LEONARDO MADIO AND MARTIN QUINN

USER-GENERATED CONTENT, STRATEGIC MODERATION, AND ADVERTISING[1]

Abstract. 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.

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.

[1] The authors thank Luis Abreu, Malin Arve, Luca Ferrari, David Henriques, Laurent Linnemer, Christian Peukert, Carlo Reggiani, PatrickWaelbroeck 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), and the *Giorgio Rota Best Paper Award Conference* (Centro Einaudi, Turin, 2020). Leonardo acknowledges nancial 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.

Advertisers' exposure on these platforms is not risk-free. As most contents are generated or uploaded by users, it lacks external and professional validation (Allcott and Gentzkow 2017). As a result, it is often the case that online material is inappropriate, harmful, or even illegal. Recent estimates suggest that approximately 4-10% of display advertising does not meet brand safety requirements and the majority of content can be classied at a moderate risk level (Plum 2019). The recent story of social media platforms is full of examples and scandals, which raised several concerns on howplatforms deal with what is posted online. In June 2020, several inuential 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 dealing with protests for failure over content moderation. 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 usergenerated content, e.g., racist, extremist, and unsafe content.[3] Subsequently, they suspended their marketing campaign: some reduced their ad expenditure up to 70% in light of the extensive user market coverage the platform had. Others, instead, returned to the platform after a temporary pullback. The reason was distinctly expressed by the Association of National Advertisers, who argued that that because of such scandals, "reputation […] can be damaged or severely disrupted".[4]

To contain the scandals, YouTubewas forced to intervene by tightening its moderation policy, by shutting down 400 channels (including popular YouTubers such as PewDiePie), and by removing thousands of comments and videos. These interventions were 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 –

[2] See *The Brands Pulling Ads From Facebook Over Hate Speech*, «The New York Times» (https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html).
[3] See *YouTube Adpocalypse*, «Fandom» (https://youtube.fandom.com/wiki/YouTubeAdpocalypse). See also *A timeline of the YouTube brand safety debacle*, «Digitalcontentnext», March 31, 2017 (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.
[5] See e.g., https://support.google.com/youtube/answer/9194476.

Oversight Board – to decide which content could 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 prices that advertisers pay to reach 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-of. On the one hand, the platform has incentives to invest in content moderation to create a safe environment for advertisers. 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 revenue. 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 advertiser's risk of being displayed next to unsafe content, but it also allows to reach a larger 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]

We find that the marginal gains from moderation depend on the direct and indirect effects that a stronger moderation policy entails. The direct (positive) impact leads to more impressions, which may create a disutility for users if ads are not informative. The indirect (negative) effect leads to fewer users on the platform and, as a result, fewer impressions. Interestingly, such a trade-of depicts a non-monotonic relationship between the optimal content moderation policy and the price advertisers pay to be on the platform. When the cost of moderating content is sufficiently small, the platform always increases its moderation effort if advertiser sensitiveness to brand risk

[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., *Facebook' sorry' for distressing suicide posts on Instagram*, BBC, 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 to help advertisers create a list of possibly sensitive videos.

[7] In other cases, such as YouTube, strict regulation on cannabis and rearm-related content fuelled new niche platforms such as TheWeedTube.com and Full30.com. See *After the porn ban, Tumblr users have ditched the platform as promised*, «The Verge», March 14, 2019 (https://www.theverge.com/2019/3/14/18266013/tumblrporn-ban-lost-users-down-traffic). See also *The road to becoming a weedtuber isn't easy*, «Leafbuyer», November 10, 2018 (https://www.leafbuyer.com/blog/weedtube/).

increases. Notwithstanding, the ad price is U-shaped in the brand risk and the highest price is set for very high or minimal brand risk. The reason is that when brand risk is small, advertisers care more about the customer reach and, hence, the platform can set a very high price. On the contrary, when brand risk is very high, the platform prefers to moderate all content and set a very high price to compensate for its investment.

The relevance of moderation costs in shaping platform behaviour also emerges when these costs are very large. This is the case – for example – of small entrant platforms which may face signicantly high cost for moderating content due to scarcity of past data or lack of state-of-the start equipment. Likewise, it can also be the case of language barriers or when the manifestly unlawful content and not-manifestly unlawful – but still harmful for advertisers – content becomes narrow. We find that when moderation costs are sufficiently high, instead, content moderation decreases has an inverted U-shaped relationship, such that it initially increases up to the point in which moderation becomes so costly that the platform finds it optimal to disinvest. In other words, the platform stops moderating content because it gets too expensive to accommodate advertiser preferences without losing customers. In this case, the ad price always decreases with brand risk.

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 hosted on the platform: safe and unsafe ones. The first type always benefits users and advertisers. 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.[9] We model the presence of brand safety issues as the net value that advertisers obtain from joining a platform with a certain amount of unsafe content. However, the platform can indirectly control their presence and virality of unsafe content by investing in content moderation (and changing their terms and conditions for its users) such as hiring human content moderators and investing in monitoring

[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).
[9] 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).

and AI-based content moderation. The stricter a platform content moderation policy, the lower the share of inappropriate content, the smaller the brand risk advertisers may face.

Our main analysis is performed by looking at the strategies of a monopolist platform and results hold in very general settings. A natural variation of our model is to consider how platform competition influences the incentives to invest in content moderation. We therefore present a Hotelling setup in which two (horizontally) differentiated platforms compete for user attention. In such a scenario, as platforms become more substitutable from the consumer perspective (e.g., more intense competition, lower switching costs), platforms react accordingly by lowering their content moderation effort and increasing or reducing the price advertisers pay to place their ads. The rationale is that as competition intensies, the marginal users become more valuable from the consumer perspective which can be attracted by lowering content moderation and reducing the nuisance they face in the presence of ad impression. If content moderation is sufficiently costly, platforms prefer to be more lenient with unsafe content and charge more advertisers because of the larger customer audience ensured. On the contrary, a more tolerant content moderation policy is associated with a lower ad price if content moderation is not very expensive. Indeed, this would compensate advertisers for possible brand safety issues.

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 common" in light of dangerous misinformation and hate speech circulating on social media platforms.[10] Specifically, we find that the introduction of a fixed tax per ad placed on the platform has twofold effects. First, it reduces the incentives to invest in content moderation. Second, it can lead to a higher or lower price than in an environment with tax-free ads. The reason is that there is a first-order pass-through of the tax on the ad price. However, due to the lower content moderation, there is a second-order effect such that the ad price decreases (to compensate advertisers for the increased brand risk). Depending on the prevailing effect – which is linked to the cost function's convexity – the price can either increase or decrease.

[10] P. Romer, *A tax that could fix Big Tech*, «New York Times», May 7, 2019 (https://www.nytimes.com/2019/05/06/opinion/tax-facebook-google.html).

Our results provide implications for marketers and policymakers. As discussed, brand safety has become paramount in recent years and major brands coordinated their actions to induce platforms to tackle the problems of content moderation. However, these actions are unlikely to be successful if, on the other side of the market, there is a demand for controversial, viral, or potentially harmful content. The same problem would arise in the presence of users reluctant to forms of control of their expression online, especially for content whose identification can be challenging for automated tools. While our model accounts for the direct negative externality that the presence of potentially harmful content entails, our results can be relevant to discuss the platforms' incentives when the harmed party is external to the platform environment. For example, this can be the case of inappropriate content that may cause long-term negative externalities for society, e.g., fake news impacting election outcomes (Allcott and Gentzkow 2017) or leading to vaccine hesitancy (Carrieri *et al.* 2019). In European Union and in the United States, policymakers have started considering upgrades of the current liability regimes applied to online intermediaries and stricter regulation may impose to platforms procedural obligations and duties at least concerning manifestly unlawful content and hate speech.[11] Similarly, code of conducts on disinformation may reduce the extent to which advertisers may be exposed to unsafe content.

A second result drawing policy implications concerns the typical concern characterizing markets with strong network externalities and the winner-takes-all scenarios (see, e.g., Furman *et al.* 2019). Our results suggest that absent regulatory tools or changes in 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 distortion regarding both ad pricing and content type and configure a market failure.

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 2018) and concerns the media outlet provisions of news and other content. Other

[11] In the US, platforms are considered hosting service providers and, hence, exempted from liability (US Communication Decency Act Section 230). Under the European E-Commerce Directive (2000), platforms can benefit from a conditioned liability exemption, depending on the knowledge standard of the illegal activity carried out on the platform and their passive role in the distribution of the information. In 2020, the European Commission has launched the Digital Services Act to upgrade liability rules for platforms.

studies in the marketing literature looks at UGC in their forms 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 advertisers' aversions to brand-risk has signicant 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.[12] The ad-targeting literature is perhaps the closest to the spirit of our study. This literature generally assumes a better match between the user's preference and the advertisers' 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.[13] 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 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.

[12] The literature on advertising and media has mainly focused on 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-saws (Anderson and Peitz 2020).
[13] For a review, see e.g., Gentzkow *et al.* 2015.

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) study competition on content quality (real or fake news) between media outlets and find 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) investigate media bias in a market in which platforms sell content to readers and profit from advertisers. They give the power to platforms to change the accuracy of the news. Such a lever can have a signicant effect as a lack of accuracy 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 profits. In this case, the platform might influence that match by moderating content (more) carefully.

In the framework of media bias, Mullainathan and Shleifer (2005) show that when newspapers compete for user demand, there is an incentive to exaggerate media bias. Similarly to ours, Gal-Or *et al.* (2012) study 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 intensies. This results in more slanting to soften competition and greater polarization of readers. In our model, instead, when competition intensifies, content moderation becomes more tolerant 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) study Facebook's decision in 2016 to ban ads linking to external websites fabricating fake news. They find that the ban was effective: fake news declined more on Facebook than on Twitter after the policy. Rao (2018) documents 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) motivate our analysis on platform heterogeneity in moderation policies. They show that Facebook was more prone than Twitter in banning fake and false news, underlying platform heterogeneity.

ARTICLE STRUCTURE. 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. Such an assumption can be justied 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. 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.[14]

Users can consume two types of content: a mass 1 of safe content and a mass (m) of unsafe content. The former, which identifies professional videos and news, pictures of vacations and pets, entail positive benefits for both users and advertisers. For advertisers, one can imagine a positive match value when impressions are just next to these contents. The latter, instead, identifies controversial and possibly harmful content. For instance, these can be borderline comments which userswant 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 selects and which is identified by the parameter $m \in [0, 1]$, 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 dangerous), whereas with $m = 1$the platform moderates all content.

THE PLATFORM. There is an ad-funded platform that charges a zero price to users and lets advertisers (acting on behalf of brands) pay for launching an ad campaign at price *p*. We assume that advertisers do not compete for ad space and they launch at

[14] See e.g., *Additional changes to YouTube partner*, YouTube (https://youtubecreators.googleblog.com/2018/01/additional-changes-to-youtube-partner.html.

most one ad campaign. We denote the number of advertisers joining the platform by $a(m,p)$. The platform maximizes profits by choosing the price p and investing in costly content moderation $C(m)$. We assume that content moderation is suciently 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 $C(m)$. 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 platform costs are likely to be much larger as requiring additional investments in text analysis.[15] Similarly, while AI tools and filters based on tags and keywords can have benefits, some content may require post-human moderation, therefore leading to much higher prices. All these costs are taken into account by a platform when choosing ad prices and content moderation policies. Platform's profits can then be summarized as follows:

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

INTERNET USERS. There is a unit mass of Internet users. Each user is identied by the duple (u,ϕ) that captures her taste for "safe", u, and "unsafe" content, $\varnothing$. Specically, we assume that the preference for safe content is distributed according to the following parameter $u \in [0,\overline{u}]$. Users are also differentiated according to their taste ϕ for unsafe content, with $\phi \in [u,\overline{u}], \overline{u} > 0$. Note that while the sign of ϕ is positive, the sign of ϕ is unspecified. When this is negative, (some) users gain from content moderation, whereas when it is positive, all users dislike content moderation. Moreover, at this stage, we do not put any restrictions on the distribution function form of u and ϕ and assume both distributions are independent of one another. Moreover, we also assume that users dislike ads as being perceived as a nuisance cost, with $\gamma > 0$ identifying this parameter. The total nuisance cost to which users are

[15] 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.

therefore exposed is then equal to $\Upsilon \times a(p, m)$. The utility of the users when joining a platform is

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

ADVERTISERS. There is also a unit mass of advertisers that decide whether to launch their ad campaign on the platform depending on their long- and short-term profitability. The utility of an advertiser can be expressed as follows:

$$V = \pi(n, \Omega) - p, \tag{3}$$

where $\pi(n, \Omega)$ captures the profitability of the ad campaign and p is the price paid to the platform. For short-term profitability, $\pi(n, \cdot)$, we intend revenues obtained from the interaction with the n users the platform attracts. Online interactions yield a stream of (exogenous) revenue r. The higher r, the larger the advertisers' cross-network externalities. Thisway, $r \times n$ can represent revenues from individual clicks or the possibility to obtain short-term after-market transactions when users buy products in-store or online. For long-term profitability, $\pi(\cdot, \Omega)$, we denote the impact of a brand's (long-term) reputation. As discussed in the introduction, advertisers are increasingly concerned about the impact of scandals on their reputation. As many marketers argued when urging digital platforms to tackle misinformation, racism, and hate speech, this impact is not directly channeled through a reduction in sales or click rates but via reputation which contributes to a significant share of a firm's value (Jovanovic 2020).

Unlike the previous literature dealing with the ad market, we assume that brands (via advertisers) care about the suitability of the environment in which impressions appear. Hence, advertisers benefit from the presence of a mass of safe content according to a parameter $\upsilon \in [0, \overline{\upsilon}]$ but face a disutility, $\lambda \in [0, \overline{\lambda}]$ from the presence of a mass (m) of unsafe content. Formally, can be expressed as

$$\Omega = 1 \times \upsilon - \lambda\theta(m). \tag{4}$$

Note that this very general specification captures the large heterogeneity across advertisers' benefit from being displayed just next to a 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 high). On the contrary, unsafe content can have a small impact on advertisers promoting gambling websites (i.e., a small λ). Hence, advertisers' utility in Equation (3) can be written as follows:

$$V = rn + \upsilon - \lambda\theta(m) - p. \tag{5}$$

TIMING. The timing of the game is as follows. In the first stage, the platform maximizes profits and chooses both ad price and the 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 game is solved backward and the equilibrium concept is subgame perfect.

2.1 Optimal content moderation

We first compute the level of activity on the platform. Following Rochet and Tirole (2003) and using equations (2-5), the number of users joining the platform can be written as and the number of advertisers placing their ads as $n = \Pr(U \geq 0)$. Formally, this implies

$$\begin{aligned} a &= \Pr(\upsilon - \lambda\theta(m) + rn - p \geq 0) \equiv D^a(p,n,m), \\ n &= \Pr(u - \gamma a + \phi\theta(m) \geq 0) \equiv D^n(a,m), \end{aligned} \tag{6}$$

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 $n = d^n(p,m)$.[16] In the first stage, the platform chooses m and p to maximize $\Pi = d^a(p,m)p - C(m)$. Denote by Ψ the *elasticity of profit with respect to moderation* such that

$$\Psi = \underbrace{\frac{\partial D^a}{\partial m}}_{\text{Brand safety effect}} + \overbrace{\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}}^{\text{Eyeball effect}}. \tag{7}$$

[16] For more details, see Rochet and Tirole (2003).

From the first-order conditions, we obtain:

$$\begin{aligned} p =& -\frac{d^a(p,m)}{\frac{\partial d^a}{\partial p}} = -\frac{d^a(p,m)(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})}{\frac{\partial D^a}{\partial p}}, \\ C'(m) =& p\frac{\partial d^a}{\partial m} = p\Big(\frac{\Psi}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}\Big), \end{aligned} \tag{8}$$

and the marginal gain from moderation, which implicitly defines the optimal level of content moderation, is summarized by the lemma below.

Lemma 1. Define the following as the platform marginal revenue from content moderation

$$MR(m^*,p^*) = -\frac{d^a(m^*,p^*)\Psi}{\frac{\partial D^a}{\partial p}}. \tag{9}$$

The optimal content moderation is implicitly defined by the following expression such that marginal costs equal marginal revenue:

$$\begin{aligned} C'(m^*) =& -\frac{d^a(m^*,p^*)\Psi}{\frac{\partial D^a}{\partial p}}. \\ MC(m^*) =& MR(m^*,p^*). \end{aligned} \tag{10}$$

Note that the optimal moderation policy is chosen in away such that the marginal gains/revenues (MR) from moderation equal the marginal costs (MC). Importantly, due to the multisidedness of the market, the latter account for the price the platform selects, the effect that content moderation has on advertiser demand, and how consumer and advertiser react to increased brand safety (via $\boldsymbol{\Psi}$). One can easily see that the larger $\boldsymbol{\Psi}$, the larger the gains from content moderation. In the limit case in which $\boldsymbol{\Psi < 0}$, the platform sets $m^* = 0$ and advertisers are associated with potentially harmful content.

To shed some further light on how the advertiser sensitiveness to brand risk impacts equilibrium outcomes, we present some simple comparative statics on how the optimal price and moderation react to an increase in brand risk (a higher $\bar{\lambda}$). As brand risk is only contained in $\boldsymbol{\Psi}$ and demand forms, we investigate how p and m changes with $\boldsymbol{\Psi}$ – the elasticity of platform profits with respect to content moderation. The next proposition summarizes the main findings and highlights the relevance of moderation costs.

Proposition 1. *There exists a cut-off*

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

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

Proof. See Appendix A.

Specifically, when moderation costs are suciently small, a platform can easily adjust its moderation effort depending on how many users and advertisers can attract and the surplus to be extracted. This way, when brand risk increases – and the advertiser's willingness-to-pay to advertise decreases – the platform optimally adjusts its moderation effort. As a result, the moderation effort m^* monotonically increases with Ψ up to the point in which full moderation $m^* = 1$ is ensured. When Ψ is large enough, the brand safety effect largely outweighs the eyeball effect. In practice, the platform prefers to have fewer users to safeguard advertisers' reputation. For instance, platforms may be more meticulous when attracting advertisers promoting luxury goods or charities, whose reputational losses from scandals might be substantial.

By contrast, when moderation costs are sufficiently high, if advertiser brand risk increases, the platform faces increasingly high costs to satisfy their requests and this is not compensated by a large market capture on the consumer side. As a result, the moderation effort is inverted U-shaped: it monotonically increases to the point in which accommodating advertiser preferences becomes too expensive in terms of investments and in the number of consumers exiting the platform.

Then, the moderation effort m^* starts decreasing with Ψ to the point in which contents are no longer moderated, $m^* = 0$. Interestingly, for very high brand risk with high moderation cost, the platform prefers not to moderate content at all.

A similar discussion also applies to the effect of Ψ on ad prices, which presents some nonmonotonicity. To see why, we first analyze the situation where moderation costs are small. In this case, moderation increases with Ψ and the ad price is U-shaped. Namely, the platform 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 sells a very large number of user eyeballs to advertisers and given the low risk of being exposed to harmful material, the price can be high aswell. At the other extreme, the platform sacrices some audience and meets

the moderation requests of advertisers which - 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, nudity and sexual content, intellectual property rights violation as well as the veracity of the news. The ad price reaches a minimum when

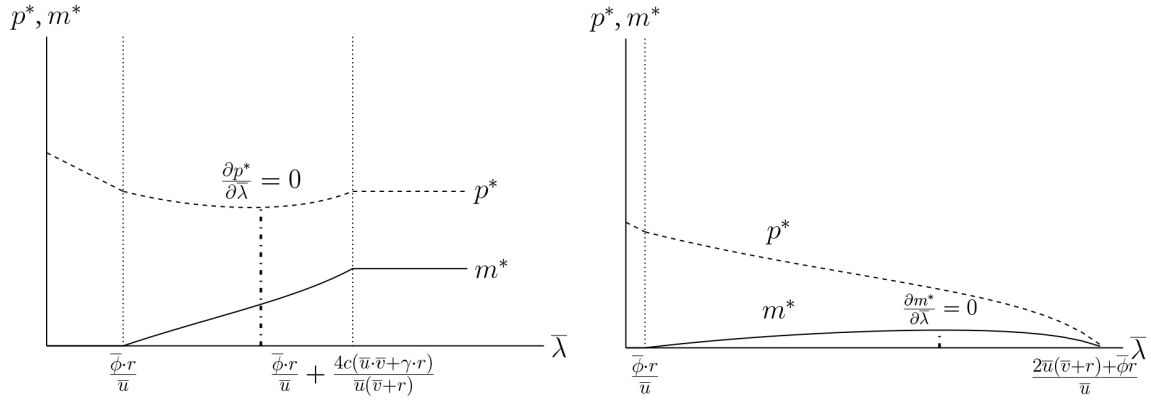
$$-2\frac{\partial D^a}{\partial p}d^a(m^*,p^*)=\frac{\Psi\dfrac{\partial D^a}{\partial \Psi}\dfrac{\partial D^a}{\partial p}}{1-\dfrac{\partial D^a}{\partial n}\dfrac{\partial Dn}{\partial a}},$$

where p^* reaches a minimum in $\overline{\lambda}$ 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 suciently small, the platform increases moderation, but the way it increases does not compensate advertisers for the consumers who exit the platform. As a result, the price goes down.

However, suppose advertisers become too sensitive to unmoderated content. In that case, the marginal revenues from increased moderation become lower than the marginal costs of moderation (in terms of intrinsic moderation costs and consumers exit), so the platform starts reducing moderation and compensates the advertisers for the very high risk of being exposed to unsuitable content. In turn, the ad price decreases.

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 advertiser aversion to brand risk when moderating costs are small (left) and large (right).

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)



3. PLATFORM COMPETITION

A natural variation of the benchmark model is the introduction of 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 advertising revenues and on the provision of UGC. In this subsection, we study a model of platform competition in the presence of full market coverage. We then study how the intensity of competition influences content moderation policy and ad pricing strategies.

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 $i = 1, 2$ for the entire ad campaign and a_i represents the number of advertisers deciding to buy a space on the website. Hence, platforms' profits are defined as follows

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

Consistently with the previous literature (Anderson *et al.* 2016), we let advertisers multihome. Hence, platforms characterize a competitive bottleneck and each platform becomes the only way to reach unique users. This implies that platforms compete for attracting consumers. We assume that platforms are symmetric and we look for a

symmetric equilibrium. As in the presence of a monopolist, advertisers are defined by a duple $(\upsilon, \lambda) \in [0, \overline{\upsilon}] \times [0, \lambda]$, with a uniform distribution of υ and λ. This setting is an adaptation of a two-dimensional differentiation as in Anderson and Gans (2011), Economides (1986) and Vandenbosch and Weinberg (1995). Their utility when patronizing platform *i* is

$$V_i = \upsilon + rm_i - \lambda\theta(m_i) - p_i. \tag{11}$$

There is a unit mass of users and we assume that the market is fully covered. Users are are uniformly and independently distributed on a line of unit length; they are identied by a duple relative to their relative preference for platform *i(j)*, defined by their position x 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., ϕ in Section 2, is set equal to zero.[17]

The utility of a user located at x and joining platform *i* is as follows:

$$U_i = u + \phi\theta(m_i) - \gamma a_i - \tau \mid x - l_i \tag{12}$$

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

The timing of the game is as before. 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 Internet users decide which platform to join. We look for a symmetric equilibrium.

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.

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

[17] This simplifying assumption allows us to focus on the case in which advertisers' and users' preferences over moderation are conicting - 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.

By solving the model backward, the number of advertisers is

$$a_i(n_i, n_j) = 1 + \frac{rn_i - \overline{\lambda}\theta(m_i) - p_i}{\overline{v}}, \tag{13}$$

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

$$n_i(a_i, a_j) = \frac{1}{2} + \frac{\overline{\phi}(\theta(m_i) - \theta((m_j)) - 2\gamma(a_i - a_j)}{4\tau}, \qquad n_j = 1 - n_i. \tag{14}$$

Rearranging the above expression, we obtain the second-period market shares of platform *i* for both advertisers and users, respectively denoted by d_{ai} and d_{ni}. In the first stage, platforms make their decision on moderation and ad price simultaneously and non-cooperatively to maximize

$$\max_{p_i, m_i} \Pi_i = d_{a^i}(p_i, p_j, m_i, m_j)p_i - C(m_i). \tag{15}$$

We assume that profits are well-behaved as long as $C(m_i)$. is suffciently convex. For the sake of simplicity, we let moderation costs be quadratic (i.e., $C(m_i) = cm_i^2 / 2$). As in the benchmark case, it is also assumed that $\theta(m) = 1 - m$.

$$\lambda_{11} \equiv \frac{\overline{\phi} r \overline{v}}{(\tau \overline{v} + \gamma r)}, \quad \lambda_{12} \equiv \frac{4c(4\tau\overline{v} + 3\overline{\gamma}r)}{(2\overline{v} + r)(2\overline{v}\tau + \gamma r)} + \frac{\overline{\phi} r \overline{v}}{2\tau\overline{v} + \gamma r},$$

Define the following cut-of value of λ:

$$\lambda_{11} \equiv \frac{\overline{\phi} r \overline{v}}{(\tau \overline{v} + \gamma r)}, \quad \lambda_{12} \equiv \frac{4c(4\tau\overline{v} + 3\overline{\gamma}r)}{(2\overline{v} + r)(2\overline{v}\tau + \gamma r)} + \frac{\overline{\phi} r \overline{v}}{2\tau\overline{v} + \gamma r},$$

and by solving for the symmetric equilibrium, we can state the following proposition.

Proposition 2. *When competition takes place, a symmetric equilibrium exists as follows.*

(a) When $\overline{\lambda} \leq \lambda_{11}$, *the platform enforces no moderation, m* = 0, and sets the following ad price*

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

(b) For any $\lambda_{11} < \lambda < \lambda_{12}$*, Nash equilibrium outcomes are an interior solution such that*

$$p_i^* = p_j^* = \frac{4c\bar{v}(2\bar{v} + r - \bar{\lambda})(\tau\bar{v} + \gamma r)}{4c\bar{v}(4\tau\bar{v} + 3\gamma r) + \bar{\lambda}(\bar{\phi} r\bar{v} - \bar{\lambda}(2\bar{v}\tau + \gamma r))},$$
$$m_i^* = m_j^* = \frac{(2\bar{v} + r - \bar{\lambda})(\bar{\lambda}(2\tau\bar{v} + \gamma r) - \bar{\phi} r\bar{v})}{4c\bar{v}(4\tau\bar{v} + 3\gamma r) + \bar{\lambda}(\bar{\phi} r\bar{v} - \bar{\lambda}(2\bar{v}\tau + \gamma r))}.$$

c) When $\bar{\lambda} \geq \lambda_{12}$*, platforms set* $m^*_i = m^*_j = 1$ *and*

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

Proof. See Appendix A.

The analysis of the symmetric equilibrium shows that when there exists a sufficiently small brand risk, that is, whenever $\bar{\lambda}$ is sufficiently low, platforms enforce no moderation policies - i.e. $m_i^* = 0$ - and set low ad prices to attract as many advertisers as possible. For intermediate brand risk, $\lambda_{11} < \lambda < \lambda_{12}$, the optimal moderation policy is an equilibrium solution, $m_i^* \in [0,1]$. In this case, platforms can increase their ad price as advertisers are willing to pay more given the reduced brand risk.[18] Finally, when brand risk is high, the platforms respectively enforce a full moderation policy – $m_i^* = 1$ – and set a very high ad price.

However, such results may depend on the intensity of competition in the market, which in our case is proxied by the degree of platform differentiation. To better grasp this effect, we compute derivatives of m_i^* and p_i^* with respect to τ. When τ decreases, product differentiation reduces and, in turn, competition intensies. The following proposition summarizes this result.

Proposition 3. *Let* $\tilde{c} := \frac{\bar{\lambda}(\bar{\lambda}\gamma + \bar{v}\bar{\phi})}{4\bar{v}\gamma}$ *, for any* $m_i^* \in [0,1]$*, fiercer market competition on the user side leads to (a) lower content moderation; (b) a price reduction (increase) for suciently small (large) moderation costs (*$c < (>)\tilde{c}$ *) and (c) fewer ads displayed to users.*

[18] Note that, in both cases, a condition to ensure a non-negative price is such that $\bar{\lambda} < r + 2\bar{v}$. This implies that advertiser's brand risk is not as high (i.e., low enough $\bar{\lambda}$) relative to advertiser gain from being on the platform, i.e., $r + 2\bar{v}$ is large). This ensures that the market exists.

Proof. See Appendix A.

Proposition 3 shows that when competition for users becomes ercer, 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. Onthe other hand, they can reduce the number of ads and, therefore, their nuisance. In equilibrium, the mechanism works as follows. When moderation is sufficiently expensive, content moderation is already low. In this case, the onlyway to attract users is to reduce the number of ads by increasing the ad price. This increases platforms' profits and attracts additional users. 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 with a reduction in the price. In turn, this 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. Different is the effect on the advertiser side: as competition gets fiercer, the number of ads placed on each platform decreases regardless of the pricing strategy. As ads are considered a nuisance cost, this turns out to increase the user welfare. The above proposition also has a relevant implication. Typically, fostering more competition in the market is advocated by policy-makers 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.

Similarly, authorities other than competition ones are concerned with potential societal externalities stemming from the uncontrolled presence of UGC. For instance, negative externalities can result from misinformation, hate speech. Absent other interventions or the possibility to enforce platform liability given the existing framework in Europe and the US, well-intended policy measures aimed at increasing competition in the market are likely to generate negative externalities. First, advertisers would face a higher brand risk without observing any demand expansion.[19] Second, negative societal externalities may arise if UGCs are perceived as harmful from policy-makers, though not always persecutable in courts.

[19] In our framework, this depends on the Hotelling structure of the model and of the full market coverage assumption.

4. DISCUSSIONS 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 started cosidering the introduction of a tax on online ads to create a fairer environment. More related to the aim of this paper, in 2019, the Nobel Prize laureate Paul Romer proposed the introduction of a tax on digital ads as a measure to induce social media platforms to limit misinformation.

To shed some light on the possible unintended effects of the introduction of such a tax, we modify our benchmark model, and we assume that the Government (exogenously) imposes a tax f on each ad. As a result, the Government can raise af, which drives platform profits to be $\Pi = d^a(p,m)(p-f) - C(m)$. As taxes impact the platform's marginal profits, we expect it to affect the price advertisers pay and accordingly, the content moderation decided by the platform.

Using the Implicit Function Theorem and the Cramer's rule, the introduction of a tax on digital ads determines the following results.

Proposition 4. *The ad price increases (decreases) with a tax on platform's revenues*

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

The optimal moderation policy also decreases with f such that

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

Proof. See Appendix A.

Proposition 4 underlines very interesting results. First, 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. This is common in optimal taxation theory as the platform is just passing the tax into the price.[20] However, a second-order effect implies a reduction in the platform moderation effort, which, in turn, decreases the ad price. One effect dominates the other depending on moderation cost. Specifically,

[20] See Mankiw *et al.* (2009) for a complete review on optimal taxation theory.

when moderation costs are suciently low, the negative indirect effect dominates as moderation decreases faster with a tax. In this case, ad price also decreases faster that it increases with the direct effect. When moderation costs are high enough, the opposite is true.

Second, the moderation policy always decreases with a tax. This is because the tax directly reduces the marginal revenues from advertisers. Hence, the higher the tax, the lower the marginal revenue from moderation enforcement, the lower the moderation effort. All in all, as in the benchmark model, the effect on moderation is aligned with advertisers' interests. As content moderation is relaxed, also advertisers place fewer ads. To better understand the above mechanisms, in Appendix B, we provide an example with a uniform distribution of preferences.

4.2 Endogenous content creation

In the main specication, 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 (i.e., 1) and unsafe content (i.e., $\theta(m)$). As before, we assume that the platform accepts all safe materials, whereas it moderates the unsafe ones.

Indeed, we explicitly model the presence of content creators among the users, who obtains a utility $U_\theta = u_\theta + nk - m$ when creating unsafe content,[21] 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 may be heterogeneous on the support $[0, \overline{u}_\theta]$ with $u_\theta < 1 - k$ 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 from moderation are equal to

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

[21] 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.

with

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

The above expression is an augmented version of the one presented in Lemma 1. However, it differs in the inclusion of indirect network externalities stemming from content creators' presence and their interest for a broad audience base. Such a result suggests that a platform is pursuing a stricter content moderation policy while pleasing risk-averse advertisers may dissatisfy both creators and users. This is entirely in line with the "Tumblr spiral". After the acquisition of Yahoo from Verizon and the 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.

4.3 Targeting

So far, we have not considered the possibility that the platform(s) can target users. 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 eventually control γ. If γ were considered equal to 0, such that ads are 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: ad prices would always be reduced when competition intensifies. When the competition for users becomes more intense, the platform no longer needs to compete by lowering the nuisance costs to users by increasing the ad price. As a result, a more intense competition leads to a reduction of both content moderation and ad price.

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".[22] This may ensure some forms of safeguards for brands and marketers. However, targeting is far from perfect (Nielsen 2018), and better

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

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 targeted moderation is imperfect.

4.4 Other Applications

OFFLINE NEWS OUTLETS. Our setting can provide more general insights into content moderation 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. 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.[23] Such a story well fits the trade-off that traditional media outlets may face when producing content. We discuss this by making two contributions.

Consider a news outlet that only produces professional content that is sufficiently attention grabbing to be attractive for users but also allows advertisers to place their ads. Hence, this outlet would strategically choose the sensitivity of materials to produce to balance user attraction and advertisers' exit. 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, we may imagine that producing professional content is cheaper than moderating thousands of comments and posts online. In this case, our framework indicates that competition between outlets would make content quality going down while the ad price goes up.

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 such a case, the aggregator would directly balance user and

[23] See The Co-operative Group, *An update on our advertising policy* (https://blog.coop.co.uk/2017/03/23/anupdate-on-our-advertising-policy/).

advertiser preferences when choosing the type of content to produce and display to (safely) monetize users' eyeballs. Indeed, a content aggregator would need to balance users' attraction strategically and advertisers' brand safety concerns. Such a set-up allows us to endogenize the platform's design choice that consists of accepting or not UGC to be displayed on the platform. Depending on platform moderation costs and production costs, an outlet may be keener on introducing UGC or not on the platform. For instance, a high-end fashion website may only attract advertisers with high brand safety. In this case, we conjecture that when moderation cost is higher than production cost, such a website would prefer to produce its content rather than allowing moderate too costly UGC.

TV REALITY SHOWS. The framework we depict can also be applied to TV reality shows, such as the famous The Big Brother, which 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 of 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.[24]

5. MAIN HIGHLIGHTS AND CONCLUSIONS

The digital revolution has changed the production of media content. Whereas in the past, thesewere mostly produced by professionals (e.g., journalists), the advent of social mediawebsites has given users control over production and difusion of content. In most cases, this happened without any external and professional validation and created concerns among advertisers and marketers. This article studies the economic implications of such a situation and underlines the trade-of faced by a social media platform when strategically enforcing content moderation. In the following, we disentangle the importance of our results for both managers and policymakers.

MANAGERIAL IMPLICATIONS. This article provides a rationale for the signicant heterogeneity across platforms in tackling illegal, harmful, or disputed content. We argue that content moderation policies are rather platform-specific as depending on

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

the overall platform elasticity concerning moderation. This is due to the type of users and advertisers each social media attracts. Hence, we provide managerial implications for both brands and platforms.

First, the two-sidedness of the market is crucial for both advertisers and platforms. On the one hand, a platform should consider its moderation cost, audience type, and advertisers type when deciding to invest in content moderation. Only in this case, the platform will be able to balance advertising price and content moderation strategy that together maximize advertisers' willingness to pay. On the other hand, advertisers must be able to detect platform choice in scouting both moderation technology and audience type depending on their nature. For instance, old brands with inherited reputation should pay more attention to platforms pursuing lax content moderation and may decide to advertise only if the short term revenues do not jeopardize their brand image. On the contrary, young brands may care less about brand safety, hence maximizing short term revenue without imperiling their long term strategy.

Second, our results underline the importance of moderation costs in ad price and content moderation decisions. As shown in its moderation report, Facebook admits facing a cost to moderate that is idiosyncratic to countries, depending on language, culture, and characteristics.[25] Our analysis shows that for a monopolist moderation, costs can lead the platform to react differently to increase in brand risk - like the one presented in recent protests by advertisers.

More importantly, the advertisers push for more brand safety may not be supported by Big Tech if moderation costs are very large. This may lead to reduced content moderation if content moderation becomes very costly, perhaps because of too many content to be analyzed or different languages to be considered. Paradoxically, such an outcome is more likely to arise the more advertisers become concerned about brand risk. On the contrary, it is in the interest of the platform to accommodate advertisers' requests if moderation costs are sufficiently small.

Third, our results showthat absent platform liability, competition between platforms plays a crucial role. Specically, we find that as competition intensies, a social media platform would always decrease its content moderation, but must be more sophisticated about its price strategy. A first situation arises when moderation is expensive. In that case, content moderation is already low and the only way for the platform to compete for user attention is by reducing the nuisance costs users face.

[25] A summary of the report can be found here https://transparency.facebook.com/community-standards enforcement.

One way to do this is to increase the ad price, which reduces advertiser demand. Alternatively, the platform could also invest in targeting technologies. However, such a solution is unlikely to provide satisfactory results as content moderation and targeting typically exhibit type-I and type-II errors. A second situation emerges when moderation can be achieved at a low cost. In that case, content moderation is high and the platform may allow for a lot more unsafe content as it is very efficient in attracting new users. In this case, as advertisers bear some risks, the platform may be willing to reduce the ad price.

POLICY IMPLICATIONS. The above-described results are also of paramount relevance not only for marketers but also for policymakers. Social media moderation policies are not neutral and this article highlights that their decisions depend on the trade-of between generating revenues from advertisers and capturing user attention. At different institutional levels, it is widely debated what platforms should do to prevent the difusion of illegal content and misinformation going on social mediawebsites as their effects could be detrimental to society. For instance, 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 launches 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.[26] In 2018 the German Bundestag passed a law requiring platforms to remove hate speech within 24 hours or face nes of up to 50 million euro (see e.g., CEPS 2018). In this respect, we also discuss more broadly how well-intended policies aimed at stimulating more competition in digital markets might have the (unintended) effect of lowering platform incentives to invest in content moderation. Our results show that increasing competition between platforms is likely to reduce their moderation effort and distort pricing strategies on the advertising market.

In addition, we study the impact of an often advocated policy measure like the digital tax on advertising revenues. This was adopted in France, Germany, Italy, and recently supported by the Nobel Prize laureate Paul Romer. This article shows that these well-intended measures may have the perverse effect of reducing moderation

[26] See *Illegal content on online platforms*, «Digital Single Market» (https://ec.europa.eu/digital-singlemarket /en/illegal-content-online-platforms). Similarly, see e.g., T. Kaeseberg on «VoxEu», December 12, 2019, *Promoting competition in platform ecosystems* (https://voxeu.org/article/promoting-competitionplatform-ecosystems).

effort for the platform, thereby increasing the relevance of the current problem faced by democracies and advertisers. More complex settings of our setup may provide further insights. For instance, platform reputation may represent away to mitigate negative externalities fromUGC and induce more responsible actions. Similarly, the recent Cambridge Analytica scandal pushed Facebook to intervene to regain its user's trust. Moreover, an extension of this work may also consider a mixed business model and the incentives of these platforms to deal with content moderation when also users pay a subscription price.

References

Allcott H. and Gentzkow M. (2017), *Social media and fake news in the 2016 election*, «Journal of Economic Perspectives», 31, 2, pp. 211-236

Allcott H., Gentzkow M. and Yu C. (2019), *Trends in the difusion of misinformation on social media*, NBER Working Paper No. 25500

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, pp. 189-222

Anderson S.P. and De Palma A. (2013), *Shouting to be heard in advertising*, «Management Science», 59, 7, pp. 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, pp. 34-54

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

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

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, pp. 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, pp. 1574-1590

Bergemann D. and Bonatti A. (2011), *Targeting in advertising markets: implications for off-line versus online media*, «The RAND Journal of Economics», 42, 3, pp. 417-443

Besley T. and Prat A. (2006), *Handcufs for the grabbing hand? media capture and government accountability*, «American Economic Review», 96, 3, pp. 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, pp. 417-443

CEPS (2018), *Germany's netzDG: A key test for combatting online hate*, Centre for European Policy Studies (CEPS)

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, pp. 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, pp. 345-354

Chintagunta P.K., Gopinath S. and Venkataraman S. (2010), *The effects of online user reviews on movie box oce 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. (2018), *Social media and news: Attention capture via content bundling*, Mimeo

Economides N. (1986), *Nash equilibrium in duopoly with products defined by two characteristics*, «The RAND Journal of Economics», 17, 3, pp. 431-439

Ellman M. and Germano F. (2009), *What do the papers sell? a model of advertising and media bias*, «The Economic Journal», 119, 537, pp. 680-704

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

Furman J., Coyle D., Fletcher A., Marsden P. and McAuley D. (2019), *Unlocking digital competition*, Report of the Digital Competition Expert Panel, March

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

Gentzkow M. and Shapiro J.M. (2006), *Media bias and reputation*, «Journal of Political Economy», 114, 2, pp. 280-316

Gentzkow M., Shapiro J.M. and Stone D.F. (2015), *Media bias in the marketplace: Theory*, in *Handbook of Media Economics*, Elsevier, vol. 1, pp. 623-645

GlobalWebIndex (2019), *Social media flagship report* (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, pp. 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*, Elsevier, vol. 1, pp. 563-592

Mankiw N.G., Weinzierl M. and Yagan D. (2009), *Optimal taxation in theory and practice*, «Journal of Economic Perspectives», 23, 4, pp. 147-174

Mullainathan S. and Shleifer A. (2005), *The market for news*, «American Economic Review», 95, 4, pp. 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*, Elsevier, vol. 1, pp. 445-530

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, pp. 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, pp. 990-1029

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

Vandenbosch M.B. and Weinberg C.B. (1995), *Product and price competition in a twodimensional vertical differentiation model*, Marketing Science, 14(2):224–249.

Xiang Y. and Sarvary M. (2007), *News consumption and media bias*, «Marketing Science», 26, 5, pp. 611-628

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

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

Appendix A

Proof of Lemma 1

The number of users who decide to join 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 $n = d^n(m, p) \equiv D^n(m, p)$. Then, we can solve the model in the first stage of the game whereby the platform chooses m and p to maximize profits:

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

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

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

and with content moderation

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

The above expressions can be rearranged to obtain

$$\begin{aligned} \frac{\partial d^a}{\partial p} &= \frac{\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} < 0, \quad \frac{\partial d^n}{\partial p} = \frac{\frac{\partial D^n}{\partial a}\frac{\partial D^a}{\partial p}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} < 0 \\ \frac{\partial d^a}{\partial m} &= \frac{\frac{\partial D^a}{\partial m} + \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}0, \quad \frac{\partial d^n}{\partial m} = \frac{\frac{\partial D^n}{\partial m} + \frac{\partial D^n}{\partial a}\frac{\partial D^a}{\partial m}}{1 - \frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}0. \end{aligned} \tag{18}$$

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

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

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

$$MR(m^*) = -\frac{d^a(m^*,p^*)(\frac{\partial D^a}{\partial m}+\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial m})}{\frac{\partial D^a}{\partial p}} = -\frac{d^a(m^*,p^*)\Psi}{\frac{\partial D^a}{\partial p}}, \tag{19}$$

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 costs.

Proof of Proposition 1

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 understand how market parameters change in Ψ.

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

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

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

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

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

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

Using (18) and exploiting $\frac{\partial d^a}{\partial \Psi} = \frac{\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} < 0$, we have

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

Rearranging the above expressions, we have

$$-\frac{\partial D^a}{\partial \Psi} = 2\frac{\partial D^a}{\partial p}\frac{\partial p^*}{\partial \Psi} + \frac{\partial m^*}{\partial \Psi}\Psi,$$

and

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

Following the Implicit Function Theorem, then have

$$\begin{pmatrix} 2\frac{\partial D^a}{\partial p} & \Psi \\ \frac{\Psi\frac{\partial D^a}{\partial p}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} & \frac{\partial D^a}{\partial p}C''(m) + \frac{\Psi^2}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}\begin{pmatrix} \frac{\partial p^*}{\partial \Psi} \\ \frac{\partial m^*}{\partial \Psi} \end{pmatrix} = \begin{pmatrix} -\frac{\partial D^a}{\partial \Psi} \\ -d^a(m^*,p^*) - \frac{\Psi\frac{\partial D^a}{\partial \Psi}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}} \end{pmatrix}. \tag{20}$$

Using the Cramer's rule, we then get

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

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

The denominator of both terms is equal to

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

which is positive (ensuring convexity in costs) if, and only if,

$$C''(m) > -\frac{\Psi^2}{(1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a})2\frac{\partial D^a}{\partial p}}. \tag{21}$$

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

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

So that

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

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 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 $\frac{\partial m^*}{\partial \bar{\lambda}}$, we then have

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

So that

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

As the denominator is positive, the effect depends on the numerator, such that $sign(\frac{\partial m^*}{\partial \Psi}) > 0$ if $-2\frac{\partial D^a}{\partial p}d^a(m^*,p^*) > \frac{\Psi \frac{\partial D^a}{\partial \Psi}\frac{\partial D^a}{\partial p}}{1-\frac{\partial D^a}{\partial n}\frac{\partial D^n}{\partial a}}$. Hence, $\frac{\partial m^*}{\partial \Psi} > 0$ if Ψ is low enough while $d^a(m^*,p^*)$ is large enough. Recall that convexity in costs need to be satisfied. Rearranging (21), 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 $\frac{\partial m^*}{\partial \Psi}$ go to zero, then

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

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

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 $C''(m) < \tilde{C}$:

- 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 2

Consider the case in which platforms compete. First, we must assume $c > \underline{c} = \frac{(\overline{\lambda}(2\tau\overline{v}+r\gamma)-r\overline{v}\overline{\phi})^2}{16\overline{v}(\tau\overline{v}+r\gamma)(2\tau\overline{v}+r\gamma)}$. In the second period of the game, by using (11) and (12), we can determine the number of users and

advertisers in platform i as

$$n_i = \frac{1}{2} + \frac{\overline{v}\overline{\phi}(\theta(m_i) - \theta(m_j)) + \gamma(2(p_j - p_i) + \overline{\lambda}(l(m_i) - l(m_j)))}{4(\tau\overline{v} + \gamma \cdot r)} \tag{22}$$

$$a_i = 1 - \frac{\overline{v}(r\overline{\phi}(\theta(m_i) - \theta(m_j)) + 2\tau(r + 2p_i - l(m_i)\overline{\lambda})) + r\gamma(2(r - p_i - p_j) - \overline{\lambda}(l(m_i) + l(m_j))}{4\overline{v}(\tau\overline{v} + \gamma r)} \tag{23}$$

Using these expressions in (15), from the first-order conditions we can solve for the symmetric equilibrium prices and moderation policies $p_i^* = p_j^*$ and $m_i^* = m_j^*$. For ease of exposition, let us define the following critical values of λ.

$$\lambda_{11} \equiv \frac{\overline{\phi} r \overline{v}}{(\tau\overline{v} + \gamma r)}, \quad \lambda_{12} \equiv \frac{4c(4\tau\overline{v} + 3\overline{\gamma}r)}{(2\overline{v} + r)(2\overline{v}\tau + \gamma r)} + \frac{\overline{\phi} r \overline{v}}{2\tau\overline{v} + \gamma r}.$$

The optimal values of m^* and p^* for a symmetric equilibrium. Formally, when $\overline{\lambda} < \lambda_{11}$, $m_i^* = m_j^* = 0$, 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 $\lambda_{11} < \overline{\lambda} < \lambda_{12}$, 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 $\overline{\lambda} \geq \lambda_{12}$, $m_i^* = m_j^* = 1$ and

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

Proof of Proposition 3

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 $-\frac{\partial m^{SH}}{\partial \tau} < 0$.

Proof of Proposition 4

Once again, consider the problem of a platform for a given tax $f < p^*$. Assuming interior solutions, the first-order conditions are:

$$\begin{aligned}\frac{\partial \Pi}{\partial p}(p^*, m^*) =&(p^* - f)\frac{\partial d^a}{\partial p} + d^a(p^*, m^*) = 0,\\ \frac{\partial \Pi}{\partial m}(p^*, m^*) =&(p^* - f)\frac{\partial d^a}{\partial m} - C'(m^*) = 0\end{aligned} \tag{24}$$

for some $(p, m) = (p^*, m^*)$. To understand the effect of f on p^* and m^*, differentiate the above system of equations with respect to f so as to obtain:

$$\begin{aligned}\frac{\partial d^a}{\partial p}(\frac{\partial p^*}{\partial f} - 1) + \frac{\partial d^a}{\partial f} =&0,\\ \frac{\partial d^a}{\partial m}(\frac{\partial p^*}{\partial f} - 1) - C''(m)\frac{\partial m^*}{\partial f} =&0\end{aligned} \tag{25}$$

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, we can define the above system of equations as

$$\begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \frac{\partial d^a}{\partial m} & -C''(m) \end{pmatrix} \begin{pmatrix} \frac{\partial p^*}{\partial f} \\ \frac{\partial m^*}{\partial f} \end{pmatrix} = \begin{pmatrix} \frac{\partial d^a}{\partial p} \\ \frac{\partial d^a}{\partial m} \end{pmatrix} \tag{26}$$

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

$$\frac{\partial p^*}{\partial f} = \frac{\det\begin{pmatrix} \frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \frac{\partial d^a}{\partial m} & -C''(m) \end{pmatrix}}{\det\begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \frac{\partial d^a}{\partial m} & -C''(m) \end{pmatrix}} \tag{27}$$

$$\frac{\partial m^*}{\partial f} = \frac{\det\begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial p} \\ \frac{\partial d^a}{\partial m} & \frac{\partial d^a}{\partial m} \end{pmatrix}}{\det\begin{pmatrix} 2\frac{\partial d^a}{\partial p} & \frac{\partial d^a}{\partial m} \\ \frac{\partial d^a}{\partial m} & -C''(m) \end{pmatrix}}. \tag{28}$$

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

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

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.

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

We show that allowing users to multihome does not change the platform strategies. As a result, prices and content moderation efforts are identical to those arising with singlehoming users and, as such, the

effect of more intense competition on equilibrium outcome does not change. In practice, increasing the fierceness of the competition by reducing transportation costs leads to less moderation in equilibrium. Similarly, we also verify that the effect of intense competition the equilibrium ad price depends on moderation cost and can go both ways.

Appendix B

To get a better picture of the trade-offs present 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. 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 (see equation (4)). 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. Here, we present the assumptions on users and advertisers' preferences that allow us to compute equilibria. 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 dig into the nature of the net gains from brand safety, Ω, as defined by (4). Hence, 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.

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

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:

$$d^a(m,p) = \frac{\theta(m)(r\overline{\phi} - \overline{u}\overline{\lambda}) + 2\overline{u}(r + \overline{v} - p)}{2(\overline{uv} + \gamma r)}. \tag{29}$$

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

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

$$c > \frac{\Psi^2}{8\overline{u}(\overline{uv} + \gamma r\overline{u})} \tag{30}$$

which is equivalent to the one in (21). Using $\theta(m) = 1 - m$, we obtain the following equilibrium outcomes

$$m^* = \begin{cases} 0 \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \leq 0, \\ \frac{\Psi(2\overline{u}(\overline{v}+r)-\Psi)}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2} \text{ if } \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}r \geq 0. \\ 1 \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \geq \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r}. \end{cases} \tag{31}$$

$$p^* = \begin{cases} \frac{\overline{u}(\overline{v}+r)-\Psi}{2\overline{u}} \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \leq 0, \\ \frac{2c(\overline{uv}+\gamma r)(2\overline{u}(\overline{v}+r)-\Psi)}{8c\overline{u}(\overline{uv}+\gamma r)-\Psi^2} \text{ if } \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r} > \overline{\lambda}\overline{u} - \overline{\phi}r \geq 0. \\ \frac{\overline{v}+r}{2} \text{ if } \overline{\lambda}\overline{u} - \overline{\phi}r \geq \frac{4c(\overline{uv}+\gamma r)}{\overline{v}+r}. \end{cases} \tag{32}$$

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 (32).

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 (32).

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 (30) 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

$$c < \frac{\overline{u}(\overline{v}+r)^2}{2(\overline{uv}+\gamma r)} \equiv \tilde{c} \tag{33}$$

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 (33). 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 $\frac{\partial p^*}{\partial \Psi} < 0$.

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.

$$\begin{aligned} \frac{dm}{df} =& \frac{1}{2(\overline{uv} + \gamma r)}\left\{-1 + \frac{\partial p}{\partial f}\right\}, \\ \frac{dp}{df} =& \frac{1}{2} + \frac{1}{4}\left(\overline{\lambda} - \frac{r\overline{\phi}}{\overline{u}}\right)\frac{\partial m}{\partial f}, \end{aligned} \tag{34}$$

which then can be solved as follows:

$$\begin{aligned} \frac{dm}{df} =& -\frac{2\overline{u}f}{8c\overline{u}(\overline{uv} + \gamma r) - \Psi^2}, \\ \frac{dp}{df} =& \frac{4c\overline{u}(\overline{uv} + \gamma r) - \Psi^2}{8c\overline{u}(\overline{uv} + \gamma r) - \Psi^2}, \end{aligned} \tag{35}$$

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.

By substituting (35) into (5), the total effect on advertisers is negative, i.e.,

$$\frac{da}{df} = -\frac{4c\overline{u}}{8c\overline{u}(\overline{uv}+\gamma r) - \Psi^2} < 0. \tag{36}$$