CESifo
AREA CONFERENCES 2020

Economics of Digitization

Munich, 19–20 November 2020

User-generated content, strategic moderation, and advertising

Leonardo Madio and Martin Quinn

CESifo

User-generated content, strategic moderation, and advertising*

Leonardo Madio[1†] Martin Quinn[2,‡]

[1]University of Padova,

[2]Católica Lisbon School of Business & Economics.

This Version: November 2020

Social networks act as "attention brokers" and stimulate the production of user-generated content to increase user activity on a platform. When ads are displayed in unsuitable environments (e.g., disputed material), advertisers may face a backlash. This article studies the incentive for an ad-funded platform to invest in content moderation and its impact on market outcome. We find that if moderation costs are sufficiently small (large), the ad price is U-shaped (decreasing) in brand risks and the optimal content moderation always increases (is inverted U-shaped). When platforms compete for user attention, content moderation decreases as competition intensifies and this constitutes a market failure. Finally, well-intended policy measures, such as taxation of platform ad revenues, alter incentives to invest in content moderation and this might lead to the spread of harmful content.

Keywords: Advertising; content moderation; user-generated content; platforms.

JEL Classification: L82; L86; M3.

*The authors thank Luis Abreu, Malin Arve, Luca Ferrari, David Henriques, Laurent Linnemer, Christian Peukert, Carlo Reggiani, Patrick Waelbroeck for helpful comments on a previous draft. We are also grateful to seminar participants in Lisbon, Paris Saclay, Telecom ParisTech, UK OFCOM, at the Workshop on Platforms E-commerce and Digital Economics (CREST, 2019), at the Conference on Auctions, Competition, Regulation, and Public Policy (Lancaster, 2019), the 17th ZEW ICT Conference (Mannheim, 2019), the EARIE (Barcelona, 2019), the Giorgio Rota Best Paper Award Conference (Centro Einaudi, Turin, 2020), online SIEP Conference (2020). Leonardo acknowledges financial support from the "MOVE-IN Louvain" Incoming Fellowship Programme and the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 670494). The usual disclaimer applies.

†University of Padova, Department of Economics and Management, Via del Santo, 33, 35123 Padova, Italy. Email: `leonardo.madio@unipd.it`. Other Affiliations: CESifo Research Network

‡Católica Lisbon School of Business & Economics, Palma de Cima, 1649-023 Lisboa; Email: martinquinn@ucp.pt. Other affiliations: Chair *Value and Policies of Personal Information (CVPIP).*

1 Introduction

Online activities represent nowadays an essential part of citizens' life. In 2018 alone, Internet users spent 2.8 million years online, and most of this traffic (33% of the total time spent online) was generated by social media accounts (GlobalWebIndex, 2019). Social media websites such as Facebook, YouTube, Instagram, Snapchat, TikTok, and many others, act as "attention brokers": they encourage users to spend more time online and monetize their attention with advertisements (ads). The more time spent on a social media website, the higher the number of profitable interactions with advertisers, the higher the platform's profit.

Advertisers' exposure on these platforms is not risk-free. As most content is generated or uploaded by users, it lacks external and professional validation (Allcott and Gentzkow, 2017) and can possibly be illegal or harmful. Recent estimates suggest that approximately 4-10% of display advertising does not meet brand safety requirements and the majority of content can be classified at a moderate risk level (Plum, 2019).[1] The recent story of social media platforms is full of stories and scandals, which cast doubts about platforms' moderation policies. For example, in June 2020, several influential brands and advertisers, ranging from Adidas to BestBuy, from Unilever to Coca-Cola, started boycotting - pulling their ads from - Facebook for its failure to create a safe environment for advertisers.[2]

Facebook was not the only platform experiencing such protests. Between 2017 and 2019, YouTube went through the so-called "The Adpocalypse". Big advertisers such as Clorox, Disney, Epic Games, Hasbro, McDonald's, Nestlé, PepsiCo, Walmart, Starbucks, AT&T, Verizon, Volkswagen appeared just next to inappropriate user-generated content, such as racist, extremist, and unsafe content.[3] Subsequently, they suspended their marketing campaigns: some reduced their ad expenditure up to 70%; others, instead, returned to the platform after a temporary pullback. The reason was distinctly expressed by the Association of National Advertisers: because of such scandals, "*reputation [...] can be damaged or severely disrupted*".[4]

To contain the scandals, YouTube intervened by tightening its moderation policy, shutting

[1] SmartyAds defines brand safety as "*the set of measures that aim to protect the brand's image from the negative or harmful influence of inappropriate or questionable content on the publisher's site where the ad impression is served*". https://smartyads.com/glossary/brand-safety-definition

[2] See TheNewYorkTimes, 'The Brands Pulling Ads From Facebook Over Hate Speech' https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html

[3] See Fandom, 'YouTube Adpocalypse' https://youtube.fandom.com/wiki/YouTubeAdpocalypse. See also Digitalcontentnext, March 31 , 2017, 'A timeline of the YouTube brand safety debacle': https://digitalcontentnext.org/blog/2017/03/31/timeline-youtube-brand-safety-debacle/

[4] See 'Statement from ANA CEO on Suspending Advertising on YouTube', March 24, 2017: https://www.ana.net/blogs/show/id/mm-blog-2017-03-statement-from-ana-ceo

down 400 channels (including popular YouTubers such as PewDiePie), and removing thousands of comments and videos. This intervention was part of a new program launched by YouTube in 2017 to allow the monetization of *advertiser-friendly* content only.[5] Other platforms, like Facebook and Instagram, followed suit, In November 2019, Facebook announced a "brand safety" tool for advertisers and, in May 2020, the creation of an independent body - Oversight Board - to decide which content should be allowed to remain on the platform.[6]

This article explores the incentives of platforms to invest in content moderation and its interlink with the price that advertisers pay to reach final users. When content is not manifestly unlawful (e.g., hate speech, illegal content, whose presence may make the platform liable), a platform faces a challenging trade-off. On the one hand, the platform may want to invest in content moderation to create a safer environment for advertisers. This is because as the risk of being associated with unsafe content decreases with stronger moderation enforcement, advertisers' willingness to pay increases and the platform can extract more revenues. On the other hand, the platform may want to safeguard individuals' fundamental freedom of speech, and please users not willing to be monitored. This may increase the risk advertisers face of being displayed next to unsafe content, but it also allows them to reach a large audience. For instance, recent evidence showed that Tumblr, Yahoo's micro-blogging social network acquired by Verizon and later sold to WordPress, once with a high tolerance for not-safe-for-work (NSFW) content, lost nearly 30% traffic after banning porn in late 2018, and almost 99% of its market value. The ban was designed to keep "*content that is not brand-safe away from ads*".[7] This *see-saw* effect between advertiser and user preferences for content moderation highlights the interesting trade-off we described above.

In a nutshell, we find that ad prices and moderation strategy are critically interlinked and depend on the size of the moderation costs. When the cost of moderating content

[5]See e.g., https://support.google.com/youtube/answer/9194476

[6]In February 2019, Dune, Marks and Spencer, the Post Office and the British Heart Foundation charity experienced brand safety issues with Instagram as their ads appeared next to self-harm and suicide videos. See e.g., BBC, 'Facebook' sorry' for distressing suicide posts on Instagram', January 23, 2019 https://www.bbc.com/news/uk-46976753. To tackle the problem, Facebook and Instagram increased content moderation efforts. For instance, Facebook claimed actions on 3.4 million content, including terrorist propaganda, graphic violence, adult nudity, and sexual activity, hate speech, and fake accounts in the first quarter of 2018. See Facebook Community Standards Enforcement Preliminary Report, 2018. In November 2019, Facebook announced a partnership with Integral Ad Science, OpenSlate and Zefr to help advertisers create a list of possibly sensitive videos.

[7]In other cases, such as YouTube, strict regulation on cannabis and firearm-related content fuelled new niche platforms such as TheWeedTube.com and Full30.com. See The Verge, March 14, 2019, 'After the porn ban, Tumblr users have ditched the platform as promised': https://www.theverge.com/2019/3/14/18266013/tumblr-porn-ban-lost-users-down-traffic. See also Leafbuyer, 'The Road to Becoming a Weedtuber Isn't Easy', November 10, 2018: https://www.leafbuyer.com/blog/weedtube/

is sufficiently low, the platform always increases its moderation effort if advertiser sensitiveness to brand risk increases. Notwithstanding, the ad price is U-shaped in the brand risk and the highest price is set for very high or minimal level of brand risk. The reason is that when brand risk is small, advertisers care more about the market reach and, because the number of consumers is the highest, the platform can also set a very high price. On the contrary, when brand risk is very high, the platform prefers to moderate all content and charge more advertisers for the high safety ensured. For intermediate values, the interplay between user aversion to moderation and advertisers' preferences for safe content leads to lower ad prices.

The relevance of moderation costs in shaping platform behaviour also emerges when these costs are high enough, e.g., small entrant platforms which may face significantly high cost for moderating content due to scarcity of data accumulated in the past or lack of state-of-the-art equipment. Likewise, it can also be the case of language barriers or when the differentiation between manifestly unlawful content and not-manifestly unlawful - but still harmful for advertisers - content becomes narrow. In these cases, content moderation has an inverted U-shaped relationship: it initially increases up to the point in which moderation becomes so costly that the platform finds it optimal to disinvest. Differently, the ad price always decreases with brand risk as the platform fails to accommodate both advertisers' safety concerns and users' preferences for less moderation. These results hold both in the presence of a platform monopoly and under a Hotelling setup with horizontally differentiated platforms.

Our analysis builds on a two-sided market model in which a platform (i.e., a social media website) provides meaningful interactions between Internet users (who consume online content) and advertisers.[8] Users join the platform free of charge, while advertisers pay an ad price to the platform. There are two types of content the platform hosts: safe and unsafe ones. The first type always benefits users and advertisers, e.g., funny videos, pets, informative content. The second type can have some controversial effects: these contents can be valuable for (some) users while entailing a negative externality on advertisers. In other words, the presence of unsafe content creates "brand safety" issues for advertisers. We model the presence of brand safety issues in terms of the net value that advertisers obtain from joining a platform with a certain amount of unsafe content. However, the platform can indirectly control the presence and virality of unsafe content by investing in content moderation, such as hiring human content moderators and investing in monitoring and AI-based tools. The stricter a platform content moderation policy, the lower the share of inappropriate content, the smaller the

[8]See the pioneering works on two-sided markets of Rochet and Tirole (2003); Armstrong (2006). For a comprehensive discussion on the advertising-financed business model, see Anderson et al. (2016).

brand risk advertisers may face.

We begin with studying the strategies of a monopolist platform and results hold in a very general setting. A natural variation of our model is to consider how platform competition influences the incentives to invest in content moderation. We therefore present a simple yet general model of competition between two symmetric and horizontally differentiated platforms that compete for user attention. In such a scenario, we find that as platforms become more substitutable from the consumer perspective (e.g., more intense competition, lower switching costs), platforms react diminishing their content moderation effort and increasing or reducing the price advertisers pay to place their ads. The rationale is that as competition intensifies, the marginal users become more valuable for the platform and these can be attracted by relaxing content moderation and reducing the nuisance from ad impressions. If content moderation is sufficiently costly, platforms prefer to be more lenient with unsafe content and this strategy is complemented by an increase of the ad price, which further benefits consumers by reducing the number of ads. On the contrary, if content moderation is not very expensive, the platform competes more aggressively by reducing content moderation and compensates advertisers for the higher brand risk by granting them a discount on the ad price. Our results suggest that absent regulatory tools or changes in the current platform liability regimes, stimulating more competition in the market may lead platforms not to internalize fully the negative externalities linked to unsafe content. As a result, competition would introduce a distortion and configure a market failure.

In Section 4, we provide several variants of our model. Above all, we study the effect of a tax on ad revenues on the platform's optimal content moderation policy. In 2019, France adopted the so-called "GAFA tax", whereas the 2018's Nobel Prize laureate in economics put forward a proposal to tax digital ads "*to protect and restore this public commons*" in light of dangerous misinformation and hate speech circulating on social media platforms.[9] Specifically, while one may imagine that the introduction of an ad tax would be directly passed onto advertisers, we find twofold effects. First, as the ad tax reduces platform's marginal gains from moderation, it then also reduces its investment effort in content moderation. Second, it can lead to a higher or lower price than in an environment with tax-free ads. The reason is the *first-order* pass-through effect of the tax onto the ad price is complemented by a *second-order* effect that compensates advertisers for the increased brand risk. Depending on the prevailing effect, which is linked to the cost function's convexity, the ad price can either increase or decrease.

[9]New York Times. 'A Tax That Could Fix Big Tech', by Paul Romer. May 7, 2019. https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html

Related Literature. This study contributes to the scant literature on user-generated content (UGC). Most of this literature features UGC as a media problem (Yildirim et al., 2013; Zhang and Sarvary, 2014; Luca, 2015; de Corniere and Sarvary, 2020) and concerns the media outlet's provision of news and other types of content. Other studies in the marketing literature look at UGC in the form of online reviews and their impact on sales (Chevalier and Mayzlin, 2006; Chintagunta et al., 2010; Proserpio and Zervas, 2017; Chevalier et al., 2018). This literature falls short of explaining the possible side-effects of UGC on advertisers. Instead, this paper studies how brand safety influences advertisers' behavior and shows that heterogeneity in advertiser aversions to brand-risk has significant consequences for the platform optimal content moderation and ad prices.

We also add to the literature on advertising and media, which has, so far, addressed different types of questions, such as the different types of ads displayed to users (Anderson and De Palma, 2013), targeting technologies and matching (Bergemann and Bonatti, 2011; Peitz and Reisinger, 2015), overlaps in the customer base and homing decision (Ambrus et al., 2016; Athey et al., 2016; Anderson et al., 2017), ad-avoidance (Anderson and Gans, 2011; Johnson, 2013), and more generally to the media *see-saw* (Anderson and Peitz, 2020). The ad-targeting literature is perhaps the closest to the spirit of our study. This literature generally assumes a better match between the user preferences and the advertiser type. This way, the likelihood of wasteful advertising campaigns is reduced, and each customer becomes a proper market. In this article, instead, targeting is not customer-specific. Investments in moderation allow a platform to decide which segment to serve and, as a result, it attracts users and advertisers more favorable to the type of content hosted by the platform.

Moreover, this article bears some similarities with the literature on media bias, which has mainly dealt with news bias originated in the supply side or the demand side of the market. The former deals with a bias originated by advertisers, political orientations, government pressures, and lobbies (see e.g., Ellman and Germano 2009; Besley and Prat 2006). The latter depends on beliefs of targeted audiences (see e.g., Gentzkow and Shapiro 2006; Mullainathan and Shleifer 2005; Xiang and Sarvary 2007; Gal-Or et al. 2012). A major feature of this literature is that a content provider decides about the distortion of the news.[10] Our approach differs from it in at least two dimensions. First, a platform acts as a content aggregator. This implies that it is not directly involved in content creation and in choosing the direction of the bias. On the contrary, it chooses which sides of the market to please the most. Second, the platform can gain control over a content only by exercising costly moderation effort. To this end, it trades-off the benefits of ensuring a higher brand safety to advertisers with a costly effort and

[10]For a review, see e.g., Gentzkow et al. 2015.

a potential demand contraction on the user side. This way, the platform decision can entail either a supply-side or demand-side bias depending on its moderation effort.

The above aspects allow us to differentiate this contribution from that of some closely related studies on media bias. For instance, Van Long et al. (2019) studied competition on content quality (real or fake news) between media outlets and found that competition increases user polarisation. Although this underlines how content providers tailor their material and bias their news, the paper does not feature advertisers' preferences and UGC. Ellman and Germano (2009) investigated media bias in a market in which platforms sell content to readers and profit from advertisers. In their framework, platforms are allowed to influence the accuracy of news. Such a lever can have a significant effect as inaccuracy in the reporting of violent or shocking news may allow the platform to generate a better match with ads. Our article underlines a similar mechanism when considering the impact of UGC on platform's profits. In this case, the platform might influence that match by moderating content more or less carefully.

In the framework of media bias, Mullainathan and Shleifer (2005) found that when newspapers compete for user demand, there is an incentive to exaggerate media bias. Similarly to ours, Gal-Or et al. (2012) studie the competition between ad-based media outlets in the presence of heterogeneous readers and endogenous homing decisions of advertisers. Although our mechanism is reminiscent of theirs, they show that when a media outlet relies on ad revenues, there are more incentives to moderate content as this results in a higher ad price. In this way, advertisers multihome and attract moderate readers. However, the authors also show that when advertisers singlehome, newspapers become a bottleneck, and competition intensifies. This results in more slanting to soften competition and stronger polarization of readers. In our model, instead, when competition intensifies, the platform becomes more tolerant with unsafe content and the number of impressions users are exposed to decreases.

Finally, recent empirical studies support our results and show how different platforms engage in different moderation policies. For instance, Chiou and Tucker (2018) studied Facebook's decision in 2016 to ban ads linking to external websites fabricating fake news. They found that the ban was effective: fake news declined more on Facebook than on Twitter after the policy. Rao (2018) documented the effectiveness of the US Federal Trade Commission enforcement on fake news websites, showing that when these websites were shut down, consumer interest for fake news declined and was displaced by the interest for regular advertisements. Their study alongside Allcott et al. (2019) motivated our analysis on platform heterogeneity in moderation policies. They showed that Facebook was more prone than Twitter in banning fake and false news, underlying platform heterogeneity. Our setup rationalizes the decisions of platforms not to moderate

borderline content that might contain fake news and how these ultimately depend on moderation costs. Notably, such decisions have a key role in the economics of social media as influence the pricing strategies advertisers are subject to.

Outline. The article unfolds as follows. In Section 2, we present a fairly general model with a platform monopolist. The effect of platform competition on content moderation is studied in Section 3. In Section 4, we present a number of extensions. Section 5 provides concluding remarks and policy implications.

2 The Model

Consider a platform environment in which an online intermediary (e.g., social media website) connects users and advertisers. Users consume UGC available on the platform, and their attention is catered to advertisers. For simplicity, let us assume that users only consume UGC and do not engage in their production.[11] Such an assumption can be justified by the fact that a few very popular content creators generate typically viral content (e.g., popular YouTubers, influencers on Instagram) and there is a long-tail of unpopular creators with a little number of views.[12]

Users consume two types of content: a mass 1 of safe content and a mass $\theta(m)$ of unsafe content. The former, which identifies professional videos and news, pictures of vacations and pets, entails positive benefits for both users and advertisers. For advertisers, one can imagine a positive match value when impressions are just next to this type of content. The latter, instead, identifies controversial and possibly harmful material, e.g., borderline comments which users want to protect in light of their freedom of speech but can create brand safety issues for advertisers. The mass of this content depends on the moderation policy the platform enforces and that is denoted by $m \in [0, 1]$, with $\theta(0) = 1$ and $\theta(1) = 0$. When $m = 0$, there is a unit mass of unsafe content (and hence 50% of the entire platform content is potentially harmful), whereas with $m = 1$ the platform moderates all unsafe content.[13]

[11] We discuss this assumption in Section 4.

[12] For instance, on YouTube, content creators can only monetize views when reaching at least 1,000 subscribers and have streamed at least 4000 hours in the last 12 months. See e.g., YouTube 'Additional changes to YouTube partner' https://youtube-creators.googleblog.com/2018/01/additional-changes-to-youtube-partner.html

[13] Moderation can be ex-ante or ex-post. When ex-ante, for instance, all content must be validated and approved by a moderator. When ex-post concerns moderation performed after the content has circulated. Content moderation can have type-I and type-II errors, thereby leading to removal of genuine content and errors in moderating harmful content. The study of these effects would not change the main trade-off faced by the platform.

The platform. There is an ad-funded platform that charges advertisers, acting on behalf of brands, for launching an ad campaign. We assume that advertisers do not compete for an ad space and they launch at most one ad campaign. We denote the number of advertisers joining the platform by $a(m, p)$, where p is the price per ad slot. The platform chooses both the ad price p and its content moderation policy m. We assume that the cost of moderating content is sufficiently convex, such that $C'(m) > 0$, $C''(m) > 0$, and $C(0) = 0$. While it can be argued that there are economies of scale, one must consider that moderation can be increasingly challenging when the content type to be monitored becomes larger. To see why, consider a very mild content moderation policy that only checks whether a content promotes terrorism. In this case, content moderation may require a certain degree of investment. However, if the platform wants to a enforce a much stricter moderation policy, also including conspiracy theories and borderline comments - for which categorization can require more effort and capabilities than with manifestly harmful content - then costs are likely to be much larger as requiring additional investments in text analysis. Similarly, while AI tools and filters based on tags and keywords can have benefits, some content may require ex-post human moderation, thereby leading to much higher costs. All these costs are taken into account by the platform when choosing ad prices and content moderation policies. The platform's profits as defined as follows:

$$\Pi = a(p, m)p - C(m). \tag{1}$$

Internet users. There is a unit mass of Internet users. Each user is identified by the duple (u, ϕ) that captures her taste for "safe", u, and "unsafe" content, ϕ, with preferences distributed in the following intervals $u \in [0, \overline{u}]$, $\overline{u} > 0$ and $\phi \in [\underline{\phi}, \overline{\phi}]$. We do not specify the sign of $\underline{\phi}$ and $\overline{\phi}$. Note that, as advertisers benefit from content moderation, when on average users benefit from unsafe content there is a see-saw and, hence, their interests are misaligned, *ceteris paribus*. Moreover, we also assume that users dislike ads and perceive them as nuisance, with $\gamma > 0$ identifying the nuisance cost. Indeed, when joining the platform, a user identified by the duple (u, ϕ) derives the following utility:

$$U = u + \phi\theta(m) - \gamma a. \tag{2}$$

Advertisers. There is a unit mass of advertisers identified according to a duple (v, λ) that captures the preference for "safe", v, and "unsafe" content, λ on this side of the market. We assume that these preferences are distributed in the following intervals $v \in [0, \overline{v}]$, $\overline{v} > 0$ and $\lambda \in [0, \overline{\lambda}]$, $\overline{\lambda} > 0$. A higher λ implies a high brand risk for advertisers (i.e., luxury brands), whereas a higher v implies a large utility from the presence of safe

content. This affects what we call the platform's long-term profitability and which we denote by $\Omega = 1 \times v - \lambda\theta(m)$. In other words, for long-term profitability, we capture the impact of brand reputation, which - according to a recent study - can contribute to a significant share of a firm's value (Jovanovic, 2020). [14] For short-term profitability, we instead refer to the benefit the platform receives when interacting with users n, generating a stream of revenues rn. This is a measure of the cross-side network externalities. In turn, we can define the utility of an advertiser (v, λ) as follows:

$$V = \Omega + rn - p = v - \lambda\theta(m) + rn - p. \tag{3}$$

Timing. The timing of the game is as follows. In the first stage, the platform maximizes both ad price and content moderation policy. In the second stage, users choose whether to visit the platform and advertisers decide whether to place their ad. These decisions are made simultaneously and we assume that users and advertisers have fulfilled expectations on the number of participants on the opposite side of the market. The equilibrium concept is the Subgame Perfect Nash Equilibrium.

2.1 Optimal content moderation

We first compute the level of activity on the platform. Following Rochet and Tirole (2003) and using equations (2-3), the number of users joining the platform can be written as $n = \Pr(U \geq 0)$ and the number of advertisers placing their ads as $a = \Pr(V \geq 0)$. Formally, this implies

$$\begin{aligned} a &= \Pr(v - \lambda\theta(m) + rn - p \geq 0) \equiv D^a(p, n, m), \\ n &= \Pr(u - \gamma a + \phi\theta(m) \geq 0) \equiv D^n(a, m), \end{aligned} \tag{4}$$

Assuming that the above system of equations admits a unique solution that defines a and n depending on (p, m) such that $a = d^a(p, m)$ and $n = d^n(p, m)$.[15] In the first stage, the platform chooses m and p to maximize $\Pi = d^a(p, m)p - C(m)$.

[14] Note that this very general specification captures the large heterogeneity across advertisers' benefit from being displayed just next to safe/unsafe content. For instance, a large λ may represent advertisers promoting luxury goods or charities, that would have a lot to lose when associated with extreme content (i.e., $\frac{\partial\Omega}{\partial m}$ is strongly negative). On the contrary, unsafe content can have a small impact on advertisers promoting gambling websites (i.e., a small λ).

[15] For more details, see Rochet and Tirole (2003) and Appendix A.

We denote by Ψ the *elasticity of profit with respect to moderation* such that

$$\Psi = \underbrace{\frac{\partial D^a}{\partial m}}_{\text{Brand safety effect}} + \overbrace{\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}}^{\text{Eyeball effect}}. \tag{5}$$

The first term highlights the positive effects that more content moderation has on advertisers as this increases brand safety and, indeed, their willingness-to-pay. We refer to it as the *brand safety effect*. The second term identifies the indirect yet effect that content moderation has on advertisers once it is channeled by the interaction between advertisers and users. We refer to it as the *eyeball effect*. This effect depends on the role that ϕ has and, hence, whether users dislike or like on average moderation. Two different cases may arise. A first case emerges when users derive on average a positive utility from content moderation, therefore aligning their interests with those of the advertisers. Hence, the platform always has a positive incentive to moderate content. A second situation arises when instead users dislike content moderation. An interesting trade-off occurs under such circumstances, as user' preferences strikes with advertisers' brand safety concerns. Hence, the platform acts as a manager of the users and advertisers' participation in the market to maximize profits.

In the rest of the paper, we focus on the latter case. To that end, we assume for the rest of the paper that users dislike content moderation on average.

Assumption 1. *Users' demand decreases with moderation:* $\frac{\partial D^n}{\partial m} < 0$.

Indeed, the above assumption implies that, as users dislike content moderation on average, the eyeball effect is negative. Consider the stage in which the the derive the following

$$\begin{aligned} p =& -\frac{d^a(p,m)}{\frac{\partial d^a}{\partial p}} = -\frac{d^a(p,m)(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})}{\frac{\partial D^a}{\partial p}}, \\ C'(m) =& p\frac{\partial d^a}{\partial m} = p\Big(\frac{\Psi}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big), \end{aligned} \tag{6}$$

and we can state the following proposition.

Proposition 1. *The optimal content moderation is implicitly defined by* (6)*. The platform marginal gain from moderation is defined as follows*

$$MR(m^*, p^*) = -\frac{d^a(m^*, p^*)\Psi}{\frac{\partial D^a}{\partial p}}. \tag{7}$$

The above proposition highlights that the optimal moderation policy is chosen in a way such that the marginal gains (revenues) from moderation equal the marginal costs. Due to the multi-sidedness of the market, the marginal gain from moderation accounts for the price the platform selects, the effect that content moderation has on advertiser demand, and how consumers and advertisers react to increased brand safety (via Ψ). As $\frac{\partial D^a}{\partial p} < 0$, the larger Ψ, the larger the gain from content moderation. On the contrary, in the limit case in which $\Psi \leq 0$, the platform prefers not to moderate content, by choosing $m^* = 0$.

To shed some further light about how the advertiser sensitiveness to brand risk impacts on equilibrium outcomes, we present some simple comparative statics. Specifically, we study how the equilibrium outcomes change with changes in parameter $\frac{\partial D^a}{\partial m}$, which proxies the average brand risk of advertisers in our model. As the latter is contained in Ψ, we provide comparative statics of p^* and m^* with respect to Ψ.[16] More specifically, one can verify that

$$\frac{\partial p^*}{\partial \Psi} = \frac{\Psi d^a(m^*, p^*) - \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m)}{\frac{\partial D^a}{\partial p}(2C''(m)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}})}.$$

$$\frac{\partial m^*}{\partial \Psi} = -\frac{2d^a(m^*, p^*) + \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}}{(2C''(m)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}})}.$$

The next proposition summarizes the main findings and highlights the relevance of moderation costs in shaping equilibrium outcomes.

Proposition 2. *There exists a cut-off*

$$\tilde{C} \equiv -2\Big(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}\Big)\frac{d^a(m^*, p^*)^2}{(\frac{\partial D^a}{\partial \Psi})^2\frac{\partial D^a}{\partial p}},$$

such that if moderation costs are sufficiently small ($C''(m^*) < \tilde{C}$), then p^* is U-shaped, whereas m^* is monotonically increasing in Ψ. Else, if moderation costs are sufficiently large ($C''(m^*) > \tilde{C}$), p^* is monotonically decreasing in Ψ, whereas m^* is inverted U-shaped in Ψ.

Proof. See Appendix A. □

The intuition behind the above proposition is the following. When moderation costs are sufficiently small, $C''(m^*) < \tilde{C}$, a platform can easily adjust its moderation effort depending on how many users and advertisers can attract. Therefore, as accommodating

[16] Denoting by $k := \frac{\partial D^a}{\partial m} \neq 0$ and assuming a function g such that $\frac{\partial g}{\partial k} \neq 0$. In turn, we have $\frac{\partial g}{\partial \Psi} = \frac{\frac{\partial g}{\partial k}}{\frac{\partial \Psi}{\partial k}} = \frac{\partial g}{\partial k}$.

advertisers' requests for brand safety is quite cheap, the optimal level of moderation provided by the platform monotonically increases with Ψ (i.e. when investing in moderation is more profitable), up to the point in which the brand risk is so high that full moderation is enforced. In this case, the brand safety effect largely outweighs the eyeball effect and, therefore, the platform prefers to have fewer users but more satisfied (and remunerative) advertisers. This situation matches the case of a platform being very meticulous because attracting luxury advertisers or charities, whose reputation loss from scandals might be substantial if associated with harmful content.

Things differ when moderation costs are sufficiently expensive, $C''(m^*) > \tilde{C}$. In this case, if advertiser brand risk increases, the platform faces increasingly high cost to accommodate the advertisers' request for more moderation and such costs are not compensated by the larger market reach on the consumer side. As a result, the moderation effort has a bell shape. It monotonically increases to the point where $-2d^a(m^*, p^*) = \frac{\Psi \frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}$, which underlines that accommodating advertisers' preferences becomes too expensive in terms of investments and the number of consumers leaving the platform because of moderation. Then, the moderation effort m^* starts decreasing with Ψ up to the point in which no content is moderated anymore, $m^* = 0$. In that case, as brand safety is high and there is no content moderation, no advertiser joins the platform and translates into a market failure as no trade takes place.

A similar discussion also applies to the non-monotonic effect of Ψ on ad prices. To see why, suppose moderation costs are sufficiently small such that it is quite cheap for the platform to accommodate any advertiser request. In this case, the platform sets relatively high prices for low and high values of Ψ and these correspond to when no moderation or full moderation is enforced. The reason is that at one extreme, the platform caters a very large number of user eyeballs to advertisers that do not perceive brand risk as a major issue (Ψ is quite low). As a result, the ad price is high. At the other extreme, the platform sacrifices some audience and meets the moderation requests of advertisers that - given their high willingness-to-pay for content moderation - also pay a very high price. For intermediate values of Ψ, that is, when the brand safety effect is not much more significant than the eyeball effect, the platform sets an intermediate level of moderation. This mild content moderation may feature controls of flags of some disputed content, such as 'hate speech', violence, sexual content, intellectual property rights infringements. The ad price reaches a minimum when $\Psi d^a(m^*, p^*) = \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m)$, and so the platform mediates the divergence between the two sides of the market by granting advertisers a price discount. In turn, p^* is convex in Ψ.

On the contrary, when moderation costs become too large. the optimal ad price is always decreasing in Ψ. The reason is twofold. First, when brand risk is sufficiently small, the

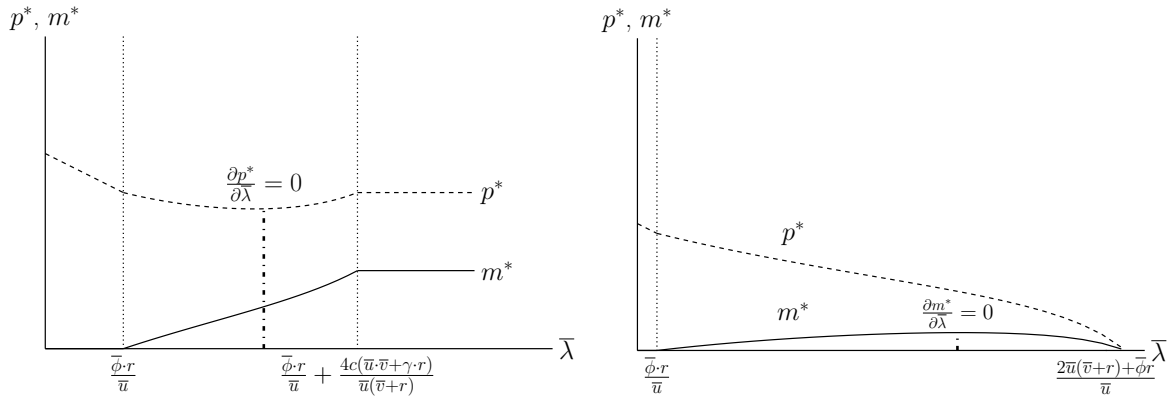

Figure 1: Example with a uniform distribution of preference: Effect of $\overline{\lambda}$ on p^* and m^* when c is small (left) and large (right)

platform increases moderation, but the way it increases does not compensate advertisers for a large number of users leaving the platform. So, the platform prefers to lower the price. When advertisers become too sensitive to unmoderated content, and hence Ψ gets quite large, the marginal revenues from increased moderation become lower than the marginal cost of moderation. It follows that the platform reduces its ad price again. In turn, the ad price decreases monotonically in Ψ.

The above discussion emerges prominently in Figure 1 - where advertiser and user preferences follow a uniform distribution (see Appendix B). The two figures present how the optimal price and content moderation react to advertisers' aversion to brand risk when moderating costs are small (left) and large (right).

3 Platform competition

Whereas platforms often exhibit forms of monopolization in their natural market, they also compete for user attention in several other markets. For instance, although their services can be regarded as sufficiently differentiated from the user perspective, YouTube competes with Facebook for user attention. In this section, we present a model of platform competition in the presence of full market coverage and study how the intensity of competition shapes market outcomes.

Platforms are located at the endpoints of a Hotelling-line of unit distance. Platform 1 is located at coordinate 0, whereas Platform 2 at coordinate 1. A platform i sets a price p_i with $i = 1, 2$ and $j \neq i$, for the entire ad campaign and a_i represents the number of advertisers deciding to buy a space on the website. Hence, platform i's profits are defined as follows

$$\Pi_i = a_i p_i - C(m_i).$$

Throughout the analysis, we assume that platforms are ex-ante symmetric and, given the multiplicity of equilibria that typically characterize markets with network effects, we look at an ex-post symmetric configuration.

On the advertiser side, consistently with the previous literature (Anderson et al., 2016), we let advertisers multihome. Namely, as there exists a competitive bottleneck, each platform becomes the only way to reach users. As in the presence of a monopolist, advertisers are defined by a duple (v, λ) and their utility when patronizing platform i is

$$V_i = v_i + rn_i - \lambda\theta(m_i) - p_i$$

On the user side, we assume a fully covered market. Users are are independently distributed on a line of unit length; they are identified by a duple relative to their relative preference for platform i, defined by their position y on a Hotelling line and by their aversion to moderation ϕ. Once again, we focus on the most interesting case in which a see-saw effect arises, that is, users are averse to moderation, whereas advertisers derive utility from it.[17] Indeed, the utility of a user located at y and joining platform i is as follows:

$$U_i = u_i + \phi\theta(m_i) - \gamma a_i + T_i(\tau, y)$$

where $T_1(\tau, y) = -\frac{\tau y}{2}$ and $T_2(\tau, y) = \frac{\tau y}{2}$, with $y \in \{\underline{y}, \overline{y}\}$ and $\overline{y} = -\underline{y}$, the user relative preference for platform 2.

In the first stage of the game, platforms compete by simultaneously and non-cooperatively choosing ad prices and content moderation policies. In the second stage of the game, advertisers decide whether to place an ad on both platforms or stay out of the market, whereas users decide which platform to join.

We can now solve the model by backward induction. The number of users who decide to join platform i is $n_i = \Pr(U_i \geq U_j)$, whereas the number of advertisers is $a_i = \Pr(V_i > 0)$. We assume that this system of equations admits a unique solution that defines a_i and n_i depending on

$$\begin{aligned} a_i &= d_i^a(p_i, m_i) \equiv D_i^a(p_i, p_j, m_i, m_j) \\ n_i &= d_i^n(p_i, p_j, m_i, m_j) \equiv D_i^n(p_i, p_j, m_i, m_j). \end{aligned} \tag{8}$$

As we assume full market coverage, then $n_j = 1 - n_i$ implies $D_j^n = 1 - D_i^n$, and by symmetry $\frac{\partial D_j^n}{\partial(.)} = -\frac{\partial D_i^n}{\partial(.)}$ and $\frac{\partial d_j^n}{\partial(.)} = -\frac{\partial d_i^n}{\partial(.)}$. We also require that the above system of equations admits a unique solution that defines a_i, a_j, n_i and n_j depending on

[17] Note that the same insights we found would also apply if users, on average, were to dislike content moderation.

(p_i, p_j, m_i, m_j) such that $a_i = d_i^a(p_i, p_j, m_i, m_j)$ and $n_i = d_i^n(p_i, p_j, m_i, m_j)$. In the first stage, platform i chooses m_i and p_i to maximize $\Pi_i = d^a(p_i, p_j, m_i, m_j)p_i - C(m_i)$.

We can now study how, in a symmetric setting, content moderation and prices impact on the demands of both sides of the market. While the proof is formally relegated to Appendix A, we shall note that platform's i elasticity to its own moderation effort changes when competition is introduced. As discussed in the presence of a platform monopolist, such a parameter played a key role in characterizing the incentive to invest in content moderation and shaping the price and moderation functions. Under competition, we denote the platforms' elasticity to moderation, Ψ_i, for $i = 1, 2$ as follows

$$\Psi_i = \frac{\partial D_i^a}{\partial m_i} + \frac{\partial D_i^a}{\partial n_i}\Big(\frac{\partial D_i^n}{\partial m_i} - \frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i}\Big).$$

Note that relative to when the platform enjoys a monopoly position, Ψ_i now also accounts for an additional term $-\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i} > 0$. As a result, the (negative) eyeball effect (under Assumption 1) can be mitigated compared to the monopolistic setup. This is because an increase in platform i's moderation level creates disaffection for some users, some of them move to platform j, which has now rooms to further increase its moderation level. Such an effect is then denoted by $-\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i} > 0$, given symmetry, and in turn it mitigates the magnitude of the negative eyeball effect. [18] Finally, consistently with Assumption 1, we assume that users tend to dislike moderation

Assumption 2. *Users' demand decrease with moderation, other things equal:* $\frac{\partial D_i^n}{\partial m_i} < 0$ *for* $i = 1, 2$.

By symmetry, the above assumption also implies that users demand at platform i increases with the moderation level at platform $j \neq i$.

We can now solve for the symmetric equilibrium. Optimal prices and content moderation are implicitly determined by the solution of the following system of equations:

$$\begin{aligned} p_i^* &= -\frac{d_i^a(p_i^*, p_j, m_i, m_j)}{\frac{\partial d_i^a}{\partial p_i}} = -\frac{d_i^a(p_i^*, p_j, m_i, m_j)(1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})}{\frac{\partial D_i^a}{\partial p_i}(1 - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})}, \\ C'(m_i^*) &= p_i\frac{\partial d_i^a}{\partial m_i} = p_i\frac{\Psi_i}{(1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})}. \end{aligned} \tag{9}$$

We can then state the following.

[18] Note that this is a by-product of the full market assumption, as disaffected users cannot walk away from the two platforms.

Proposition 3. *In a symmetric equilibrium, platforms choose content moderation and prices as in* (9)*. The platform i's marginal gain from moderation is defined as follows*

$$MR_i(m_i^*) = -\frac{d_i^a(p_i^*, p_j^*, m_i^*, m_j^*)\Psi_i}{\frac{\partial D_i^a}{\partial p_i}(1 - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})}.$$

While Proposition 3 bears some similarities with Proposition 1, it also exhibits differences. First, the negative effect of an ad price on advertisers' demand (at the denominator) is now exacerbated by the extent of the competition for user attention. In other words, more ads drive more consumers to another platform, this reduces the number of advertisers placing their ads on that platform, and therefore how much the platform can profit in this side of the market. Formally, this effect arises because $\frac{\partial D_i^a}{\partial p_i}$ is then scaled up by $(1 - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})$. Indeed, the denominator increases, thereby reducing the marginal gain and hence the incentives to invest in moderation, other things equal.

Second, as discussed, the shape of Ψ_i now accounts for an additional positive effect resulting from competition for user attention against a symmetric platform. An increase in platform i's moderation induces some of its users to join platform j, which increases advertisers' demand for platform j and in turn fosters users from platform j to join platform i. As a consequence, the number of ads might increase, driving up Ψ_i and hence the gains from moderation. Notably, it is possible to have a situation in which stricter content moderation attracts more users, because of competitive forces. Such an effect differs starkly from what we observed in the case of a platform monopolist, where an increase in moderation could simply prompt a reduction in user demand and hence advertisers, without opposite forces playing a role. Indeed, keeping advertisers' demand constant, competition might lead to higher or lower gains from moderation compared to the monopoly setting analyzed above, other things equal. In Appendix B, we show how these forces impact on equilibrium outcomes in the presence of uniform distribution of preferences.

To shed some further light on how platform competition changes incentives to moderate content and charge advertisers, in what follows we provide some comparative statics relative to the intensity of competition in the market, which in our case is proxied by the degree of platform differentiation. Starting from the equilibrium outcomes, implicitly determined by (9), we compute derivatives of m_i^* and p_i^* with respect to τ. When τ decreases, product differentiation reduces and, in turn, competition intensifies. For example, competition can intensifies because of lower barrier to entry or because of enforcement of policies aimed at making users more mobile and sensitive to marginal changes in platform strategies.

We find the following effects.

$$\begin{aligned}\frac{\partial p_i^*}{\partial \tau} &= \frac{d_i^a(p_i^*, p_j^*, m_i^*, m_j^*)\left(\left(2\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}\right)\frac{\partial \Upsilon}{\partial \tau} + \frac{\partial D_i^a}{\partial m_i}(1-2\Upsilon)\frac{\partial \Psi_i}{\partial \tau}\right)}{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)(\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}(2-3\Upsilon))} \\ \frac{\partial m_i^*}{\partial \tau} &= -\frac{d_i^a(p_i^*, p_j^*, m_i^*, m_j^*)\left((2-3\Upsilon)\frac{\partial \Psi_i}{\partial \tau} + 3\Psi_i\frac{\partial \Upsilon}{\partial \tau}\right)}{(1-\Upsilon)(\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}(2-3\Upsilon))},\end{aligned} \tag{10}$$

where $\Upsilon = \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}$. Once again, moderation costs appear to play a prominent role in determining the sign of the effect of more (or less) competition on ad prices. More specifically, we find that there exists a critical value which determines the sign of the partial effect on the equilibrium ad price.

Proposition 4. *There exists a critical value of moderation costs,*

$$\tilde{C} \equiv -\frac{\frac{\partial D_i^a}{\partial m_i}[(1-2\Upsilon)\frac{\partial \Psi_i}{\partial \tau} + 2\Psi_i\frac{\partial \Upsilon}{\partial \tau}]}{\frac{\partial D_i^a}{\partial p_i}\frac{\partial \Upsilon}{\partial \tau}},$$

such that fiercer leads to a price reduction if moderation costs are sufficiently small ($C''(m_i^*) < \tilde{C}$) and a price increase if moderation costs are sufficiently large ($C''(m_i^*) > \tilde{C}$). Regardless of moderation costs, fiercer market competition leads to less content moderation.

Proof. See Appendix A. □

Proposition 4 shows that when competition for users becomes fiercer, platforms have two ways to attract more users. On the one hand, they can relax their moderation policy and, hence, please users with a strong aversion to content moderation. On the other hand, they can reduce the number of ads and, therefore, the nuisance they are exposed to. In equilibrium, the mechanism works as follows. When moderation is sufficiently expensive, content moderation is already low and, as competition intensifies, platforms need to find another instrument to win the marginal user. Hence, the platform can induce exit of advertisers with low willingness-to-pay by increasing the ad price and mitigating the nuisance users face. On the contrary, when moderation is less expensive, the moderation policy is already quite strict. As competition intensifies, the platform prefers to reduce content moderation to attract more users and compensate advertisers for the sharp change in content moderation offering a price reduction. This, in turn, mitigates the advertisers' exit. These two forces are complements to reach the goal of attracting users when competition intensifies but, due to the symmetry of the market,

in equilibrium it does not bring about additional users and the platforms obtain equal market shares.

4 Discussion and extensions

4.1 Impact of policy tools: a tax on digital revenues

In recent years, several countries in Europe (e.g., France, Germany, Italy) have considered to tax on online revenues. More related to the aim of this paper, in 2019, the Nobel Prize laureate Paul Romer proposed to tax digital ads as a measure to induce social media platforms to limit misinformation. To shed some light on the possible unintended effects of such a policy, we modify our benchmark model. We assume that a platform monopolist is subject to an exogenously-imposed tax f on ad revenues, implying the Government raising af. The net profit of the platform is then equal to

$$\Pi = d^a(p, m)(p - f) - C(m).$$

As taxes impact the platform's marginal profits, we expect them to affect the price advertisers pay and accordingly, the content moderation decided by the platform. While the proof is formally relegated to the Appendix, we shall note that the introduction of an ad tax has a non-neutral effect on equilibrium outcomes in a way that:

$$\frac{\partial p^*}{\partial f} = \frac{\frac{\partial D^a}{\partial p} C''(m^*)(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2}{2\frac{\partial D^a}{\partial p} C''(m^*)(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2}$$

$$\frac{\partial m^*}{\partial f} = -\frac{\frac{\partial D^a}{\partial p}\Psi(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})^{-1}}{2C''(m^*)\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2},$$

where Ψ is defined as in (5). From the above expressions, we can conclude the following.

Proposition 5. *The optimal moderation policy always decreases with a fixed ad tax. There exists a critical value of moderation costs*

$$\tilde{C^*} \equiv -\frac{\Psi^2}{\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})},$$

such that if moderation costs are sufficiently large, $C''(m^*) > \tilde{C}$, an ad tax leads to an increase of ad prices. Else, for sufficiently small moderation costs, $C''(m^*) < \tilde{C}$, the introduction of an ad tax leads to a reduction in ad prices.

Proof. See Appendix A. □

Proposition 5 underlines very interesting results. First, the introduction of an ad tax always leads to less content moderation. This is because the tax directly reduces the marginal revenues from advertisers. Hence, the higher the tax, the lower the marginal gains from moderation and, consequently, the lower the incentive to moderate content.

Second, the ad price may increase or decrease with the tax depending on the cost of moderation. When the tax increases, a first-order effect drives the ad price up. However, a second-order effect implies a reduction in the platform moderation effort, which, in turn, decreases the ad price. One effect prevails over the other depending on moderation costs. When moderation costs are sufficiently low, the indirect effect dominates the direct one as content moderation decreases faster with a tax. As a result, advertisers are granted a price discount to be compensated for the high brand risk they face. When moderation costs are high enough, instead, the opposite holds. The direct effect dominates the indirect one and, therefore, advertisers pay a higher tax when an ad tax is introduced. To better understand the above mechanisms, in Appendix B, we provide an example with a uniform distribution of preferences.

4.2 Endogenous content creation

So far, we have assumed exogenous content creation. In this section, we relax this assumption and consider the case in which agents can also create content. This implies endogenizing the volume of both safe and unsafe content.

We explicitly model the presence of unsafe content creators among the users, who obtain utility $U_\theta = u_\theta + nk - m$ when creating content[19], with m being the platform moderation policy, u_θ his willingness to create an inappropriate content, and nk payoffs being the network effect from being exposed to n users on the platform. Such utility from content creation u_θ is be heterogeneous on the support $[0, \overline{u_\theta}]$ with $\overline{u_\theta} < 1 - k$ such that, if $m = 1$ (full moderation), all content creators make negative utility. Hence, the number of endogenously created content, θ, would be equal to $\theta = P(u_\theta + nk - m > 0)$. As in the benchmark model, the marginal gains (MR) from moderation are equal to

$$MR(m^*, p^*) = -\frac{d^a(m^*, p^*)\Psi}{\frac{\partial D^a}{\partial p}},$$

[19]Unsafe materials often generate virality. These can be any sensationalist or attention-grabbing content produced by creators in social networks and community platforms like Youtube, e.g., Conspiracy Theories, No-Vax comments, etc.

where

$$\Psi = \frac{\partial D^a}{\partial m}(1 - \frac{\partial D^n}{\partial \theta}\frac{\partial D^\theta}{\partial n}) + \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial \theta}\frac{\partial D^\theta}{\partial m}.$$

The above expression is an augmented version of the one presented in equation (5). However, it differs in the inclusion of indirect network externalities stemming from the presence of content creators and their interest in a broad audience base. This has an interesting and novel impact on the marginal gain from moderation. Specifically, an increase in content moderation implies an increase in the number of ads placed on the platform, which in turn leads some users to exit. A lower user demand reduced the incentives for content providers to post unsafe content on the platform, and this further drives down the demand of users with strong preferences for unsafe content.

Such an effect well suits the so-called "Tumblr spiral". After the acquisition of Yahoo from Verizon and the very stringent policy on content moderation to make the platform brand-safe, many creators and users decided to leave the platform, and the stock value of the former $1.1 billion-platform plunged to only $3 million - the price paid by Automatic, the owner of Wordpress. The above discussion rationalizes the effect of such a policy on Ψ and, hence, on the marginal gain from moderation.

4.3 Targeting

In the benchmark model, we have not considered the possibility that the platform(s) can target users and therefore operate ad hoc content moderation. Targeting can arise in different ways. First, ads can be targeted to users in a way that does not cause any distress and nuisance. This implies that the platform can possibly control the size of γ. If γ were considered equal to 0, such that ads were neutral to users, our main results would go through as in the benchmark model. An important difference, however, would be present when considering the case of competing platforms, that is, ad prices would always decrease with fiercer competition in the market. This is because when competition for users becomes more intense, the platform no longer needs to compete by reducing the number of ads (given $\gamma = 0$) via a price increase. In turn, more intense competition would only lead to a first-order effect on prices and, hence, a price reduction.

The second form of targeting can be related to better matching between advertisers and content. In this market, advertisers typically create lists of keywords they want (or do not) to be associated with. For instance, according to IAS Insider, the most blocked keywords by advertisers in November 2019 included "shooting, explosion, dead, bombs, etc".[20] This may ensure some forms of safeguard for brands and marketers. However,

[20]See IAS Insider, https://insider.integralads.com/the-20-most-blocked-keywords-in-november-2019/

targeting is far from perfect (Nielsen, 2018), and better precision may require investment costs which are very similar to the one used in our model. As the main trade-off remains unchanged, our model also encompasses a setup in which targeting is imperfect and for which higher investments in content moderation lead to higher brand safety for advertisers.

Finally, even if we assume a perfect targeting technology, the probability to incur reputation costs (due to word-of-mouth or negative buzz) might be too high not to moderate content at all. A reinforcing reason might emerge in assuming that users strongly favoring unsafe content are unmatchable with any advertisers. For example, users with very strong preferences for conspiracy theories would hardly be matched with advertisers, as the latter might not want to target, on average, these users. As a result, brand safety issues would continue to hold.

4.4 Other Applications

Offline news outlets. Our setting can offer insights regarding content moderation policies also arising in other markets. For instance, consider a (traditional) media outlet hosting content. Typically, these outlets have full control over the type of content they display. Such a practice differs from platforms that do not control content production. However, even professional content can feature a divergence between the interests of the users and those of the advertisers. Such a situation also relates to Ellman and Germano (2009) as it showed how media outlets may have incentives to bias the accuracy of news. In September 2016, following the online campaign "Stop Funding Hate" related to the presence of disputed content on migrants, several advertisers such as The Body Shop, Plusnet, Walkers, and many others announced that they would stop advertising on *The Daily Mail* and *The Sun*. Others, like the Co-operative Group, preferred to maintain their adverts as driving up sales.[21] Such a story well fits the trade-off that traditional media outlets may face when producing or reporting content.

Consider now an ad-funded news outlet that only produces professional content that is sufficiently attention-grabbing to be attractive for users. Hence, this outlet would strategically choose the sensitivity of materials to produce to balance user attraction and advertiser preferences. Whereas investments in content moderation are not needed in this case as there are no UGC available on the platform, content production may still be costly. The better (or, the more professional) the content, the higher the cost, the safer it can be for advertisers. However, one may imagine that producing professional

[21] See The Co-operative Group, 'An update on our advertising policy' https://blog.coop.co.uk/2017/03/23/an-update-on-our-advertising-policy/.

content is cheaper than moderating thousands of online user-generated content. Suppose these costs are sufficiently small, our framework suggests that more competition between outlets is likely to lower content quality and increase the price advertisers would have to pay.

Content aggregators. Our study can also provide applications for content aggregators that host both first-party (i.e., professional content) and third-party (i.e., UGC) content. In this case, the aggregator would directly balance user and advertiser preferences when choosing the type of content to produce and display to (safely) monetize users' eyeballs. Such a setup allows us to endogenize the platform's design choice that consists of accepting or not UGC to be displayed on the platform. Depending on moderation cost and production cost, an outlet may be keener on introducing UGC on the platform. For instance, a high-end fashion website may only attract advertisers with high brand safety, and not allow reviews and comments. In this case, we conjecture that when moderation cost is higher than production cost, such a website would prefer to only produce its content rather than moderating third-party UGC.

TV shows. The framework we depict can also be applied to TV reality shows, such as the famous *The Big Brother*. Oftentimes, these shows are sponsored by advertisers and feature the presence of a group of (unprofessional) contestants. While viewers might like some of the houseguests' scandals, which keep the reality game alive after years, this might not always be the case for advertisers that sponsor the program with their products. For instance, in Italy, in 2018, several different sponsors, including Nintendo, decided to give up their partnership with the TV show after bullying in the house.[22] A similar story also occurred in France, with advertisers boycotting a tv show because of sensitive content.[23] It follows that media producers are in the situation of balancing two starkly different preferences and decide how to moderate what is shown on TV and the price to charge to advertisers.

5 Main highlights and conclusions

The digital revolution has changed the production of media content. Whereas in the past, these were mostly produced by professionals (e.g., journalists), the advent of social

[22]Blitzquotidiamo.com, May 4, 2018, 'Grande Fratello, la grande fuga degli sponsor: niente acqua, shampoo e Nintendo': https://archivio.blitzquotidiano.it/tv/grande-fratello-fuga-sponsor-acqua-nintendo-2876635/

[23]L'Express.fr, October 10, 2019: https://www.lexpress.fr/actualite/medias/eric-zemmour-boycotte-par-des-annonceurs-sur-paris-premiere-et-bientot-sur-cnews$_2$102554.*html*

media websites has given users control over the production and diffusion of content. In most cases, this happens without any external and professional validation, which creates concerns among advertisers and marketers worldwide.

This article studied the trade-off faced by a social media platform when strategically enforcing content moderation and provided a rationale for the significant heterogeneity across platforms in tackling illegal, harmful, or disputed content. Our results underlined the importance of moderation costs in ad price and content moderation decisions. We found that the size of moderation costs can lead the platform to react differently to an increase in brands' aversion to risk - like the one presented in recent protests by advertisers. More importantly, the advertiser demand for more brand safety may not be supported by Big Tech if moderation costs are very large, e.g, because of a large amount of content to be checked or different languages to be considered.[24] Paradoxically, such an outcome is more likely to arise the more advertisers become concerned about brand risk, giving rise, in an extreme case, to a market failure in which no user joins the platform. On the contrary, it is in the interest of the platform to accommodate advertisers' requests if moderation costs are sufficiently small. The non-monotonicity of the platform's optimal choices of content moderation and prices, their link to brand safety issues, and the role of moderation costs may therefore help explaining heterogeneity across platforms in dealing with content moderation.

The above-described results are relevant not only for marketers but also for policymakers. Social media moderation policies are not neutral regarding market functioning and this article highlighted that the platform's choice depends on the trade-off between generating revenues from advertisers and capturing user attention. At different institutional levels, it is widely debated what platforms should do to prevent the diffusion of illegal content and misinformation going on social media websites as their effects could be detrimental to society, e.g., fake news impacting election outcomes (Allcott and Gentzkow, 2017) or leading to vaccine hesitancy (Carrieri et al., 2019). The European Commission recently issued a recommendation on how tackling effectively illegal content online (EU, 2018) stressing how platforms need to "*exercise a greater responsibility in content governance*" and, in 2020, it launched the Digital Services Act with plans to revise the EU E-Commerce Directive, change the liability regimes of online intermediaries, and regulate content moderation and algorithms.[25] Indeed, understanding platforms' strategies when dealing

[24] As shown in its moderation report, Facebook states facing moderation costs that are idiosyncratic to countries, depending on language, culture, and other characteristics. A summary of the report can be found here https://transparency.facebook.com/community-standards-enforcement.

[25] See Digital Single Market, 'Illegal content on online platforms': https://ec.europa.eu/digital-single-market/en/illegal-content-online-platforms. Similarly, see e.g., Thorsten Kaeseberg on VoxEu, December 12, 2019, 'Promoting competition in platform ecosystems': https://voxeu.org/article/promoting-competition-platform-ecosystems.

with harmful content becomes crucial to identify the best policy measures to adopt.

In this respect, we discussed more broadly how well-intended public policies may have unintended effect on moderation effort by platforms. First, we showed that more competition in digital markets might lead platforms to enforce a slacker content moderation. Typically, fostering more competition in the market is advocated by policymakers and regulatory agencies. For instance, this could translate in lowering barriers to entry, reducing switching costs, facilitating data portability, larger compatibility across platforms, or having non-exclusive access to essential inputs. Absent other interventions we therefore showed that a fiercer platform competition in the market are likely to generate negative societal externalities. Second, we studied the impact of an often advocated policy measure like the digital tax on advertising revenues. This was discussed in France, Germany, Italy, and recently supported by the Nobel Prize laureate Paul Romer. Finally, we verified that such well-intended measures may have the perverse effect of reducing moderation effort for the platform, thereby increasing the relevance of the current problem faced by democracies and advertisers.

References

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Ambrus, A., Calvano, E., and Reisinger, M. (2016). Either or both competition: A "two-sided" theory of advertising with overlapping viewerships. *American Economic Journal: Microeconomics*, 8(3):189–222.

Anderson, S. P. and De Palma, A. (2013). Shouting to be heard in advertising. *Management Science*, 59(7):1545–1556.

Anderson, S. P., Foros, Ø., and Kind, H. J. (2017). Competition for advertisers and for viewers in media markets. *The Economic Journal*, 128(608):34–54.

Anderson, S. P. and Gans, J. S. (2011). Platform siphoning: Ad-avoidance and media content. *American Economic Journal: Microeconomics*, 3(4):1–34.

Anderson, S. P. and Peitz, M. (2020). Media see-saws: winners and losers on media platforms. *Journal of Economic Theory.*, Volume 186.

Anderson, S. P., Waldfogel, J., and Stromberg, D. (2016). *Handbook of Media Economics, vol 1A*. Elsevier.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691.

Athey, S., Calvano, E., and Gans, J. S. (2016). The impact of consumer multi-homing on advertising markets and media competition. *Management Science*, 64(4):1574–1590.

Bergemann, D. and Bonatti, A. (2011). Targeting in advertising markets: Implications for offline versus online media. *The RAND Journal of Economics*, 42(3):417–443.

Besley, T. and Prat, A. (2006). Handcuffs for the grabbing hand? Media capture and government accountability. *American Economic Review*, 96(3):720–736.

Carrieri, V., Madio, L., and Principe, F. (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from italy. *Health Economics*, 28(1377-1382):417–443.

Chevalier, J. A., Dover, Y., and Mayzlin, D. (2018). Channels of impact: User reviews when quality is dynamic and managers respond. *Marketing Science*, 37(5):688–709.

Chevalier, J. A. and Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354.

Chintagunta, P. K., Gopinath, S., and Venkataraman, S. (2010). The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957.

Chiou, L. and Tucker, C. E. (2018). Fake news and advertising on social media: A study of the anti-vaccination movement. *NBER Working Paper No. 25223*.

de Corniere, A. and Sarvary, M. (2020). Social media and the news: Content bundling and news quality. *TSE Working Paper, n. 20-1152*.

Ellman, M. and Germano, F. (2009). What do the papers sell? As model of advertising and media bias. *The Economic Journal*, 119(537):680–704.

EU (2018). Commission recommendation of 1.3.2018 on measures to effectively tackle illegal content online. *C(2018) 1177 final*.

Gal-Or, E., Geylani, T., and Yildirim, T. P. (2012). The impact of advertising on media bias. *Journal of Marketing Research*, 49(1):92–99.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Gentzkow, M., Shapiro, J. M., and Stone, D. F. (2015). Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier.

GlobalWebIndex (2019). Social media flagship report. *Available at https://www.globalwebindex.com/hubfs/Downloads/Social-H2-2018-report.pdf*.

Johnson, J. P. (2013). Targeted advertising and advertising avoidance. *The RAND Journal of Economics*, 44(1):128–144.

Jovanovic, B. (2020). Product recalls and firm reputation. *American Economic Journal: Microeconomics*. Forthcoming.

Liu, Y.-H. (2018). The impact of consumer multi-homing behavior on ad prices: Evidence from an online marketplace. *Mimeo*.

Luca, M. (2015). User-generated content and social media. In *Handbook of Media Economics*, volume 1, pages 563–592. Elsevier.

Mullainathan, S. and Shleifer, A. (2005). The market for news. *American Economic Review*, 95(4):1031–1053.

Nielsen (2018). Nielsen digital ad ratings: Benchmarks and findings through 2h 2016, Europe.

Peitz, M. and Reisinger, M. (2015). The economics of Internet media. In *Handbook of Media Economics*, volume 1, pages 445–530. Elsevier.

Plum (2019). Online advertising in the UK. *Report commissioned by the UK Department of Digital, Culture, Media Sport*.

Proserpio, D. and Zervas, G. (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science*, 36(5):645–665.

Rao, A. (2018). Deceptive claims using fake news marketing: The impact on consumers. *Available at SSRN 3248770*.

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.

Van Long, N., Richardson, M., and Stähler, F. (2019). Media, fake news, and de-bunking. *Economic Record*, 95(310):312–324.

Xiang, Y. and Sarvary, M. (2007). News consumption and media bias. *Marketing Science*, 26(5):611–628.

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

Zhang, K. and Sarvary, M. (2014). Differentiation with user-generated content. *Management Science*, 61(4):898–914.

Appendix A

Proof of Proposition 1

Consider the utility of users and advertisers as defined by (2-3). The number of users joining the platform is $n = \Pr(U \geq 0)$, whereas the number of advertisers is $a = \Pr(V \geq 0)$. Following Rochet and Tirole (2003), the demands can be expressed as follows:

$$\begin{aligned} a &= \Pr(v - \lambda\theta(m) + rn - p \geq 0) \\ n &= \Pr(u - \gamma a + \phi\theta(m) \geq 0) \end{aligned}$$

Assume that the above system of equations admits a unique solution that defines a and n depending on (p, m) such that $a = d^a(p, m) \equiv D^a(p, m)$ and $n = d^n(m, p) \equiv D^n(m, p)$. We can solve the model in the first stage of the game: the platform chooses m and p to maximize $\Pi = d^a(p, m)p - C(m)$.

In what follows, we first look at how demands on both sides of the market change with ad prices and moderation. The derivatives of d^a and d^n with respect to p and m can be deduced from those of D^n and D^a as in the following expressions

$$\begin{aligned} \frac{\partial d^a}{\partial p} &= \frac{\partial a}{\partial p} + \frac{\partial D^a}{\partial n}\frac{\partial d^n}{\partial p} \\ \frac{\partial d^n}{\partial p} &= \frac{\partial D^n}{\partial p} + \frac{\partial D^n}{\partial a}\frac{\partial d^a}{\partial p} \end{aligned}$$

$$\begin{aligned} \frac{\partial d^a}{\partial m} &= \frac{\partial D^a}{\partial m} + \frac{\partial D^a}{\partial n}\frac{\partial d^n}{\partial m} \\ \frac{\partial d^n}{\partial m} &= \frac{\partial D^n}{\partial m} + \frac{\partial D^n}{\partial a}\frac{\partial d^a}{\partial m} \end{aligned}$$

The above expressions can be rearranged to obtain the following results

$$\begin{aligned} \frac{\partial d^a}{\partial p} &= \frac{\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} < 0, \qquad \frac{\partial d^n}{\partial p} = \frac{\frac{\partial D^n}{\partial a}\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} < 0 \\ \frac{\partial d^a}{\partial m} &= \frac{\frac{\partial D^a}{\partial m} + \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \lessgtr 0, \quad \frac{\partial d^n}{\partial m} = \frac{\frac{\partial D^n}{\partial m} + \frac{\partial D^n}{\partial a}\frac{\partial D^a}{\partial m}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \lessgtr 0. \end{aligned} \tag{11}$$

Consider the maximization problem of the platform when choosing m and p simultaneously. From the first-order conditions, and using the above expressions, it follows that

$$\begin{aligned} p =& -\frac{d^a(p,m)}{\frac{\partial d^a}{\partial p}} = -\frac{d^a(m,p)(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})}{\frac{\partial D^a}{\partial p}}, \\ C'(m) =& p\frac{\partial d^a}{\partial m} = p\Big(\frac{\frac{\partial D^a}{\partial m}+\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big). \end{aligned}$$

Denote MR the marginal gain from moderation such that $MR = p^*\frac{\partial d^a}{\partial m}|_{m=m^*}$. By using p as defined above, we then have

$$MR(m^*) = -\frac{d^a(m^*,p^*)(\frac{\partial D^a}{\partial m}+\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m})}{\frac{\partial D^a}{\partial p}} = -\frac{d^a(m^*,p^*)\Psi}{\frac{\partial D^a}{\partial p}}, \tag{12}$$

where

$$\Psi = \frac{\partial D^a}{\partial m}+\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}$$

represents the platform's elasticity to moderation. The optimal level of content moderation is implicitly defined by the following expression $MR(m^*) = C'(m^*)$, where the latter term accounts for the marginal cost of moderation.

Proof of Proposition 2

Consider how the equilibrium variables change with marginal gains from moderation (i.e either changes in brand risk or changes in user's preference for moderated contents). Let us first consider how demands on both sides of the market react. This boils down to the analysis of how equilibrium outcomes change with Ψ.

To begin with, consider the problem of the platform and recall that, from the first-order conditions, there exists a duple $(p,m) = (p^*,m^*)$ satisfying the following two expressions:

$$\begin{aligned} 0 &= p^*\frac{\partial D^a}{\partial p} + d^a(m^*,p^*)(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}), \\ 0 &= C'(m^*)\frac{\partial D^a}{\partial p} + d^a(m^*,p^*)\Psi. \end{aligned}$$

By differentiating the above expressions with respect to Ψ, we have the following:

$$\begin{aligned} 0 &= \frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi} + \frac{\partial d^a(m^*,p^*)}{\partial \Psi}(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}), \\ 0 &= C''(m^*)\frac{\partial m^*}{\partial \Psi}\frac{\partial D^a}{\partial p} + \frac{\partial d^a(m^*,p^*)}{\partial \Psi}\Psi + d^a(m^*,p^*). \end{aligned}$$

Using the chain rule, $\frac{\partial d^a(p^*,m^*)}{\partial \Psi} = \frac{\partial d^a}{\partial \Psi} + \frac{\partial d^a}{\partial m}\frac{\partial m^*}{\partial \Psi} + \frac{\partial d^a}{\partial p}\frac{\partial p^*}{\partial \Psi}$, we then have

$$0 = \frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi} + (1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})\Big(\frac{\partial d^a}{\partial \Psi} + \frac{\partial d^a}{\partial m}\frac{\partial m^*}{\partial \Psi} + \frac{\partial d^a}{\partial p}\frac{\partial p^*}{\partial \Psi}\Big),$$
$$0 = C''(m^*)\frac{\partial m^*}{\partial \Psi}\frac{\partial D^a}{\partial p} + \Big(\frac{\partial d^a}{\partial \Psi} + \frac{\partial d^a}{\partial m}\frac{\partial m^*}{\partial \Psi} + \frac{\partial d^a}{\partial p}\frac{\partial p^*}{\partial \Psi}\Big)\Psi + d^a(m^*, p^*).$$

Using (11) and exploiting $\frac{\partial d^a}{\partial \Psi} = \frac{\frac{\partial D^a}{\partial \Psi}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}$, we have

$$0 = 2\frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi} + \frac{\partial m^*}{\partial \Psi}\Psi + \frac{\partial D^a}{\partial \Psi},$$
$$0 = C''(m^*)\frac{\partial m^*}{\partial \Psi}\frac{\partial D^a}{\partial p} + \Big(\frac{\partial D^a}{\partial \Psi} + \Psi\frac{\partial m^*}{\partial \Psi} + \frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi}\Big)\frac{\Psi}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} + d^a(m^*, p^*).$$

In turn, this implies the following

$$-\frac{\partial D^a}{\partial \Psi} = 2\frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi} + \frac{\partial m^*}{\partial \Psi}\Psi,$$
$$-d^a(m^*, p^*) - \frac{\partial D^a}{\partial \Psi}\frac{\Psi}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} = \frac{\Psi}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\frac{\partial p^*}{\partial \Psi}\frac{\partial D^a}{\partial p} + \frac{\partial m^*}{\partial \Psi}\Big(C''(m^*)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big),$$

By using the Implicit Function Theorem, then we have

$$\begin{pmatrix} 2\frac{\partial D^a}{\partial p} & \Psi \\ \frac{\Psi\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & \frac{\partial D^a}{\partial p}C''(m) + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}\begin{pmatrix} \frac{\partial p^*}{\partial \Psi} \\ \frac{\partial m^*}{\partial \Psi} \end{pmatrix} = \begin{pmatrix} -\frac{\partial D^a}{\partial \Psi} \\ -d^a(m^*, p^*) - \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}. \quad (13)$$

Using the Cramer's rule, in turn, this implies the following

$$\frac{\partial p^*}{\partial \Psi} = \frac{\det\begin{pmatrix} -\frac{\partial D^a}{\partial \Psi} & \Psi \\ -d^a(m^*, p^*) - \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & \frac{\partial D^a}{\partial p}C''(m^*) + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}}{\det\begin{pmatrix} 2\frac{\partial D^a}{\partial p} & \Psi \\ \frac{\Psi\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & \frac{\partial D^a}{\partial p}C''(m) + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}}$$

$$\frac{\partial m^*}{\partial \Psi} = \frac{\det \begin{pmatrix} 2\frac{\partial D^a}{\partial p} & -\frac{\partial D^a}{\partial \Psi} \\ \frac{\Psi \frac{\partial D^a}{\partial p}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & -d^a(m^*,p^*) - \frac{\Psi \frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}}{\det \begin{pmatrix} 2\frac{\partial D^a}{\partial p} & \Psi \\ \frac{\Psi \frac{\partial D^a}{\partial p}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & \frac{\partial D^a}{\partial p}C''(m^*) + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}}$$

The denominator of both terms is equal to

$$\frac{\partial D^a}{\partial p}\Big(2C''(m^*)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big),$$

which is positive if, and only if,

$$C''(m) > -\frac{\Psi^2}{(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})2\frac{\partial D^a}{\partial p}}. \tag{14}$$

The numerator of $\frac{\partial p^*}{\partial \Psi}$ is

$$-\frac{\partial D^a}{\partial \Psi}\Big(\frac{\partial D^a}{\partial p}C''(m^*) + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big) + \Psi\Big(d^a(m^*,p^*) + \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big),$$

which can be rearranged in

$$\Psi d^a(m^*,p^*) - \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m^*).$$

In turn, we have

$$\frac{\partial p^*}{\partial \Psi} = \frac{\Psi d^a(m^*,p^*) - \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m^*)}{\frac{\partial D^a}{\partial p}(2C''(m^*)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}})}.$$

As the denominator is positive, the effect depends on the numerator, such that $sign(\frac{\partial p^*}{\partial \Psi}) > 0$ if $\Psi d^a(m^*,p^*) > \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m^*)$. Note that case of $m > 0$ ($\Psi > 0$), the LHS and the RHS are both positive. Hence, $\frac{\partial p^*}{\partial \Psi}$ is positive if Ψ and $d^a(m^*,p^*)$ is large enough while $C''(m)$ is low enough. Note that for $\Psi < 0$, we have $m^* = 0$, $C(m^* = 0) = 0$, which drives $\frac{\partial p^*}{\partial \Psi} < 0$.

Turning to the numerator of $\frac{\partial m^*}{\partial \overline{\lambda}}$, we then have

$$-2\frac{\partial D^a}{\partial p}\Big(d^a(m^*,p^*) + \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big) + \frac{\Psi\frac{\partial D^a}{\partial p}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\frac{\partial D^a}{\partial \Psi},$$

which can be rearranged in

$$-\frac{\partial D^a}{\partial p}(2d^a(m^*,p^*)+\frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}).$$

In turn, we have

$$\frac{\partial m^*}{\partial \Psi}=-\frac{2d^a(m^*,p^*)+\frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}}{(2C''(m^*)\frac{\partial D^a}{\partial p}+\frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}})}.$$

Considering the negative sign before the expression and that the denominator is negative, the total effect depends on the sign of numerator. Hence, $sign(\frac{\partial m^*}{\partial \Psi})>0$ if $2d^a(m^*,p^*)>-\frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}$. As $\frac{\partial D^a}{\partial \Psi}<0$, $\frac{\partial m^*}{\partial \Psi}>0$ if Ψ is low enough while $d^a(m^*,p^*)$ is large enough. Recall that convexity in costs need to be satisfied. Rearranging (14), this requires

$$\Psi<\Psi^c\equiv\sqrt{-2C''(m^*)\frac{\partial D^a}{\partial p}(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})}.$$

Call Ψ^p and Ψ^m the critical value of Ψ such that the numerators of $\frac{\partial p^*}{\partial \Psi}$ and $\frac{\partial m^*}{\partial \Psi}$, respectively, are both equal to zero, then

$$\Psi^p\equiv\frac{\frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}C''(m^*)}{d^a(m^*,p^*)}$$

$$\Psi^m\equiv=-2\frac{d^a(m^*,p^*)(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})}{\frac{\partial D^a}{\partial \Psi}}.$$

Denote by

$$\tilde{C}\equiv-2\Big(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}\Big)\frac{d^a(m^*,p^*)^2}{(\frac{\partial D^a}{\partial \Psi})^2\frac{\partial D^a}{\partial p}}.$$

We find that when :

- $C''(m^*)>\tilde{C}\Longrightarrow\Psi^p>\Psi^c>\Psi^m$
- $C''(m^*)<\tilde{C}\Longrightarrow\Psi^p<\Psi^c<\Psi^m$

As a result, if $C''(m^*)<\tilde{C}$, then

- If $0<\Psi\leq\Psi^p$, then $\frac{\partial p^*}{\partial \Psi}<0$ and $\frac{\partial m^*}{\partial \Psi}>0$;
- If $\Psi^p<\Psi\leq\Psi^c$, then $\frac{\partial p^*}{\partial \Psi}>0$ and $\frac{\partial m^*}{\partial \Psi}>0$;

Hence, m^* always increases with Ψ in the relevant parameter space, whereas p^* is U-shaped in Ψ.

Next, if $C''(m^*)>\tilde{C}$, then

- If $0 \leq \Psi^m$, then $\frac{\partial p^*}{\partial \Psi} < 0$ and $\frac{\partial m^*}{\partial \Psi} > 0$;
- If $\Psi^m < \Psi \leq \Psi^c$, then $\frac{\partial p^*}{\partial \Psi} < 0$ and $\frac{\partial m^*}{\partial \Psi} < 0$;

Hence, in the relevant space, p^* always declines with Ψ while m^* is inverted U-shaped in Ψ.

Proof of Proposition 3

Suppose platforms are ex-ante symmetric and compete for user attention. In what follows, we first look at how demands on both sides of the market change with ad prices and moderation.

Recall that the number of users joining platform 1 is $n_1 = \Pr(U_1 \geq U_2)$ with $n_2 = 1 - n_1$, whereas the number of advertisers is $a_1 = \Pr(V_1 > 0)$ and $a_2 = \Pr(V_2 > 0)$. Assume that the above system of equations admits a unique solution that defines a_i and n_i depending on (p_i, p_j, m_i, m_j) such that $a_i = d_i^a(p_i, m_i) \equiv D_i^a(p_i, p_j, m_i, m_j)$ and $n_i = d_i^n(p_i, p_j, m_i, m_j) \equiv D_i^n(p_i, p_j, m_i, m_j)$, for $i = 1, 2$, with $i \neq j$. As $D_j^n = 1 - D_i^n$, it follows that $\frac{\partial D_j^n}{\partial(.)} = -\frac{\partial D_i^n}{\partial(.)}$ and $\frac{\partial d_j^n}{\partial(.)} = -\frac{\partial d_i^n}{\partial(.)}$.

The derivatives of d_i^a and d_i^n with respect to p_i, p_j and m_i, m_j can be deduced from those of D_i^n and D_i^a. Specifically, the effect of prices on platforms' demands on both sides of the market is as follows

$$\begin{aligned}
\frac{\partial d_i^a}{\partial p_i} &= \frac{\partial D_i^a}{\partial p_i} + \frac{\partial D_i^a}{\partial n_i}\frac{\partial d_i^n}{\partial p_i} \\
\frac{\partial d_i^a}{\partial p_j} &= \frac{\partial D_i^a}{\partial n_i}\frac{\partial d_i^n}{\partial p_j} \\
\frac{\partial d_i^n}{\partial p_i} &= \frac{\partial D_i^n}{\partial a_i}\frac{\partial d_i^a}{\partial p_i} + \frac{\partial D_i^n}{\partial a_j}\frac{\partial d_j^a}{\partial p_i} \\
\frac{\partial d_i^n}{\partial p_j} &= \frac{\partial D_i^n}{\partial a_i}\frac{\partial d_i^a}{\partial p_j} + \frac{\partial D_i^n}{\partial a_j}\frac{\partial d_j^a}{\partial p_j}.
\end{aligned}$$

Similarly, the effect of content moderation on platforms' demands on both sides of the

market is equal to:

$$\begin{aligned}
\frac{\partial d_i^a}{\partial m_i} &= \frac{\partial D_i^a}{\partial m_i} + \frac{\partial D_i^a}{\partial n_i}\frac{\partial d_i^n}{\partial m_i} \\
\frac{\partial d_i^n}{\partial m_i} &= \frac{\partial D_i^n}{\partial m_i} + \frac{\partial D_i^n}{\partial a_i}\frac{\partial d_i^a}{\partial m_i} + \frac{\partial D_i^n}{\partial a_j}\frac{\partial d_j^a}{\partial m_i} \\
\frac{\partial d_i^a}{\partial m_j} &= \frac{\partial D_i^a}{\partial n_i}\frac{\partial d_i^n}{\partial m_j} \\
\frac{\partial d_i^n}{\partial m_j} &= \frac{\partial D_i^n}{\partial m_j} + \frac{\partial D_i^n}{\partial a_i}\frac{\partial d_i^a}{\partial m_j} + \frac{\partial D_i^n}{\partial a_j}\frac{\partial d_j^a}{\partial m_j}
\end{aligned}$$

The above expressions can be rearranged to obtain the following

$$\begin{aligned}
\frac{\partial d_i^a}{\partial p_i} &= \frac{\frac{\partial D_i^a}{\partial p_i}(1 - \frac{\partial D_j^a}{\partial n_j}\frac{\partial D_j^n}{\partial a_j})}{1 - \frac{\partial D_j^a}{\partial n_j}\frac{\partial D_j^n}{\partial a_j} - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i} + \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_j^a}{\partial n_j}(\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_j^n}{\partial a_j} - \frac{\partial D_i^n}{\partial a_j}\frac{\partial D_j^n}{\partial a_i})}, \\
\frac{\partial d_i^a}{\partial m_i} &= \frac{(\frac{\partial D_i^a}{\partial m_i} + \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial m_i})(1 - \frac{\partial D_j^a}{\partial n_j}\frac{\partial D_j^n}{\partial a_j}) + \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_j}\frac{\partial D_j^a}{\partial n_j}\frac{\partial D_j^n}{\partial m_i}}{1 - \frac{\partial D_j^a}{\partial n_j}\frac{\partial D_j^n}{\partial a_j} - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i} + \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_j^a}{\partial n_j}(\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_j^n}{\partial a_j} - \frac{\partial D_i^n}{\partial a_j}\frac{\partial D_j^n}{\partial a_i})}.
\end{aligned} \tag{15}$$

By assuming symmetry, we can simplify as follows:

$$\begin{aligned}
\frac{\partial d_i^a}{\partial p_i} &= \frac{\frac{\partial D_i^a}{\partial p_i}(1 - \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})}{1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}}, \\
\frac{\partial d_i^a}{\partial m_i} &= \frac{\frac{\partial D_i^a}{\partial m_i} + \frac{\partial D_i^a}{\partial n_i}(\frac{\partial D_i^n}{\partial m_i} - \frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i})}{1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}} \\
\frac{\partial d_i^a}{\partial p_j} &= -\frac{\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial p_i}}{1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}} \\
\frac{\partial d_i^a}{\partial m_j} &= -\frac{\frac{\partial D_i^a}{\partial n_i}(\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i} + \frac{\partial D_i^n}{\partial m_i})}{1 - 2\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}}.
\end{aligned} \tag{16}$$

For the sake of notation, we denote by $1 - 2\Upsilon$ the denominator of the two expressions above, where $\Upsilon \equiv \frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i}$.

We are now in the condition to state the maximization problem of platform i and the optimal choice of m_i and p_i simultaneously in the first period of the game. From the

first-order conditions, and using the above expressions, one can easily obtain:

$$\begin{aligned} p_i =& -\frac{d_i^a(p_i,p_j,m_i,m_j)}{\frac{\partial d_i^a}{\partial p_i}} = -\frac{d_i^a(p_i,p_j,m_i,m_j)(1-2\Upsilon)}{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)}, \\ C'(m_i) =& p_i\frac{\partial d_i^a}{\partial m_i} = p_i\frac{\frac{\partial D_i^a}{\partial m_i}+\frac{\partial D_i^a}{\partial n_i}(\frac{\partial D_i^n}{\partial m_i}-\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i})}{(1-2\Upsilon)}, \end{aligned}$$

from which we can identify MR_i, the marginal gain from moderation for platform i:

$$MR_i(m_i^*) = -\frac{d_i^a(p_i,p_j,m_i,m_j)\Psi_i}{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)} \tag{17}$$

where

$$\Psi_i = \frac{\partial D_i^a}{\partial m_i}+\frac{\partial D_i^a}{\partial n_i}(\frac{\partial D_i^n}{\partial m_i}-\frac{\partial D_i^n}{\partial a_i}\frac{\partial D_i^a}{\partial m_i}).$$

The optimal level of content moderation is implicitly defined by the following expression $MR_i(m_i^*) = C'(m_i^*)$. This concludes the proof.

Proof of Proposition 4

To start with, recall the first-order conditions such that there exists a duple $(p,m) = (p^*,m^*)$ satisfying the following two expressions:

$$\begin{aligned} 0 =& p_i^*\frac{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)}{(1-2\Upsilon)} + d_i^a(p_i^*,p_j^*,m_i^*,m_j^*) = 0, \\ 0 =& \frac{d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)\Psi_i}{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)} - C'(m_i^*) \end{aligned}$$

By differentiating them with respect to τ, we have the following

$$\begin{aligned} 0 =& \frac{\partial D_i^a}{\partial p_i}\frac{1}{1-2\Upsilon}\Big(\frac{2p_i^*(1-\Upsilon)\frac{\partial\Upsilon}{\partial\tau}}{(1-2\Upsilon)} - p_i^*\frac{\partial\Upsilon}{\partial\tau} + \frac{\partial p_i^*}{\partial\tau}(1-\Upsilon)\Big) + \frac{\partial d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)}{\partial\tau}, \\ 0 =& -\frac{1}{(1-2\Upsilon)}\Big(d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)\frac{\partial\Psi_i}{\partial\tau} + \frac{2d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)\Psi_i\frac{\partial\Upsilon}{\partial\tau}}{(1-2\Upsilon)} + \Psi_i\frac{\partial d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)}{\partial\tau}\Big) \\ & - C''(m_i^*)\frac{\partial m_i^*}{\partial\tau}. \end{aligned} \tag{18}$$

Using the chain rule, we know that $\frac{\partial d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)}{\partial \tau}$ can be decomposed as follows:

$$\frac{\partial d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)}{\partial \tau} = \frac{\partial d_i^a}{\partial m_i}\frac{\partial m_i^*}{\partial \tau} + \frac{\partial d_i^a}{\partial p_i}\frac{\partial p_i^*}{\partial \tau} + \frac{\partial d_i^a}{\partial m_j}\frac{\partial m_j^*}{\partial \tau} + \frac{\partial d_i^a}{\partial p_j}\frac{\partial p_j^*}{\partial \tau},$$

Exploiting the fact that $\frac{\partial d_i^a}{\partial \tau} = \frac{\partial D_i^a}{\partial n_i}\frac{\partial d_i^n}{\partial \tau}$ and $\frac{\partial d_i^n}{\partial \tau} = \frac{\partial D_i^n}{\partial \tau} + \frac{\partial D_i^n}{\partial a_i}\frac{\partial d_i^a}{\partial \tau} + \frac{\partial D_i^n}{\partial a_j}\frac{\partial d_j^a}{\partial \tau}$, we then have

$$\frac{\partial d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)}{\partial \tau} = \frac{1}{1-2\Upsilon}\Big(\Psi_i\frac{\partial m_i^*}{\partial \tau} + \frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)\frac{\partial p_i^*}{\partial \tau} - \Psi_j\frac{\partial m_j^*}{\partial \tau} - \Upsilon\frac{\partial D_i^a}{\partial p_i}\frac{\partial p_j^*}{\partial \tau}\Big).$$

Solving (18), we have the following two results:

$$\frac{\partial p_i}{\partial \tau} = \frac{d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)\Big(\Big(2\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}\Big)\frac{\partial \Upsilon}{\partial \tau} + \frac{\partial D_i^a}{\partial m_i}(1-2\Upsilon)\frac{\partial \Psi_i}{\partial \tau}\Big)}{\frac{\partial D_i^a}{\partial p_i}(1-\Upsilon)(\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}(2-3\Upsilon))}$$

$$\frac{\partial m_i}{\partial \tau} = -\frac{d_i^a(p_i^*,p_j^*,m_i^*,m_j^*)\Big((2-3\Upsilon)\frac{\partial \Psi_i}{\partial \tau} + 3\Psi_i\frac{\partial \Upsilon}{\partial \tau}\Big)}{(1-\Upsilon)(\frac{\partial D_i^a}{\partial m_i}\Psi_i + C''(m_i^*)\frac{\partial D_i^a}{\partial p_i}(2-3\Upsilon))}$$

Recall that we have assumed sufficiently convex moderation cost. As this implies that $C''(m_i^*) > -\frac{(1-\Upsilon)\frac{\partial D_i^a}{\partial m_i}\Psi_i}{\frac{\partial D_i^a}{\partial p_i}(2-3\Upsilon)}$, the denominator of $\frac{\partial m_i^*}{\partial \tau} < 0$ while that of $\frac{\partial p_i^*}{\partial \tau}$ is positive because of the presence of $\frac{\partial D_i^a}{\partial p_i} < 0$.

To study the sign of the numerators, recall that $\Upsilon < 0$, $\frac{\partial \Upsilon}{\partial \tau} > 0$, and $\Psi_i > 0$. By looking at numerator of $\frac{\partial p_i^*}{\partial \tau}$, one can see that there exists a critical value of $C''(m_i^*)$ such that if

$$C''(m_i^*) \leq -\frac{\frac{\partial D_i^a}{\partial m_i}(1-2\Upsilon)\frac{\partial \Psi_i}{\partial \tau} + 2\frac{\partial D_i^a}{\partial m_i}\Psi_i\frac{\partial \Upsilon}{\partial \tau}}{\frac{\partial D_i^a}{\partial p_i}\frac{\partial \Upsilon}{\partial \tau}} \equiv \tilde{C}$$

then the numerator is negative and $\frac{\partial p_i^*}{\partial \tau} > 0$. Otherwise, the numerator is positive, and $\frac{\partial p_i^*}{\partial \tau} < 0$ for $C''(m_i^*) > \tilde{C}$. In turn, this implies that as τ decreases, p_i^* decreases if $C''(m_i^*)$ is sufficiently small and increases otherwise.

It remains to prove the effect of τ on the optimal moderation policy m_i^*. As the denominator of $\frac{\partial m_i}{\partial \tau}$ is negative and the expression is introduced by a minus, m_i^* increases with τ if the numerator is positive, that is, if

$$(2-3\Upsilon)\frac{\partial \Psi_i}{\partial \tau} + 3\Psi_i\frac{\partial \Upsilon}{\partial \tau} > 0$$

and, more specifically, it suffices that $\frac{\partial \Psi_i}{\partial \tau} > 0$.

Decomposing the above expression using Ψ and Υ, we have:

$$\frac{\partial D_i^a}{\partial n_i}(2-3\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})\frac{\partial^2 D_i^n}{\partial m_i \partial \tau}+\frac{\partial D_i^a}{\partial n_i}(3\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial m_i}+\frac{\partial D_i^a}{\partial m_i})\frac{\partial^2 D_i^n}{\partial a_i \partial \tau}.$$

One can notice that the sign of the effect depends on the sign of cross-derivatives. As $D_i^n \equiv \Pr(\frac{\phi(\theta(m_1)-\theta(m_2))+\gamma(a_1-a_2)}{\tau} > y)$, we consider $D_i^n \equiv G(Z>y)$ with $Z=\frac{Y}{\tau}$ and $Y=\phi(\theta(m_1)-\theta(m_2))+\gamma(a_1-a_2)$. Hence we have $\frac{\partial D_i^n}{\partial(.)}=\frac{\partial G}{\partial Z}\frac{\partial Z}{\partial Y}\frac{\partial Y}{\partial(.)}=\frac{\partial G}{\partial Z}\frac{1}{\tau}\frac{\partial Y}{\partial(.)}$. As a consequence, we have $\frac{\partial^2 D_i^n}{\partial(.)\partial\tau}=-\frac{\partial G}{\partial Z}\frac{1}{\tau^2}\frac{\partial Y}{\partial(.)}=-\frac{1}{\tau}\frac{\partial D_i^n}{\partial(.)}$. Replacing it in the above equation, we obtain:

$$\frac{\partial D_i^a}{\partial n_i}(2-3\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial a_i})(-\frac{1}{\tau}\frac{\partial D_i^n}{\partial m_i})+\frac{\partial D_i^a}{\partial n_i}(3\frac{\partial D_i^a}{\partial n_i}\frac{\partial D_i^n}{\partial m_i}+\frac{\partial D_i^a}{\partial m_i})(-\frac{1}{\tau}\frac{\partial D_i^n}{\partial a_i}).$$

Rearranging it, we can observe the following:

$$-\frac{1}{\tau}\frac{\partial D_i^a}{\partial n_i}\underbrace{\left(2\frac{\partial D_i^n}{\partial m_i}+\frac{\partial D_i^a}{\partial m_i}\frac{\partial D_i^n}{\partial a_i}\right)}_{<0}>0.$$

In turn, the numerator is positive and $\frac{\partial m_i^*}{\partial \tau}>0$, always. This concludes the proof.

Proof of Proposition 5

Once again, consider the problem of a monopolist platform as described in Section 2 and let us introduce a given tax $f<p^*$ per ad. Focusing on the interior solutions only, the first-order conditions are:

$$\begin{aligned}\frac{\partial \Pi}{\partial p}(p^*,m^*) =&(p^*-f)\frac{\partial d^a}{\partial p}+d^a(p^*,m^*)=0,\\ \frac{\partial \Pi}{\partial m}(p^*,m^*) =&(p^*-f)\frac{\partial d^a}{\partial m}-C'(m^*)=0\end{aligned} \tag{19}$$

for $(p,m)=(p^*,m^*)$. To understand the effect of f on p^* and m^*, we next differentiate the above system of equations with respect to f, which leads to the following result:

$$\begin{aligned}\frac{\partial d^a}{\partial p}(\frac{\partial p^*}{\partial f}-1)+\frac{\partial d^a}{\partial f} =&0,\\ \frac{\partial d^a}{\partial m}(\frac{\partial m^*}{\partial f}-1)-C''(m^*)\frac{\partial m^*}{\partial f} =&0\end{aligned} \tag{20}$$

As d^a depends on f only through p^* and m^*, using the chain rule, we can define the following $\frac{\partial d^a}{\partial f}=\frac{\partial d^a}{\partial m}\frac{\partial m^*}{\partial f}+\frac{\partial d^a}{\partial p}\frac{\partial p^*}{\partial f}$. As a result, the above system of equations can be

expressed as follows:

$$\begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \\ \frac{\partial d^a}{\partial m} & -C''(m^*) \end{pmatrix} \begin{pmatrix} \frac{\partial p^*}{\partial f} \\ \\ \frac{\partial m^*}{\partial f} \end{pmatrix} = \begin{pmatrix} \frac{\partial d^a}{\partial p} \\ \\ \frac{\partial d^a}{\partial m} \end{pmatrix} \tag{21}$$

By the Implicit Function Theorem and the Cramer's Rule, we get

$$\frac{\partial p^*}{\partial f} = \frac{\det \begin{pmatrix} \frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \\ \frac{\partial d^a}{\partial m} & -C''(m^*) \end{pmatrix}}{\det \begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \\ \frac{\partial d^a}{\partial m} & -C''(m^*) \end{pmatrix}} \quad \frac{\partial m^*}{\partial f} = \frac{\det \begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial p} \\ \\ \frac{\partial d^a}{\partial m} & \frac{\partial d^a}{\partial m} \end{pmatrix}}{\det \begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \\ \frac{\partial d^a}{\partial m} & -C''(m^*) \end{pmatrix}}. \tag{22}$$

Using $\frac{\partial d^a}{\partial m}$ and $\frac{\partial d^a}{\partial p}$ as previously defined, we can rewrite the denominator of both expressions as equal to:

$$-2C''(m^*)\frac{\partial d^a}{\partial p} - \left(\frac{\partial d^a}{\partial m}\right)^2 = -\frac{1}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\left(2C''(m^*)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\right).$$

The numerator of $\frac{\partial p^*}{\partial f}$ is equal to:

$$-C''(m^*)\frac{\partial d^a}{\partial p} - \left(\frac{\partial d^a}{\partial m}\right)^2 = -\frac{1}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\left(C''(m^*)\frac{\partial D^a}{\partial p} + \frac{\Psi^2}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\right).$$

Hence, we obtain

$$\frac{\partial p^*}{\partial f} = \frac{C''(m^*)\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2}{2C''(m^*)\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2},$$

which is positive if, and only if, the numerator is negative, that is, when :

$$\frac{\partial p^*}{\partial f} > 0 \Longleftrightarrow -C''(m^*)\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) > \Psi^2.$$

Hence, there exists a critical value of moderation costs such that if

$$C''(m^*) > \tilde{C} \equiv -\frac{\Psi^2}{\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})},$$

then $\frac{\partial p^*}{\partial f} > 0$. Else, for $C''(m^*) < \tilde{C}$, we then have $\frac{\partial p^*}{\partial f} < 0$. The numerator of $\frac{\partial p^*}{\partial f}$ is equal to

$$2\frac{\partial d^a}{\partial p}\frac{\partial d^a}{\partial m} - \frac{\partial d^a}{\partial m}\frac{\partial d^a}{\partial p} = \frac{\frac{\partial D^a}{\partial p}\Psi}{(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})^2}.$$

Hence, we obtain

$$\frac{\partial m^*}{\partial f} = -\frac{\frac{\partial D^a}{\partial p}\Psi(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})^{-1}}{2C''(m^*)\frac{\partial D^a}{\partial p}(1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}) + \Psi^2} < 0,$$

as the numerator is always negative. This concludes the proof.

Appendix B

To get a better picture of the trade-offs we highlighted in the benchmark model with a monopolist platform, in this Appendix we present an example with a uniform distribution of preferences of advertisers and users.

Monopolist Platform

First, let us specify the utility of users and advertisers, respectively, for a uniform distribution of preferences. An Internet user is defined by a duple $(u, \phi) \sim U[0, \overline{u}] \times U[0, \overline{\phi}]$, such that both u and ϕ are uniformly and independently distributed. We only consider the case in which users dislike moderation, but their tastes are heterogeneous, i.e., $\underline{\phi}$ in Section 2, is set equal to zero. This simplifying assumption allows us to focus on the case in which advertisers' and users' preferences over moderation are conflicting - which is the most insightful case. As a result, if some users had a negative ϕ, the platform would have a slightly higher incentive to increase content moderation effort as users' preferences would converge towards those of advertisers. We also assume that advertisers are heterogeneous: an advertiser is a duple $(v, \lambda) \sim U[0, \overline{v}] \times U[0, \overline{\lambda}]$, such that both v and λ are uniformly and independently distributed. This allows us to capture differences in the net gain from purchasing an ad campaign on a social network, depending on the long-term net gain from being on the platform.

To make the model tractable, we assume that preferences for safe content have a stronger effect on both users and advertisers than the ones for unsafe content. On the user side, $\overline{u}$ needs to be sufficiently large such that users who do not value unsafe content (say $\phi = 0$) do not refrain from using platform when their intrinsic preference for the platform, u, is large enough. Conversely, we also assume that $\overline{\phi}$ is low enough such that users who largely enjoy unsafe content ($\phi = \overline{\phi}$) may decide not to visit the social network if the intrinsic utility they get from the platform (i.e., access to a mass 1 of safe content) is not high enough. More generally, the above assumptions imply that preferences for safe content have a stronger effect on user decisions than aversion for the moderation of unsafe content.

On the advertiser side, we let $\overline{v}$ be such that advertisers not concerned with brand safety (e.g., $\lambda = 0$) will not display ads when not benefiting enough from interactions with the unit mass of safe content. We also assume that the $\overline{\lambda}$ is low enough, such that some advertisers with a strong preference for brand safety ($\lambda = \overline{\lambda}$) may still display ads when deriving high gain from being exposed to the mass 1 of safe content. More generally,

such an assumption implies that v has a stronger impact on advertisers' decisions to buy an ad than brand risk issues alone.

To provide clear insights on the optimal ad price p^* and moderation policy m^* and to compute the equilibrium, we assume that the mass of unsafe content is a linear and decreasing function of m. Similarly, we assume quadratic moderation costs.

$$\theta(m) = 1 - m \quad \text{and} \quad C(m) = c\frac{m^2}{2}, \quad \text{with} \quad c > 0.$$

With the above specifications on hold, on the second stage, advertisers decide whether to display an ad and users whether to join the platform. This results in the following demand:

$$d^a(m,p) = \frac{\theta(m)(r\overline{\phi} - \overline{u}\overline{\lambda}) + 2\overline{u}(r + \overline{v} - p)}{2(\overline{uv} + \gamma r)}. \tag{23}$$

$$d^n(m,p) = \frac{(2\overline{u} + \theta(m)\overline{\phi} - 2\gamma)\overline{v} + 2\gamma p + \theta(m)\gamma\overline{\lambda}}{2(\overline{uv} + \gamma r)}.$$

By maximizing (1) and rearranging it, we obtain

$$\begin{aligned} p(m) =& \frac{2\overline{u}(\overline{v} + r) + r\overline{\phi}\theta(m) - \overline{u}\overline{\lambda}\theta(m)}{4\overline{u}}, \\ C'(m) =& p\frac{(r\overline{\phi}\theta'(m) - \overline{u}\overline{\lambda}l'(m)}{2(\overline{uv} + \gamma r)}. \end{aligned}$$

Denote $\Psi = \overline{\lambda}\overline{u} - \overline{\phi}r$ the platform profit elasticity with respect to moderation, convexity in costs is ensured by the following expression

$$c > \frac{\Psi^2}{8\overline{u}(\overline{uv} + \gamma r\overline{u})}$$

which is equivalent to the one in (14). Using $\theta(m) = 1 - m$, we obtain the following equilibrium outcomes:

$$m^* = \begin{cases} 0 \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \leq 0, \\ \frac{\Psi(2\overline{u}(\overline{v}+r)-\Psi)}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2} \text{ if } \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}r \geq 0. \\ 1 \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \geq \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r}. \end{cases}$$

$$p^* = \begin{cases} \frac{\overline{u}(\overline{v}+r)-\Psi}{2\overline{u}} \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \leq 0, \\ \\ \frac{2c(\overline{uv}+\gamma r)(2\overline{u}(\overline{v}+r)-\Psi)}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2} \text{ if } \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}r \geq 0. \\ \\ \frac{\overline{v}+r}{2} \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \geq \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r}. \end{cases} \tag{24}$$

First, we begin to prove the corner solutions. Suppose the platform chooses a no moderation policy such that $m^* = 0$. This happens whenever $2\overline{u}(\overline{v}+r) < \Psi$ and $\Psi > 0$ or $2\overline{u}(\overline{v}+r) > \Psi$ and $\Psi < 0$. One can easily note that former would imply negative price, which is not possible, while the latter is instead compatible with positive user and advertisers demand. Hence $m^* = 0$ when $\Psi \leq 0$. Suppose now the platform enforces full moderation, $m^* = 1$. This happens for any $\frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}$. In both case, p^* is retrieved by substituting $m = \{0, 1\}$ into $p(m)$ to obtain the expression defined in the first and the third lines in (24). Second, consider an interior solution such that m^* belongs to $(0, 1)$. This happens if $\frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}r \geq 0$, and so the price is the one defined in the second line in (24).

Using the equilibrium outcomes, we can then derive the profits of the platform as follows.

$$\Pi^* = \begin{cases} \frac{(\overline{v}+r)(\overline{u}(\overline{v}+r)-\Psi)}{4(\overline{uv}+\gamma r)} \text{ if } \overline{\lambda}\overline{u} < \overline{\phi}r, \\ \\ \frac{c(2\overline{u}(\overline{v}+r)-\Psi^2)}{2(8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2)} \text{ if } \frac{4c(\overline{uv}+\gamma r)+\overline{\phi}r(\overline{v}+r)}{\overline{v}+r} > \overline{\lambda}\overline{u} \geq \overline{\phi}r. \\ \\ \frac{\overline{u}(\overline{v}^2+r(2\overline{v}+r))-2c(\overline{uv}+\gamma r)}{4(\overline{uv}+\gamma r)} \text{ if } \overline{\lambda}\overline{u} \geq \frac{4c(\overline{uv}+\gamma r)+\overline{\phi}r(\overline{v}+r)}{\overline{v}+r}. \end{cases}$$

In what follows, we provide support for the results in the general model and, more specifically, for those in Proposition 1. Hence, we study how p^* and m^* vary with Ψ.

First, notice that numerators for both m^* and p^* cancels out when $\Psi = 2\overline{u}(\overline{v}+r)$. This happens as brand risk is sufficiently large such that the advertiser willingness to pay decreases. Second, rewrite condition (5) with respect to Ψ. We find that for both m^* and p^*, the numerator cancels out for higher value of Ψ than the denominator if

$$c < \frac{\overline{u}(\overline{v}+r)^2}{2(\overline{uv}+\gamma r)} \equiv \tilde{c} \tag{25}$$

Importantly, this leads to the following results. If moderation costs are sufficiently small, $c < \tilde{c}$, we find that $\frac{\partial^2 p^*}{\partial^2 \Psi} > 0$. This then implies that p^* admits a minimum in

$$\Psi = 2\overline{u}(\overline{v}-r) - 2\sqrt{\overline{u}(\overline{uv}(\overline{v}+2r) - 2c(\overline{vu}-\gamma r))}.$$

Hence, p^* is convex and U-shaped in Ψ. Similarly, we it can be easily verified that $\frac{\partial m^*}{\partial \Psi} > 0$ when $c < \frac{\Psi^2(\overline{v}+r)}{8(\overline{uv}+\gamma r)(\Psi-\overline{u}(\overline{v}+r))}$ - this is ensured by (25). To see it, consider that m^* admits an inflexion point in $\Psi = \overline{u}(\overline{v}+r)$, which is also the maximum of the numerator. Hence, for $\Psi < \overline{u}(\overline{v}+r)$, $\frac{\partial m^*}{\partial \Psi} > 0$ as the denominator decreases and the numerator converges to the maximum. For a higher value of Ψ, the numerator may pass its maximum, turning the shape of m from concave to convex. However, as the denominator is shrinking faster than the numerator, we then have $\frac{\partial m^*}{\partial \Psi} > 0$

If moderation costs are sufficiently large, $c > \tilde{c}$, the numerator of p^* and m^* cancels out for existing value of Ψ that ensure profit concavity. In this case, we find $\frac{\partial p^*}{\partial \Psi} < 0$, until a point where Ψ is so high such that no the consumer demand disappears. By the same mechanism as the previous case, we find that $\frac{\partial^2 m^*}{\partial^2 \Psi} < 0$, and that m^* admits a maximum in

$$\Psi = \frac{4c(\overline{uv}+\gamma r) - 2^{\frac{3}{2}}\sqrt{c(\overline{uv}+\gamma r)(2c(\overline{uv}+\gamma r)+\overline{u}(\overline{v}+r)^2)}}{\overline{v}+r}.$$

Hence, m^* is inverted U-shaped in Ψ.

To sum up,

- if $c \leq \tilde{c}$, p^* is convex in Ψ and $\frac{\partial m^*}{\partial \Psi} > 0$.
- if $c > \tilde{c}$, m^* is concave in Ψ and $\frac{\partial p^*}{\partial \Psi} < 0$.

Using the expressions presented in this Appendix, one can easily retreive the shapes of the curves presented in Figure 1.

Effect of a tax on ad-revenues

Assume a uniform distribution of preferences. The following results are presented. First, the effect on advertisers is such that

$$\frac{da}{df} = \frac{1}{2(\overline{uv}+\gamma r)}\left\{\Psi\frac{\partial m}{\partial f} - 2\overline{u}\frac{\partial p}{\partial f}\right\}$$

where $\Psi = \overline{\lambda}\overline{u} - \overline{\phi}r$ the *elasticity of profit with respect to moderation*. Its sign depends on the sign of $\frac{dm}{df}$ and $\frac{dp}{df}$. To see it, consider the second stage of the game. The platform chooses m and p to maximize profits. Rearranging the first order conditions, we can see that the total effect on moderation effort and prices comes from the solution of the

following system of equations.

$$\begin{aligned}\frac{dm}{df} =& \frac{1}{2(\overline{uv}+\gamma r)}\Big\{-1+\frac{\partial p}{\partial f}\Big\},\\ \frac{dp}{df} =& \frac{1}{2}+\frac{1}{4}\Big(\overline{\lambda}-\frac{r\overline{\phi}}{\overline{u}}\Big)\frac{\partial m}{\partial f},\end{aligned} \tag{26}$$

which then can be solved as follows:

$$\begin{aligned}\frac{dm}{df} =& -\frac{2\overline{u}f}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2},\\ \frac{dp}{df} =& \frac{4c\overline{u}(\overline{uv}+\gamma r)-\Psi^2}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2},\end{aligned} \tag{27}$$

The above results indicate that $\frac{dm}{df}<0$, whereas $\frac{dp}{df}>(<)0$ depending on the sign of the numerator. When c is sufficiently large, i.e., $c>\overline{c}:=\frac{\Psi^2}{4\overline{u}(\overline{uv}+\gamma r)}$, the price increases with more taxation. Else it decreases.

Platform competition

Consider the case in which platforms compete and assume a uniform distribution of tastes. Advertisers are defined by a duple $(v,\lambda)\in[0,\overline{v}]\times[0,\overline{\lambda}]$, with a uniform distribution of v and λ. Their utility when patronizing platform i is

$$V_i = v + rn_i - \lambda\theta(m_i) - p_i. \tag{28}$$

On the other side of the market, users are uniformly and independently distributed on a line of unit length; they are identified by a duple relative to their relative preference for platform $i(j)$, defined by their position y on the Hotelling line and by their aversion for moderation ϕ. The latter is assumed to be uniformly distributed in the interval $[0,\overline{\phi}]$. Note that, for tractability, we only consider the case in which users dislike moderation, but their tastes are heterogeneous, i.e., $\underline{\phi}$ in Section 2, is set equal to zero. The utility of a user located at y and joining platform i is as follows:

$$U_i = u + \phi\theta(m_i) - \gamma a_i - \tau|y-l_i|, \tag{29}$$

where $l_i\in\{0,1\}$ indicates the location of the platform on the Hotelling line.

To simplify the setting, we also assume that profits are well-behaved as long as $C(m_i)$ is sufficiently convex. For the sake of simplicity, we let moderation costs be quadratic (i.e., $C(m_i)=cm_i^2/2$).

In what follows, we look for a symmetric equilibrium. By solving the model by backward induction, the number of advertisers is

$$a_i(n_i) = 1 + \frac{rn_i - \overline{\lambda}\theta(m_i) - p_i}{\overline{v}}, \tag{30}$$

whereas the number of users 'exclusive' to each platform is

$$n_i(a_i, a_j) = \frac{1}{2} + \frac{\overline{\phi}(\theta(m_i) - \theta((m_j)) - 2\gamma(a_i - a_j)}{4\tau}, \qquad n_j = 1 - n_i. \tag{31}$$

By using (28) and (29), we can determine the number of users and advertisers in platform i as

$$n_i = \frac{1}{2} + \frac{\overline{v}\overline{\phi}(\theta(m_i) - \theta(m_j)) + \gamma(2(p_j - p_i) + \overline{\lambda}(\theta(m_i) - \theta(m_j)))}{4(\tau\overline{v} + \gamma \cdot r)} \tag{32}$$

$$a_i = 1 - \frac{\overline{v}(r\overline{\phi}(\theta(m_i) - \theta(m_j)) + 2\tau(r + 2p_i - \theta(m_i)\overline{\lambda})) + r\gamma(2(r - p_i - p_j) - \overline{\lambda}(\theta(m_i) + \theta(m_j))}{4\overline{v}(\tau\overline{v} + \gamma r)}. \tag{33}$$

From the first-order conditions, we can solve for the symmetric equilibrium prices and moderation policies $p_i^* = p_j^*$ and $m_i^* = m_j^*$. As in the benchmark case, it is also assumed that $\theta(m) = 1 - m_i = 1 - m_j$.

For ease of exposition, let us define the following critical values of λ.

$$\lambda_{\underline{comp}} \equiv \frac{\overline{\phi}r\overline{v}}{(\tau\overline{v} + \gamma r)}, \quad \lambda_{\overline{comp}} \equiv \frac{4c(4\tau\overline{v} + 3\overline{\gamma}r)}{(2\overline{v} + r)(2\overline{v}\tau + \gamma r)} + \frac{\overline{\phi}r\overline{v}}{2\tau\overline{v} + \gamma r}.$$

The optimal values of m^* and p^* for a symmetric equilibrium. Formally, when $\overline{\lambda} < \lambda_{\underline{comp}}$, $m_i^* = m_j^* = 0$, and

$$p_i^* = p_j^* = \frac{(2\overline{v} + r - \overline{\lambda})(\tau\overline{v} + \gamma r)}{4\tau\overline{v} + 3\gamma r}.$$

When $\lambda_{\underline{comp}} < \overline{\lambda} < \lambda_{\overline{comp}}$, an interior solution exists, with $m_i^* = m_j^* \in [0, 1]$ defined as follows

$$p_i^* = p_j^* = \frac{4c\overline{v}(2\overline{v} + r - \overline{\lambda})(\tau\overline{v} + \gamma r)}{16c\tau\overline{v}^2 + ((\overline{\lambda\phi} + 12c\gamma)r - 2\overline{\lambda}^2\tau)\overline{v} - \gamma\overline{\lambda}^2 r},$$
$$m_i^* = m_j^* = \frac{(2\overline{v} + r - \overline{\lambda})(\overline{\lambda}(2\tau\overline{v} + \gamma r) - \overline{\phi}r\overline{v})}{16c\tau\overline{v}^2 + ((\overline{\lambda\phi} + 12c\gamma)r - 2\overline{\lambda}^2\tau)\overline{v} - \gamma\overline{\lambda}^2 r}.$$

When $\overline{\lambda} \geq \lambda_{\overline{comp}}$, $m_i^* = m_j^* = 1$ and

$$p_i^* = p_j^* = \frac{(2\overline{v} + r)(\tau\overline{v} + \gamma r)}{4\tau\overline{v} + 3\gamma r}.$$

The above expressions represent the Nash equilibrium outcomes under competition between two symmetric platforms. Note that a condition to ensure a non-negative price is that $\overline{\lambda} < r + 2\overline{v}$. This implies that advertiser's brand risk is not as high (i.e., low enough $\overline{\lambda}$) relative to advertiser gain from being on the platform, i.e., $r + 2\overline{v}$ is large). This ensures that the market exists.

To shed some further light on the effect of competition intensity on equilibrium outcomes. Denote $\tilde{c} := \frac{\overline{\lambda}(\overline{\lambda}\gamma + \overline{v}\overline{\phi})}{4\overline{v}\gamma}$ an intermediate moderation cost which satisfies the assumption of convexity, we compute the (negative) derivatives of m_i^* and p_i^* with respect to τ. As a remark, we know that $\overline{\lambda} < r + 2\overline{v}$ to ensure non-negative prices when m^*. Then, call $p^{SH} := p_i^* = p_j^*$, $m^{SH} := m_i^* = m_j^*$. It follows that

$$-\frac{\partial m^{SH}}{\partial \tau} < 0, \qquad -\frac{\partial a^{SH}}{\partial \tau} < 0,$$

and

$$-\frac{\partial p^{SH}}{\partial \tau} = \frac{4cr\overline{v}^2(r + 2\overline{v} - \overline{\lambda})(4c\overline{v}\gamma - \overline{\lambda}(\gamma\overline{\lambda} + \overline{v}\overline{\phi}))}{(4c\overline{v}(4\tau\overline{v} + 3r\gamma) - \overline{\lambda}(2\tau\overline{v}\overline{\lambda} + r\gamma\overline{\lambda} - r\overline{v}\overline{\phi}))^2},$$

with the latter expression being negative for $c < \tilde{c}$, and positive otherwise.

Note that if $\gamma = 0$ (no nuisance from ads), then $-\frac{\partial p^{SH}}{\partial \tau} < 0$ and $-\frac{\partial m^{SH}}{\partial \tau} < 0$. This concludes the proof.

Competition with multihoming users

In social networks, given the absence of a real monetary cost paid by users, multihoming decisions are most dominant. In this section, we relax the assumption of singlehoming users to verify how content moderation changes when this is the case. One can easily note that when some users multihome, multihoming advertisers might interact with the same users twice, thereby placing wasteful ads. This might force the platform to reduce the ad price or please advertisers with tight content moderation policy. On the other hand, the presence of multihoming users relaxes the competition between platforms for the marginal consumer and this reduces the incentive to engage in a lax moderation policy.

To understand the optimal business strategies, we present the following variations of the model with singlehoming users. The utility of singlehoming users is as the

benchmark model, whereas the utility of the multihoming consumer on the platform i is $U_i^{mh} = (1+\sigma)u + \phi(\theta(m_i) + \theta(m_j)) - \gamma(a_i + a_j) - \tau$, with $\sigma \in (0,1)$ represents the marginal utility that consumers gain when joining the second platform. By equating the utility of the consumer singlehoming on platform i and multihoming, we can derive the following critical values which identify the users indifferent between singlehoming on platform 1 and multihoming and singlehoming on platform 2 and multihoming.

$$\tilde{x}_1 \equiv 1 - \frac{u\sigma + \phi\theta(m_2) - \gamma a_2}{\tau}, \qquad \tilde{x}_2 \equiv \frac{u\sigma + \phi\theta(m_1) - \gamma a_1}{\tau}$$

which implies that the demand of platform 1 is equal to $n_1 := \frac{1}{\overline{\phi}} \int_0^{\overline{\phi}} \tilde{x}_2 \mathrm{d}\phi$, and $n_2 := \frac{1}{\overline{\phi}}(\overline{\phi} - \int_0^{\overline{\phi}} \tilde{x}_1 \mathrm{d}\phi)$. However, as the multihoming advertisers observe some users twice, the value of their interactions is generated only once. This implies that multihoming users, n^{MH} account for only half of the value on each platform. Such an assumption finds justification in a recent work by Liu (2018). Indeed, an advertiser joining platform i obtains the following utility

$$V_i = v + r(n_i^{SH} + \frac{n^{MH}}{2}) - \lambda\theta(m_i) - p_i$$

where $n_1^{SH} = \frac{1}{\overline{\phi}} \int_0^{\overline{\phi}} \tilde{x}_1 \mathrm{d}\phi$, $n_2^{SH} = \frac{1}{\overline{\phi}}(\overline{\phi} - \int_0^{\overline{\phi}} \tilde{x}_2 \mathrm{d}\phi)$, and $n^{MH} = \frac{1}{\overline{\phi}} \int_0^{\overline{\phi}} (\tilde{x}_2 - \tilde{x}_1) \mathrm{d}\phi$.

Using the above expressions, we can determine the number of advertisers joining each platform. Platforms maximize profits as in the benchmark model and concavity requires a sufficiently large c. Note that the same condition is required when considering singlehoming users. In equilibrium, when $\overline{\lambda}$ is such that an interior solution exists, we obtain the same results as with singlehoming consumers. Hence,

$$p^{MH} = p^{SH}, \qquad m^{MH} = m^{SH}.$$

with $m^{MH} = m_i^* = m_j^* \in [0,1]$ and with $\overline{\lambda} < r + 2\overline{v}$ as in the main model. As the equilibrium outcome remains unchanged, also the derivatives are unchanged. Hence, allowing for multihoming consumers does not change our main results.