_i f_i(\theta_i) \mathrm{d}\theta_i = \lambda_i E^{\theta_i \sim F_i}[\theta_i]$$

where the first equality is obtained by the integration by parts. Ultimately, we can convert the platform's problem to the following auxiliary maximization problem, whose solution

[21] In fact, the reformulated problem (CS') also differs from Myerson's problem in the sense that the sum of $x_p(h,v)$ and $x_d(h,v)$ must be always 1 (i.e., the right to decide whether to keep or remove the content must be granted among the plaintiff and the defendant) while the auctioneer in Myerson (1981) could hold back the item if all bids are low. However, this difference does not make any effect as the consumer-surplus-maximizing platform would not choose $x_d(h,v) + x_p(h,v) < 1$ even if allowed, because $\lambda_i(1 - F_i(\theta_i))/f_i(\theta)$ in the integrand of (CS') is non-negative at all θ_i.

coincides with the solution of the original problem: (CS′):

$$\max_{x_d, x_p, \overline{C}_d(0), \overline{C}_d(0)} \sum_{i=p,d} \left\{ \int_0^{\overline{\theta}} \int_0^{\overline{\theta}} \lambda_i E^{\theta_i \sim F_i}[\theta_i] x_i(\theta_i, \theta_{-i}) \mathrm{d}F_i(\theta_i) \mathrm{d}F_{-i}(\theta_{-i}) - \overline{C}_i(0) \right\} \tag{CS″}$$

$$\text{subject to} \quad \overline{C}_d(0) \geqslant 0, \quad \overline{C}_p(0) \geqslant 0, \quad 0 \leqslant x_p(h,v) = 1 - x_d(h,v) \leqslant 1 \quad \forall (h,v) \in \Theta^2,$$

$$\int_0^{\overline{\theta}} x_i(\theta_i, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i}) \text{ is non-decreasing in } \theta_i \quad \forall i = p, d.$$

The solution for the maximization problem (CS″) is straightforward: $\overline{C}_i(0) = 0$ for both $i = p$ and d, and

$$(x_p(h,v), x_d(h,v)) = (x_p^c(h,v), x_d^c(h,v)) = \begin{cases} (1,0) & \text{if } \lambda_p E^{h \sim F_p}[h] > \lambda_d E^{d \sim F_d}[v], \\ (0,1) & \text{if } \lambda_p E^{h \sim F_p}[h] \leqslant \lambda_d E^{d \sim F_d}[v]. \end{cases} \tag{24}$$

Note that $(x_p^c(h,v), x_d^c(h,v))$ maximizes the integrand of the objective function in (CS″) pointwise; furthermore, $(x_p^c(h,v), x_d^c(h,v))$ makes the integral in the second constraints in (24) constant for both $i = p$ and d, and thus, trivially non-decreasing. Thus, $(x_p^c(h,v), x_d^c(h,v))$ in (24) indeed solves (CS″).

By rewriting the optimal allocation rule (24) in terms of the allocation rule, we finally obtain the following consumer optimal mechanism:

$$\begin{aligned} P^c(h,v) &\equiv 1 \quad \forall (h,v) \qquad \text{if } \lambda_p E^{h \sim F_p}[h] > \lambda_d E^{d \sim F_d}[v], \\ P^c(h,v) &\equiv 0 \quad \forall (h,v) \qquad \text{if } \lambda_p E^{h \sim F_p}[h] \leqslant \lambda_d E^{d \sim F_d}[v]. \end{aligned}$$

In other words, the platform decides whether to remove the content based on the comparison between the weighted expected values of h and v: the platform removes content for sure if the weighted expected value of the plaintiff's type h is larger than the weighted expected value of the defendant's type v, and vice versa. Alternatively, we may interpret P^c as the consumer optimal moderation mechanism allocating the "exclusive right" to decide whether to keep or remove the contents to the agent with a larger weighted expected value of θ_i. Importantly, it does not depend on the realization of private types h and v which agent would win this right. The platform assigns the right only based on the ex

ante expected values of h and v.

Proposition 5. *The platform can maximizes the weighted consumer surplus by removing the content iff* $\lambda_p E^{h\sim F_p}[h] > \lambda_d E^{v\sim F_d}[v]$ *and keeping the content iff* $\lambda_p E^{h\sim F_p}[h] \leqslant \lambda_d E^{v\sim F_d}[v]$, *independent of the report of agents' private types.*

The primary motivation for the simple allocation rule P^c is to minimize the agent's costs that incur throughout the moderation process. Indeed, substituting $\overline{P}$ and $\overline{Q}$ in (F_2) and (F_3) with P^c and $Q^c = 1 - P^c$,[22] we can verify that the cost rule in the consumer optimal mechanism is zero

$$\overline{C}_p(h) = \overline{C}_d(v) = 0 \quad \forall h, v.$$

However, compared to the error-free allocation rule P^e, the consumer optimal allocation rule P^c suffers both higher type-I and type-II errors.

As discussed in the introductory section, the consumer-optimal procedure is essentially a zero-cost flagging procedure similar to spam filters used by several platforms. This automated systems are designed to flag the potentially unwanted or dangerous content.

This filtering technology was used by Twitter: the platform was notified of the dispute and it picked a winner based on the average value and harm.

Our analysis here sheds lights on the cost and the benefit of Twitter's zero-cost flagging procedure: compared to the error-free allocation, it increases the risk of both type-I and type-II error but minimizes the disutilities that consumers may incur during the moderation process. The platform could eliminate these type-I and type-II errors by implementing P^e instead, but it would hurt the consumer surplus as a whole as the implementation of P^e would require a large amount of cost for the agents.

[22] We also impose $\overline{C}_i(0) = 0$ for both $i = p$ and d without loss, which must be always the case in the consumer optimal mechanism; see the footnote **??**.

5 Conclusion

Content moderation has become an indispensable aspect of online platforms today, with different platforms experimenting with a wide range of methods. A central question in content moderation is how to design the procedures for moderating content. This paper provides theoretical insights into precisely this question and analyzes the efficiency of different procedures, leveraging the tools offered by game theory and mechanism design.

We begin by comparing the relative merits of the four popular moderation procedures: centralization, flagging, notice and counter-notice, and voting. Each procedure can be analyzed as a bargaining game between the plaintiff and the defendants. Counter-notice and voting reduce moderation errors because both of them partially decentralize the moderation decision to the users. Centralization is error-prone but cost-saving.

To identify the most efficient procedure among all possible ones, we apply the mechanism design theory to generalize our model of content moderation. We solve for two notions of social optimality: an error-free mechanism that implements the efficient allocation rule, and a consumer-optimal mechanism that maximizes the sum of the expected payoffs of the plaintiff and the defendant. The error-free mechanism is a repeated counter-notice procedure similar to Wikipedia's Talk Page. In the Talk Page, the editors of the content take turns responding until one of them concedes. A consumer-optimal mechanism is a zero-cost flagging procedure common in spam filters. In this case, once the platform is notified of controversial content, it decides whether to remove the content based purely on the average value and harm.

Our theoretical findings bear policy and managerial implications. For platforms that place a high value on veracity and accuracy, they have much to learn from Wikipedia. Associating every content with a structured help page for academic discussion between users would be a good example to follow. For platforms that pay more attention to cost-effectiveness, automated moderation predicts the value and harm of the new content from historical data would be preferred. Recent advances in machine learning that improve the prediction of the trained models will be valuable for such automation. Policymakers should bear in mind that there is no one-size-fits-all approach to content moderation.

Procedural details on how content is moderated are equally important as what content is being moderated. Decentralization, when implemented properly, can improve efficiency. Centralization, on the other hand, can be good for consumer welfare. Our paper is one of the first steps toward understanding when decentralization (vs. centralization) would be socially desirable.

We conclude our paper by discussing a few limitations and promising extensions. Among other issues not contemplated, there is the possibility that platforms compete to offer content moderation procedures. Which moderation procedure will emerge in platform competition and how the choice of moderation is affected by market structure in general would be an interesting topic for future research. The study of content moderation is a challenging problem of our time, as it significantly influences the quality, amount and validity of information accessible to everyone in the digital age.

A Some Procedures at work

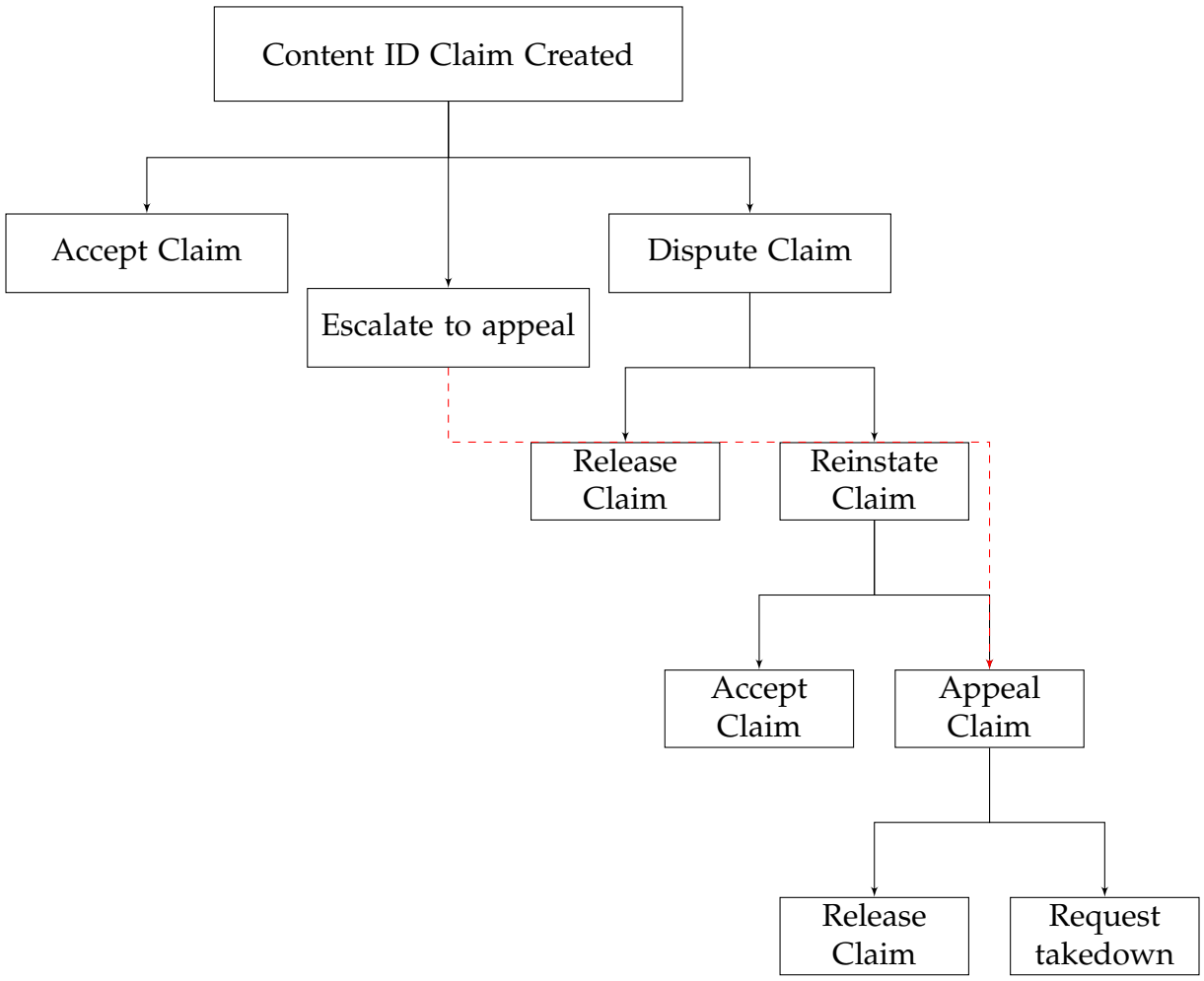


Figure 13: YouTube Content ID dispute and appeal process. Source: https://support.google.com/youtube/answer/2797454.

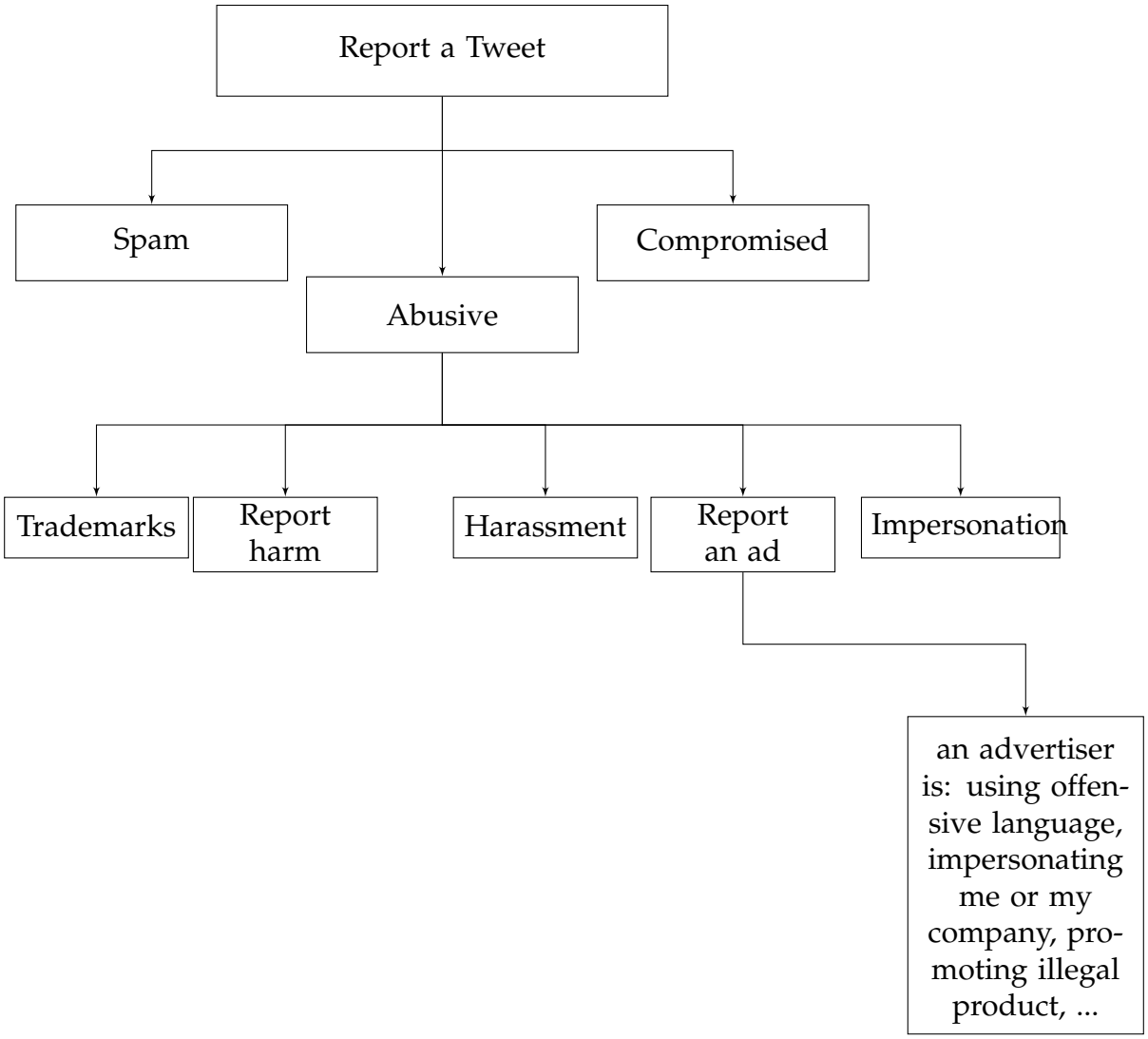


Figure 14: The process of reporting a tweet. Source: Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18(3), 410–428. https://doi.org/10.1177/1461444814543163.

B Mechanism Design with More Than Two Agents

In this section, we discuss how the results in Section 4 extend to the case with multiple plaintiffs and defendants.

B.1 Environment

Agents and Payoffs. There are $N_p \geqslant 2$ plaintiffs and $N_d \geqslant 2$ defendants, respectively indexed by $i = 1, 2, \ldots, N_p$ and $j = N_p + 1, N_p + 2, \ldots, N_p + N_d$. Let $N = N_p + N_d$ denote the total number of agents. Also, define $\mathcal{I}_p := \{1, 2, \ldots, N_p\}$ and $\mathcal{I}_d := \{N_p + k : k = 1, 2, \ldots, N_d\}$ as the sets of (indices for) plaintiffs and defendants, where $\mathcal{I} := \mathcal{I}_p \cup \mathcal{I}_d$.

We continue encoding the decisions to keep and remove the content in question with $a = 0$ and $a = 1$, respectively. Each agent's private type is denoted by θ_i, which we interpret a defendant's type (respectively, a plaintiff's type) as net benefit from the outcome $a = 0$ (respectively, net benefit from the outcome $a = 1$). In addition, c_i denotes each agent's cost (disutility) incurred throughout the moderation process. Overall, each agent i's final payoff from the moderation outcome is given as follows:

$$\begin{aligned} u_i(a, c_i; \theta_i) &= a\theta_i - c_i && \text{if } i \in \mathcal{I}_p, \\ u_i(a, c_i; \theta_i) &= (1 - a)\theta_i - c_i && \text{if } i \in \mathcal{I}_d. \end{aligned} \tag{25}$$

While the realization of each θ_i is private information of each agent, the common prior beliefs regarding these values are given by F_i, where all $\theta_1, \theta_2, \ldots, \theta_N$ are statistically independent. For each $i \in \mathcal{I}$, F_i admits differentiable density function f_i, where the hazard rates $f_i(\cdot)/[1 - F_i(\cdot)]$ are strictly increasing. We assume all F_i's have a common support denoted as $\Theta = [\underline{\theta}, \overline{\theta}]$, where $\underline{\theta} = 0$. Furthermore, we assume that $F_i = F_j = F_d$ for any two $i, j \in \mathcal{I}_p$ (i.e., F_i coincides across all plaintiffs), and also, $F_i = F_j = F_p$ for any two $i, j\mathcal{I}_p$. Let F_p and F_d denote the the common cumulative distribution functions for the plaintiffs and defendants, respectively, where $f_p := F_p'$ and $f_d := F_d'$.

Following the convention of the literature, $\theta_{-i} := (\theta_j)_{j \in \mathcal{I} \setminus \{i\}}$ generically denotes a

profile of private types of agent i's opponents. Let $F_{-i} : \Theta^{N-1} \to [0,1]$ and $f_{-i} : \Theta^{N-1} \to [0,\infty)$ denote the cumulative function and the density for θ_{-i}. In addition, $\theta = (\theta_j)_{j=1}^N$ generically denotes a profile of all agents' private types, where its cumulative distribution and density functions are denoted by $F : \Theta^N \to [0,1]$ and $f : \Theta^N \to [0,\infty)$, respectively.

Moderation Mechanism. A direct-revelation mechanism proceeds in the identical way as in Section 4. First, each agent i privately learns the realization of θ_i, and then, confidentially reports a message $\tilde{\theta}_i$ to the mechanism. Next, based on the reported types $\tilde{\theta} = (\tilde{\theta}_i)_{i=1}^N$, the platform removes the content with probability $P(\tilde{\theta}) \in [0,1]$ (equivalently, the platform keeps the content with probability $Q(\tilde{\theta}) = 1 - P(\tilde{\theta})$), and at the same time, each agent i incurs cost $c_i = C_i(\tilde{\theta})$.

As in the main text, we will generically denote a moderation mechanism as $\mu = (P, C)$, where P and $C = (C_1, \ldots, C_N)$ are referred to as the assignment rule and cost rule, respectively. By the revelation principle, we may focus on individual-rational and incentive-compatible mechanisms, also called *feasible* mechanisms. $\mathcal{M}$ denotes the set of all feasible mechanisms. Any given allocation rule P is called *implementable* (with the cost rule C) if there exists a cost rule C that makes (P, C) feasible.

We define $\overline{P}_i(\theta_i)$, $\overline{C}_i(\theta_i)$, and $\overline{Q}_i(\theta_i)$ similar to ones in the main text:

$$\overline{P}_i(\theta_i) := \int_{\theta_{-i} \in \Theta^{N-1}} P(\theta_i, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i}), \qquad \overline{Q}_i(\theta_i) := 1 - \overline{P}_i(\theta_i),$$
$$\overline{C}_i(\theta_i) := \int_{\theta_{-i} \in \Theta^{N-1}} C_i(\theta_i, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i})$$

for any $\theta_i \in \Theta$ and $i \in \mathcal{I}$. Then, $\mu = (P, C)$ is individual-rational and incentive-compatible if and only if the following four conditions, which are essentially identical to (F_1)–(F_4),

jointly hold:

$$\overline{P}_i(\theta_i) \text{ is non-decreasing in } \theta_i \text{ for any } i \in \mathcal{I}_p, \tag{$\widetilde{\mathrm{F}}_1$}$$

$$\overline{Q}_i(v) \text{ is non-decreasing in } \theta_i \text{ for any } i \in \mathcal{I}_d, \tag{$\widetilde{\mathrm{F}}_2$}$$

$$\overline{C}_i(\theta_i) = \theta_i \overline{P}_i(\theta_i) - \int_0^{\theta_i} \overline{P}_i(x)\mathrm{d}x - \overline{C}_i(0) \quad \text{for any } \theta_i \in \Theta \text{ and } i \in \mathcal{I}_p, \tag{$\widetilde{\mathrm{F}}_3$}$$

$$\overline{C}_i(\theta_i) = \theta_i \overline{Q}_i(\theta_i) - \int_0^{\theta_i} \overline{Q}_i(x)\mathrm{d}x - \overline{C}_i(0) \quad \text{for any } \theta_i \in \Theta \text{ and } i \in \mathcal{I}_d, \tag{$\widetilde{\mathrm{F}}_4$}$$

$$\overline{C}_i(0) \leqslant 0 \quad \text{for any } i \in \mathcal{I}. \tag{$\widetilde{\mathrm{F}}_5$}$$

B.2 Error-free Mechanism

With multiple plaintiffs and defendants, the error-free allocation rule generalizes to the following:

$$P^e(\theta_1, \ldots, \theta_N) = 1 - Q^e(\theta_1, \ldots, \theta_N) = \begin{cases} 1 & \text{if } \sum_{i \in \mathcal{I}_p} \theta_i > \sum_{i \in \mathcal{I}_d} \theta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Substituting $P = P^e$ in the definition of interim allocation $\overline{P}$, for any plaintiff $i \in \mathcal{I}_p$,

$$\overline{P}_i^e(\theta_i) = \int_{\theta_{-i} \in \Theta^{N-1}} I\left\{\theta_i > \sum_{j \in \mathcal{I}_d} \theta_j - \sum_{j \in \mathcal{I}_p \setminus \{\theta_i\}} \theta_j\right\} \mathrm{d}F_{-i}(\theta_{-i}) = G_p(\theta_i)$$

increases in θ_i, where G_p represents the cumulative distribution function of the random variable $\sum_{j \in \mathcal{I}_d} \theta_j - \sum_{j \in \mathcal{I}_p \setminus \{\theta_i\}} \theta_j$. Similarly, for any defendant $i \in \mathcal{I}_d$,

$$\overline{Q}_i^e(\theta_i) = \int_{\theta_{-i} \in \Theta^{N-1}} I\left\{\theta_i > \sum_{j \in \mathcal{I}_p} \theta_j - \sum_{j \in \mathcal{I}_d \setminus \{i\}} \theta_j\right\} \mathrm{d}F_{-i}(\theta_{-i}) = G_d(\theta_i)$$

also increases in θ_i, where G_d represents the cumulative distribution function of the random variable $\sum_{j \in \mathcal{I}_p} \theta_j - \sum_{j \in \mathcal{I}_d \setminus \{i\}} \theta_j$. Thus, by the argument essentially identical to the argument proceeding Proposition 4 in Section 4, we obtain the following proposition.

Proposition 6. *P^e is implementable with the cost rule $(C_i)_{i=1}^N = (C_i^e)_{i\in\mathcal{I}}$ such that*

$$C_i^e(\theta_i) = \begin{cases} \theta_i G_p(\theta_i) - \int_0^{\theta_i} G_p(x)dx & \text{for any } \theta_i \in \Theta \text{ and } i \in \mathcal{I}_p, \\ \theta_i G_d(\theta_i) - \int_0^{\theta_i} G_d(x)dx & \text{for any } \theta_i \in \Theta \text{ and } i \in \mathcal{I}_d. \end{cases}$$

The outcome of the error-free mechanism can be achieved through the following game. All agents $i \in \mathcal{I}$ simultaneously choose how much effort x_i to invest in arguing why the content in question should be kept or taken down. In the case of Wikipedia, x_i may represent the amount of effort that each editor invests in participating in the discussion on a Talk Page. Let each agent's x_i incur a disutility of $C_i^e(x_i)$, where $C_i^e(\cdot)$ is as defined in Proposition 6. In the end, the platform removes the content in question if and only if $\sum_{i\in\mathcal{I}_p} x_i > \sum_{i\in\mathcal{I}_d} x_i$; otherwise, the platform keeps the content. It is then straightforward to see that the outcome of the error-free mechanism is achieved in a Bayesian Nash equilibrium of this game.

The successful implementation of the error-free mechanism relies on the design of the cost function $C_i^e(\cdot)$. In principle, a platform can design $C_i^e(\cdot)$ by carefully manipulating the agents' user experience in the moderation process. For example, in the case of Wikipedia, the platform can make it more costly for each agent to participate in a Talk Page by manipulating the interface, which would result in a higher level of disutility $C_i^e(x_i)$ for each choice of x_i.

However, this is a challenging task. Firstly, the functional form of $C_i^e(\cdot)$ in Proposition 6 depends on the probability distributions of the agents' private types (i.e., F_p and F_d), which platforms often do not have direct access to. Secondly, even if a platform manages to obtain this information, it may not be clear exactly how a change in the agents' user experience ultimately affects their disutilities. Despite these challenges, a careful design of C_i^e is essential to implement the error-free outcome and to allow agents to credibly transfer relevant private information to platforms.

B.3 Consumer Optimal Mechanism

In this subsection, we will generalizes the results in Section 4.3 regarding the consumer optimal mechanism. Suppose that the platform's objective is to maximize the weighted consumer surplus:

$$\max_{\mu\in\mathcal{M}} \lambda_p \int_{\Theta^N} \sum_{i\in\mathcal{I}_p} [\theta_i P(\theta) - C_i(\theta)]\, \mathrm{d}F(\theta) + \lambda_d \int_{\Theta^N} \sum_{i\in\mathcal{I}_d} [\theta_i Q(\theta) - C_i(\theta)]\, \mathrm{d}F(\theta) \tag{$\widetilde{\mathrm{CS}}$}$$

where $0 \leqslant \lambda_p \leqslant 1$ and $\lambda_d = 1 - \lambda_p$ respectively stand for the weights to the plaintiffs' and the defendants' payoffs in the consumer welfare. The consumer optimal mechanism refers to the feasible mechanism that maximizes ($\widetilde{\mathrm{CS}}$), which we will denote by $\mu^c = (P^c, C^c)$.

For any feasible $\mu = (P, C)$, we can rewrite the first integral in ($\widetilde{\mathrm{CS}}$) as

$$\begin{aligned}
&\int_{\Theta^N} \sum_{i\in\mathcal{I}_p} [P(\theta)\theta_i - C_i(\theta)]\mathrm{d}F(\theta) = \sum_{i\in\mathcal{I}_p} \int_0^{\overline{\theta}} [\overline{P}(\theta_i)\theta_i - \overline{C}_i(\theta_i)]\mathrm{d}F_p(\theta_i) \\
&= \sum_{i\in\mathcal{I}_p} \left[\int_0^{\overline{\theta}} \int_0^{\theta_i} \overline{P}_i(\theta_i)\mathrm{d}\theta_i \mathrm{d}F_p(\theta_i) - \overline{C}_i(0) \right] = \sum_{i\in\mathcal{I}_p} \left[\int_0^{\overline{\theta}} \overline{P}_i(\theta_i) \frac{1 - F_p(\theta_i)}{f_p(\theta_i)} \mathrm{d}F_p(\theta_i) - \overline{C}_i(0) \right]
\end{aligned}$$

where the first follows the definition of $\overline{P}_i$ and $\overline{C}_i$; the second equality follows ($\widetilde{\mathrm{F}}_3$); the third equality follows the integration by parts. Similarly,

$$\int_{\Theta^N} \sum_{i\in\mathcal{I}_d} [\theta_i Q(\theta) - C_i(\theta)]\, \mathrm{d}F(\theta) = \sum_{i\in\mathcal{I}_d} \left[\int_0^{\overline{\theta}} \overline{Q}_i(\theta_i) \frac{1 - F_d(\theta_i)}{f_d(\theta_i)} \mathrm{d}F_d(\theta_i) - \overline{C}_i(0) \right].$$

Thus, the platform's problem ($\widetilde{\mathrm{CS}}$) can be reformulated as follows:

$$\begin{aligned}
\max_{P,Q,\overline{C}_i(0)} \quad & \int_{\Theta^N} \left[\lambda_p \sum_{i\in\mathcal{I}_p} \overline{P}_i(\theta_i) \frac{1 - F_p(\theta_i)}{f_p(\theta_i)} + \lambda_d \sum_{i\in\mathcal{I}_d} \overline{Q}_i(\theta) \frac{1 - F_d(\theta_i)}{f_d(\theta_i)} \right] \mathrm{d}F(\theta) \\
& - \sum_{i\in\mathcal{I}_p} \lambda_p \overline{C}_i(0) - \sum_{i\in\mathcal{I}_d} \lambda_d \overline{C}_i(0)
\end{aligned} \tag{$\widetilde{\mathrm{CS}}'$}$$

subject to

$$0 \leqslant P(\theta) = 1 - Q(\theta) \leqslant 0 \quad \forall \theta \in \Theta, \quad \overline{C}_i(0) \leqslant 0 \quad \forall i \in \mathcal{I},$$
$$\overline{P}_i(\theta_i) = \int_0^{\theta_i} P(x, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i}) \text{ is non-decreasing in } \theta_i \quad \forall i \in \mathcal{I}_p,$$
$$\overline{Q}_i(\theta_i) = \int_0^{\theta_i} Q(x, \theta_{-i}) \mathrm{d}F_{-i}(\theta_{-i}) \text{ is non-decreasing in } \theta_i \quad \forall i \in \mathcal{I}_d.$$

Next, we make a couple of further observations to simply ($\widetilde{CS}'$). First, the constraint $\overline{C}_i(0) \leqslant 0$ does not affect any other constraints, and thus, it is trivially optimal to set $\overline{C}_i(0) = 0$ for all i. Second, by the standard ironing argument, we may restrict attention to *constant allocation rules* such that $P(\theta) = 1 - Q(\theta)$ is constant for all $\theta \in \Theta^N$. To see this, consider any allocation rule $(P^\dagger, Q^\dagger)$ such that $(P, Q) = (P^\dagger, Q^\dagger)$ (combined with $\overline{C}_i(0) = 0$ for all i) satisfies all the constraints in ($\widetilde{CS}'$). Let

$$\overline{P}_i^\dagger(\theta_i) = \int_0^{\theta_i} P^\dagger(x, \theta_{-i}) \mathrm{d}F_{-i}(\theta_i) \quad \forall \theta_i \in \Theta \quad \forall i \in \mathcal{I}_p$$
$$\overline{Q}_i^\dagger(\theta_i) = \int_0^{\theta_i} Q^\dagger(x, \theta_{-i}) \mathrm{d}F_{-i}(\theta_i) \quad \forall \theta_i \in \Theta \quad \forall i \in \mathcal{I}_d$$

and

$$P_0^\dagger := \int_{\Theta^N} P^\dagger(\theta) \mathrm{d}F(\theta), \quad Q_0^\dagger := 1 - P_0^\dagger = \int_{\Theta^N} Q^\dagger(\theta) \mathrm{d}F(\theta)$$

denote the interim and ex ante allocations. Now, consider the case that the platform adopts the allocation rule $(P, Q) = (P_0^\dagger, Q_0^\dagger)$; that is, the platform removes the content in question with the constant probability $P_0^\dagger$ regardless of which messages that the agents submit. Clearly, $(P, Q) = (P_0^\dagger, Q_0^\dagger)$ satisfies all the constraints in ($\widetilde{CS}'$). To see that $(P, Q) = (P_0^\dagger, Q_0^\dagger)$ generates a larger consumer surplus compared to the case with $(P, Q) = (P^\dagger, Q^\dagger)$, note that

$$\begin{aligned}\int_0^{\overline{\theta}} \overline{P}_i^\dagger(\theta_i) \frac{1 - F_p(\theta_i)}{f_p(\theta_i)} \mathrm{d}F_p(\theta_i) &\leqslant \int_0^{\overline{\theta}} \overline{P}_i^\dagger(\theta_i) \mathrm{d}F_p(\theta_i) \cdot \int_0^{\overline{\theta}} \frac{1 - F_p(\theta_i)}{f_p(\theta_i)} \mathrm{d}F_p(\theta_i) \\ &= P_0^\dagger \cdot \int_0^{\overline{\theta}} \frac{1 - F_p(\theta_i)}{f_p(\theta_i)} \mathrm{d}F_p(\theta_i)\end{aligned}$$

where the first inequality holds due to Harris inequality (which is applicable in this case

because $\overline{P}_i^{\dagger}(\cdot)$ is non-decreasing and $[1-F_p(\cdot)]/f_p(\cdot)$ is non-increasing); the last equality holds by the definition of $P_0^{\dagger}$. Similarly,

$$\int_0^{\overline{\theta}} \overline{Q}_i^{\dagger}(\theta_i)\frac{1-F_d(\theta_i)}{f_d(\theta_i)}\mathrm{d}F_d(\theta_i) \leqslant Q_0^{\dagger}\cdot\int_0^{\overline{\theta}}\frac{1-F_d(\theta_i)}{f_d(\theta_i)}\mathrm{d}F_d(\theta_i).$$

Thus, the objective function of ($\widetilde{CS}'$) is larger when it is evaluated at $(P,Q)=(P_0^{\dagger},Q_0^{\dagger})$, compared to the case that it is evaluated at $(P,Q)=(P^{\dagger},Q^{\dagger})$.

Finally, it is straightforward that, among all constant allocations rules, the following maximizes the consumer welfare:

$$\begin{aligned}
P(\theta)=1 \quad \forall\theta\in\Theta^N \quad \text{if } \lambda_p\sum_{i\in\mathcal{I}_p}\int_0^{\overline{\theta}}\frac{1-F_p(\theta_i)}{f_p(\theta_i)}\mathrm{d}F_p(\theta_i) > \lambda_d\sum_{i\in\mathcal{I}_d}\int_0^{\overline{\theta}}\frac{1-F_d(\theta_i)}{f_d(\theta_i)}\mathrm{d}F_d(\theta_i),\\
P(\theta)=0 \quad \forall\theta\in\Theta^N \quad \text{if } \lambda_p\sum_{i\in\mathcal{I}_p}\int_0^{\overline{\theta}}\frac{1-F_p(\theta_i)}{f_p(\theta_i)}\mathrm{d}F_p(\theta_i) \leqslant \lambda_d\sum_{i\in\mathcal{I}_d}\int_0^{\overline{\theta}}\frac{1-F_d(\theta_i)}{f_d(\theta_i)}\mathrm{d}F_d(\theta_i),
\end{aligned}$$

or equivalently,

$$\begin{aligned}
P(\theta)=1 \quad \forall\theta\in\Theta^N \quad \text{if } \lambda_p\sum_{i\in\mathcal{I}_p}E^{\theta_i\sim F_p}[\theta_i] > \lambda_d\sum_{i\in\mathcal{I}_d}E^{\theta_i\sim F_d}[\theta_i],\\
P(\theta)=0 \quad \forall\theta\in\Theta^N \quad \text{if } \lambda_p\sum_{i\in\mathcal{I}_p}E^{\theta_i\sim F_p}[\theta_i] \leqslant \lambda_d\sum_{i\in\mathcal{I}_d}E^{\theta_i\sim F_d}[\theta_i].
\end{aligned}$$

In other words, the consumer optimal mechanism removes the content (regardless of the agents' report) if and only if the weighted sum of the plaintiffs' expected harms is larger than the weighted sum of the defendants' values from the content. The following proposition, which generalizes Proposition 5 summarizes the discussion so far.

Proposition 7. *The platform can maximizes the weighted consumer surplus by removing the content if and only if* $\lambda_p\sum_{i\in\mathcal{I}_p}E^{\theta_i\sim F_p}[\theta_i] > \lambda_d\sum_{i\in\mathcal{I}_d}E^{\theta_i\sim F_d}[\theta_i]$ *and keeping the content if and only if* $\lambda_p\sum_{i\in\mathcal{I}_p}E^{\theta_i\sim F_p}[\theta_i] \leqslant \lambda_d\sum_{i\in\mathcal{I}_d}E^{\theta_i\sim F_d}[\theta_i]$*, independent of the report of agents' private types.*

We omit the discussion about the intuition for Proposition 7 which is essentially identical to the intuition for Proposition 5.

References

Bolton, G., Breuer, K., Greiner, B., and Ockenfels, A. (2023). Fixing feedback revision rules in online markets. *Journal of Economics & Management Strategy*, 32(2):247–256.

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

Börgers, T. (2015). *An Introduction to the Theory of Mechanism Design*. Oxford University Press.

Clark, J., Faris, R., Gasser, U., Holland, A., Ross, H., and Tilton, C. (2019). Content and conduct: How english wikipedia moderates harmful speech. *The Social Science Research Network Electronic Paper Collection, Research Publication*, (2019-5).

Crawford, K. and Gillespie, T. (2016). What is a flag for? social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18(3):410–428.

De Chiara, A., Manna, E., Rubí-Puig, A., and Segura-Moreiras, A. (2021). Efficient copyright filters for online hosting platforms. *Available at SSRN 3945130*.

Douek, E. (2022). Content moderation as systems thinking. *Harv. L. Rev.*, 136:526.

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

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros Fernandez, A., Roberts, S. T., Sinnreich, A., and Myers West, S. (2023). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4).

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, 28:1.

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law & Technology*, 17:42.

Grimmelmann, J. and Zhang, P. (2023). An economic model of intermediary liability. *Berkeley Technology Law Journal, Forthcoming*.

Hua, X. (2021). Holding platforms liable. *HKUST Business School Research Paper*, (2021-048).

Jemielniak, D. (2014). *Common Knowledge?: An Ethnography of Wikipedia*. Stanford University Press.

Jiménez Durán, R. (2022). The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN*.

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131:1598.

Kwan, A. P., Yang, S. A., and Zhang, A. H. (2023). Crowd-judging on two-sided platforms: An analysis of in-group bias. *Management Science*.

Lee, W. K. and Cui, Y. (2023). Should gig platforms decentralize dispute resolution? *Manufacturing & Service Operations Management*.

Liu, Y., Yildirim, P., and Zhang, Z. J. (2022). Implications of revenue models and technology for content moderation strategies. *Marketing Science*.

Madio, L. and Quinn, M. (2021). Content moderation and advertising in social media platforms. *Available at SSRN 3551103*.

Myerson, R. B. (1981). Optimal auction design. *Mathmatcis of Operations Research*, 6:58–73.

Papanastasiou, Y., Yang, S. A., and Zhang, A. H. (2023). Improving dispute resolution in two-sided platforms: The case of review blackmail. *Management Science*.

Urban, J. M., Karaganis, J., and Schofield, B. L. (2017). Notice and takedown: Online service provider and rightsholder accounts of everyday practice. *Journal of the Copyright Society of the USA*, 64:371.

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

Zhang, P. (2021). Piracy or fair use? evaluating the welfare of software copyright takedowns: Theory and evidence from github. *Available at SSRN 4274527*.