Online Appendix
A Model of Harmful yet Engaging Content on Social Media

By George Beknazar-Yuzbashev, Rafael Jiménez-Durán, and Mateusz Stalinski

A Network Effects

In the main body of the paper, we have assumed away network effects. Here, we relax this assumption and let utility depend on aggregate engagement.[1] Specifically, we assume that the maximum net utility that individuals obtain when they use the platform is:

$$v(h,a,Q) := \max_{q} \; u(q,h,Q) - w(1+a)q. \tag{6}$$

It is easy to show that results 1, 2, 3, and 5 (and the platform's problem under an ad-based business model) remain unchanged in the presence of network effects.

Before proceeding with how Result 4 changes (i.e., how incentives under a subscription-based model change), we note that we have now two relevant aggregate demands (i.e., the aggregate engagement and the aggregate user base). We define the aggregate engagement under a subscription-based business model as:

$$\widetilde{Q}(h,\tau,Q) := \int_{\tau - v(h,0,Q)} q(h,0,Q) f(\theta) d\theta, \tag{7}$$

where $q(h,0,Q)$ solves (6) when $a = 0$ and τ is the fixed fee that the platform charges its users. To solve the fixed-point problem $Q = \widetilde{Q}(h,\tau,Q)$, note that, since engagement is strictly decreasing in the fee, the corresponding inverse demand is $\tau(Q,h)$.

The aggregate user base is now:

$$\widetilde{Q}^s(h,\tau,Q) := \int_{\tau - v(h,0,Q)} f(\theta) d\theta. \tag{8}$$

[1] Another form of network effects could be to let utility depend on the user base. However, we believe that in this context it is more plausible that users care about the aggregate engagement of others. For example, users can prefer a platform where there is more content available ($u_Q > 0$), or their time spent on the platform can be complementary with the time that others spend ($u_{qQ} > 0$).

The platform's problem is now to choose aggregate engagement, to solve:[2]

$$\max_{Q,h\in[0,1]} \tau(Q,h) \times \widetilde{Q}^s(h,\tau(Q,h),Q) - c(Q,h).$$

Result 6. *With network effects, a subscription-based social media platform can find it optimal to display harmful content if aggregate engagement increases with harmful content and if the user base is large and not too price-sensitive.*

Omitting arguments for clarity, the first-order condition with respect to h is now:

$$\frac{\partial \tau}{\partial h}\left(\widetilde{Q}^s + \tau \frac{\partial \widetilde{Q}^s}{\partial \tau}\right) + \tau \frac{\partial \widetilde{Q}^s}{\partial h} = \frac{\partial c}{\partial h}$$

The right-hand side of this expression is positive since individual engagement increases with h and assuming that marginal cost is positive. The sign of the left-hand side is ambiguous. The term $\partial\tau/\partial h$ is positive if aggregate engagement increases with harmful content. The terms inside the parenthesis can be positive if the user base $\widetilde{Q}^s$ is sufficiently large and if it is not too sensitive to the subscription fee (i.e., if $\partial\widetilde{Q}^s/\partial\tau$ is small). The term $\partial\widetilde{Q}^s/\partial h$ is negative because harmful content decreases the user base.

Intuitively, harmful content has two effects on the willingness to pay for a subscription-based platform with network effects. First, it can increase aggregate user engagement (if it is complementary enough with individual engagement), which indirectly pushes up the willingness to pay for the platform. Second, it directly decreases the willingness to pay for the platform because it decreases utility. Therefore, a subscription-based platform will offer harmful content if the first force is strong enough.

B Ex-Post Content Moderation.

In our basic specification, we assume that all the platform does is set the fraction of harmful content that it offers its users, $h \in [0,1]$, filtering the feed to the desired level—which happens unbeknownst to users. However, in addition to quietly tailoring the feed, all platforms engage in public and salient *removal* of hateful content—for example, Twitter places labels that indicate: "This Tweet is no longer available because it violated the Twitter rules." Below we modify our model to include content moderation and show how moderation can coexist with content filtering.

Suppose that the platform moderates (removes) a fraction m of harmful posts. We modify the user utility in two ways. First, they are now exposed to the fraction of unmoderated

[2] As before, we assume that second-order conditions hold.

harmful content, $h(1-m)$. Second, following Jiménez Durán (2022), we assume that individuals might derive direct utility from the removal rate. This could capture, for example, a taste for punishing rule violations (Fehr and Schmidt, 2006).

Thus, the maximum net utility that individuals obtain when they use the platform is now:

$$v(h,a,m) := \max_{q} \; u(q, h(1-m), m) - w(1+a)q, \tag{9}$$

where we assume for simplicity that $\partial u/\partial m$ is positive.

Results 1-5 and their proofs remain unchanged with this modification, but an additional finding is that the incentives to display harmful content can co-exist with the incentives to remove it.

Result 7. *An ad-driven platform can find it optimal to both offer harmful content to users and remove some of this content.*

We re-define the aggregate demand (aggregate user engagement) in this context as:

$$Q(h,a,m) := \int_{-v(h,a,m)} q(h,a,m) f(\theta) d\theta, \tag{10}$$

where $q(h,a,m)$ solves (9).

Let $a(Q,h,m)$ denote the inverse demand curve and $c(Q,h(1-m),m)$ denote the cost function of the platform. The platform's problem is now:

$$\pi^A := \max_{Q,h\in[0,1],m\in[0,1]} \; p(h)a(Q,h,m)Q - c(Q,h(1-m),m).$$

The first-order condition with respect to m is:

$$p(h)\frac{\partial a(Q,h,m)}{\partial m}Q = \frac{\partial c(Q,h,m)}{\partial m} - \frac{\partial c(Q,h,m)}{\partial h}h.$$

The first term of the right-hand side of this equation is positive assuming that the marginal cost of displaying more harmful content to users is positive. Indeed, there is anecdotal evidence that ex-post content moderation is quite resource intensive because it requires, among other inputs, to hire human moderators to review content (Gillespie, 2018). While the second term is also positive, we believe it is plausible that the first term dominates the second one—the marginal costs associated with exposing users to more harmful content are likely small relative to how much it costs to manually review millions of posts every day. For example, the German government has fined Facebook for failing to remove hateful content under its NetDG law, but the fines are not frequent and have amounted up to 2 million Euros (Bundesamt für Justiz, 2019).

Because $a(Q, h, m)$ is the inverse demand curve and since $Q(a, h, m)$ is decreasing in a, we can again say that the platform removes posts when $\partial Q(a, h, m)/\partial m$ is big enough; that is, when the removal of posts increases engagement.

To see the effect on engagement, differentiate Equation (10) with respect to m, use the implicit function theorem and the envelope theorem to get:

$$\frac{\partial Q}{\partial h} = \frac{u_{qm} - u_{qh}}{|u_{qq}|}\frac{Q}{q} + qf(-v(h, a))u_m,$$

where we have omitted some arguments for brevity. In words, the removal of posts increases engagement when it sufficiently increases utility and therefore, the user base (e.g., u_m is high because users derive substantial value from the platform punishing the producers of harmful content) or when, at the margin, moderation is sufficiently complementary with engagement relative to the complementarity between harmful content and engagement (i.e., $u_{qm} - u_{qh}$) is high enough.

References

Bundesamt für Justiz (2019). Federal Office of Justice Issues Fine against Facebook. `https://www.bundesjustizamt.de/DE/Presse/Archiv/2019/20190702_EN.html`. Accessed: 2021-09-30.

Fehr, E. and K. M. Schmidt (2006). The Economics of Fairness, Reciprocity and Altruism: Experimental Evidence and New Theories. *Handbook of the Economics of Giving, Altruism and Reciprocity 1*, 615–691.

Gillespie, T. (2018). *Custodians of the Internet*. Yale University Press.

Jiménez Durán, R. (2022). The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter. *Available at SSRN*.