PNAS

BRIEF REPORT | SOCIAL SCIENCES

OPEN ACCESS

The effectiveness of moderating harmful online content

Philipp J. Schneider[a,1] and Marian-Andrei Rizoiu[b,1,2]

Edited by Margaret Levi, Stanford University, Sanford, CA; received May 2, 2023; accepted June 28, 2023

In 2022, the European Union introduced the Digital Services Act (DSA), a new legislation to report and moderate harmful content from online social networks. Trusted flaggers are mandated to identify harmful content, which platforms must remove within a set delay (currently 24 h). Here, we analyze the likely effectiveness of EU-mandated mechanisms for regulating highly viral online content with short half-lives. We deploy self-exciting point processes to determine the relationship between the regulated moderation delay and the likely harm reduction achieved. We find that harm reduction is achievable for the most harmful content, even for fast-paced platforms such as Twitter. Our method estimates moderation effectiveness for a given platform and provides a rule of thumb for selecting content for investigation and flagging, managing flaggers' workload.

content moderation | harmful content | harm reduction | stochastic modeling

Social media platforms are the new town squares (1)—dematerialized, digital, and unregulated town squares. In 2022, Elon Musk acquired Twitter with the stated goal of preserving free speech for the future. However, alongside free speech, harmful content disseminates and prospers in this unregulated space: mis- and disinformation that spreads faster than its debunking (2), social bots that infiltrate political processes (3), hate speech against women, immigrants, and minorities (4) or viral challenges that put teens' lives at risk. In response, there have been calls for the governments to intervene and regulate. As the first move of its kind, the European Council introduced the Digital Services Act (DSA) and the Digital Markets Act (DMA) (5), EU legislation aimed at projecting the regulations of our offline world onto the digital one. It implements notice and action mechanisms (cf. Art. 16) to report harmful online content. Furthermore, the regulation introduces a process for appointing *trusted flaggers*, subject matter experts in detecting harmful content (cf. Art. 22). Once such content is flagged, platforms must promptly remove the content. However, online content is notorious for its "virality"—it spreads at high speeds and has short lifespans. Therefore, we ask about the effectiveness of this new legislation: how to quantify the likely harm caused by harmful content and how to determine the response time for effective mitigation?

In this work, we leverage state-of-the-art information spread modeling to assess the effectiveness of the DSA regulation and the EU code of conduct for countering harmful online speech. Fig. 1 conceptualizes an online discussion, where each post (● or ⬡) draws more people into the discussion and generates more posts, referred to as offspring. This phenomenon of content spreading is known as the *self-exciting property*. A harmful post (⬢) will therefore generate potentially other harmful posts (⬢ and ⬢) with a decreasing intensity, shown by the red dashed line on the *Bottom* panel of Fig. 1. How would the new EU legislation potentially stop the propagation of the harm? The core concept is to limit harmful posts' reach and the offspring generation. We denote the number of harmful, direct offspring as the potential harm—denoted as n^* and comparable in meaning to R_0, the basic reproduction number of infectious diseases (6). Content moderation is achieved by removing the harmful post (⬢) at time Δ after posting and thus stemming offspring generation after this time (⬢). In addition, we assume that any harmful direct offspring generated before Δ (⬢) are also moderated; their number defines the actual harm—labeled as n^*_Δ. The harm reduction χ is the percentage of all harmful offspring avoided, both direct and indirect—i.e., offspring of the offspring generated via the recurrent branching process. The effect of the policy heavily depends on the speed at which the discussions unfold on social networks. We quantify this using the *content half-life*, defined as the time required to generate half of the direct offspring. A recent (as of 2023) empirical investigation (7) determined the half-life of social media posts on different platforms: Twitter (24 min), Facebook (105 min), Instagram (20 h), LinkedIn (24 h), YouTube (8.8 d), and Pinterest (3.75 mo). A lower half-life means that most harm happens right after the content is posted, and content moderation needs to be performed quickly to be effective.

Author affiliations: [a]College of Management of Technology, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland; and [b]Faculty of Engineering and Information Technology, The University of Technology Sydney, Ultimo NSW 2007, Australia

Author contributions: P.J.S. and M.-A.R. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

The authors declare no competing interest.

[1] P.J.S. and M.-A.R. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: Marian-Andrei.Rizoiu@uts.edu.au.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2307360120/-/DCSupplemental.

Published August 14, 2023.

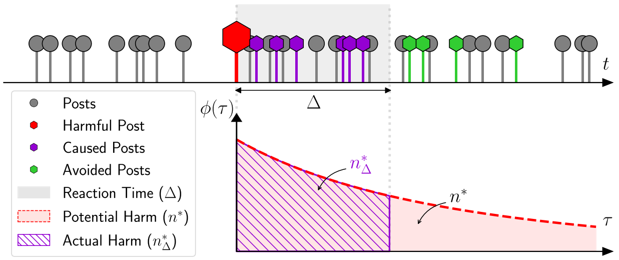
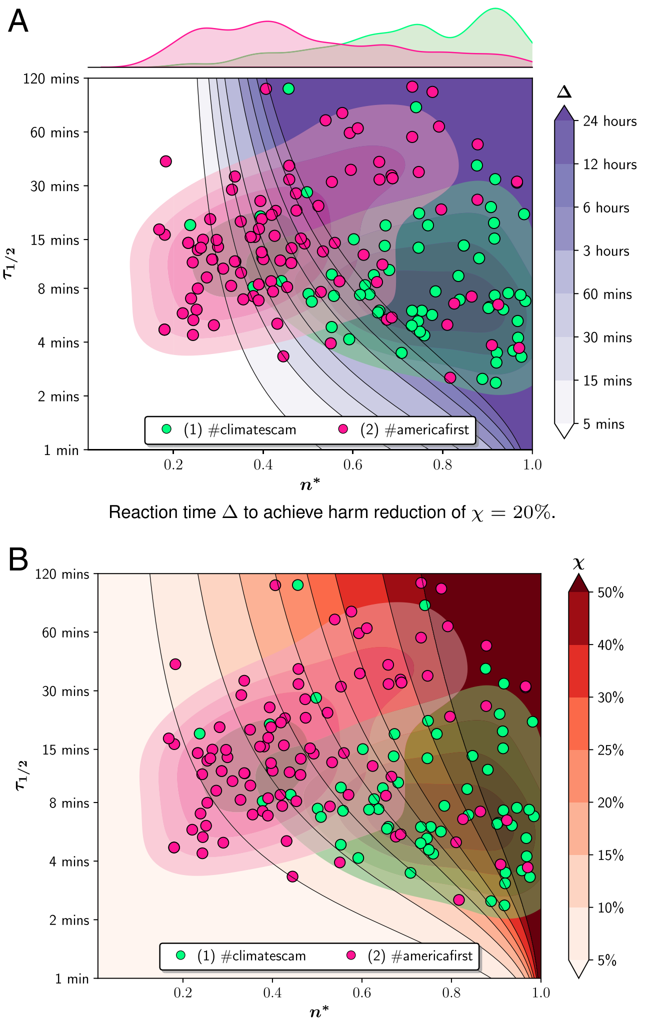


Fig. 1. Social media dynamics as self-exciting point process. Social media posts (●) include a percentage of posts considered harmful (⬡). One harmful post (⬢) likely generates n^* other harmful content at the rate $\phi(\tau)$. When the post is removed at time Δ, the likely caused harm is limited to n^*_Δ (⬢) and further harm is avoided (⬢). The harm reduction χ is the percentage of all harmful content generated directly or indirectly via self-excitation avoided via moderation.

Results

We analyze the likely effectiveness of the EU-regulated moderation as the interplay between the reaction time Δ and the harm reduction χ (Table 1). This interplay is modulated by the online content's potential harm n^* and half-life $\tau_{1/2}$. The colormap in Fig. 2*A* explores the question what is the maximum reaction time required to achieve a given harm reduction (here $\chi = 20\%$)? On the colormap, we project real-world discussions around two potentially problematic topics: climate change denial and nationalistic political views (*Materials and Methods*). We obtain the positions of the data points by estimating the process dynamics over rolling time windows. The mean content half-life for the two topics is 25.8 min. The centroids of the discussion dynamics are at ($\tilde{n}^* = 0.75$, $\tilde{\tau}_{1/2} = 7.48$ min)* for #climatescam and ($\tilde{n}^* = 0.44$, $\tilde{\tau}_{1/2} = 13.97$ min) for #americafirst. This indicates that due to the significantly higher virality for #climatescam, more time ($\Delta > 24$ h) is available for content moderation compared to the nationalism topic ($\Delta = 2.22$ h). Fig. 2*B* explores the question what is the expected harm reduction when content moderation is performed within 24 h? (as currently stipulated by the EU regulation) The harm reduction at the centroids is $\chi = 29.18\%$ and $\chi = 13.29\%$ for #climatescam and #americafirst, respectively.



Reaction time Δ to achieve harm reduction of $\chi = 20\%$.

Harm reduction χ when content is removed within $\Delta = 24$ hours.

Fig. 2. Visual representation of the dependency of the reaction time Δ (*A*) and harm reduction χ (*B*) with the potential harm n^* (*x*-axis) and content half-live $\tau_{1/2}$ (*y*-axis). Both Δ and χ increase for longer half-lives and higher virality. Real-world potentially problematic content exhibits widely highly variable dynamics. For #climatescam, we can achieve harm reduction of [15%, 50%] for $\Delta = 24$ h.

Discussion

Our work introduces two measures for the effectiveness of DSA moderation: the potential harm n^* and the content half-life $\tau_{1/2}$. For content with known n^* and $\tau_{1/2}$, we can determine the relation between the reaction time Δ and the harm reduction χ.

We make three observations. First, Fig. 2*A* shows the reaction time Δ increases with both half-life $\tau_{1/2}$ and the potential harm n^*. While the former is intuitive, the latter is significant as it indicates that the DSA-legislated moderation can be effective even on Twitter, the platform with the shortest half-life. For example, most of the discussions on #climatescam have high reaction times ($\Delta > 24$ h). Second, Fig. 2*B* shows that the harm reduction χ increases with the potential harm n^*. This somewhat counterintuitive result arises from the nonlinear interactions between χ, n^*, and $\tau_{1/2}$ (*Materials and Methods*). It is significant as it indicates that DSA moderation can effectively stop the most harmful content. Third, despite a significant difference in the distribution of potential harm χ for the two topics (see the density marginals on top of Fig. 2*A*), we see that the most harmful content can emerge in both topics. Therefore, we cannot select a single topic for moderation. Our approach can be used as a strategy to direct the manual flagging efforts toward the most harmful content ($n^* > 0.8$) that can be effectively moderated ($\chi > 50\%$). This would increase the overall effectiveness of the moderation and the flagger's workload.

The major social media platforms employ large content moderation teams, estimated to 15,000 (Facebook), 10,000 (YouTube), and 1,500 (Twitter), each moderator addressing between 600 and 800 claims daily (8). This indicates platforms

Table 1. Variables of interest for modeling content removal

Parameter	Interpretation
$\phi(\tau)$	The rate at which content generates reactions on social media.
n^*	Potential harm—The number of additional harmful posts a content generates directly.
Δ	Reaction time—Mandated time to remove flagged harmful content on social media platforms.
n^*_Δ	Actual harm—The number of direct harmful reactions a content generates prior to moderation at time Δ.
$\tau_{1/2}$	Content half-life—Time until a content generated half the direct reactions.
χ	Harm reduction—Percentage of direct and indirect harmful offspring avoided by content moderation.

*Let $\tilde{x}$ denote the median of x.

have sufficient human resources for the new legislation, despite the concerns about moderators' suitability based on social context awareness, native language, and moderation guidelines. With the DSA, the European Union seeks to make this process uniform across platforms, transparent, and regulated. The keys to its success seem to be appointing trusted flaggers, developing an effective tool for reporting harmful content across platforms, and correctly timing the reaction time for moderation. This paper provides a framework for policymakers to draft mechanisms for content moderation by indicating where to focus human fact-checking efforts and how quickly to react.

Materials and Methods

Dataset. To showcase the applicability of our methodology, we compiled a Twitter dataset comprising tweets emitted between 1 July and 31 December 2022, relating to two topics often linked to harmful content, as identified by prior literature and news media. The two topics were collected using the hashtags 1) #climatescam (479,051 posts)–controversial opinions regarding climate change (9)–and 2) #americafirst or #americansfirst (278,899 posts)–debates over key US political topics such as immigration and foreign policies (10).

Method. We estimate the relationship between moderation efforts and content-sharing dynamics using a well-established point process framework (see *SI Appendix* for a review).

Model. Events in social media are commonly modeled using self-exciting point processes, also referred to as Hawkes processes (11). A conditional intensity function $\lambda(t|\mathcal{H}_t)$ defined as

$$\lambda(t|\mathcal{H}_t) = \mu + \sum_{i:t_i<t} \phi(t - t_i; m_i), \quad t \geq 0, \qquad [1]$$

captures how historical events cause the generation of new events. In this work, events are social media postings–both harmful and nonharmful, which are assumed to share the diffusion dynamics within a topic. The background rate μ explains exogenous effects, such as users spontaneously posting new information; the kernel $\phi(\cdot)$ controls the endogenous dynamics of how users reshare and retweet the content they see. The history $\mathcal{H}_t$ stores the event times t_i when an action was performed and the follower count of the user m_i as $\{(t_i, m_i)\}_{i=1}^{N(t)}$, where $N(t)$ is the total number of events. We model contagion using a power-law function

$$\phi(\tau; m) = \kappa\, m^{\beta} (\tau + c)^{-(1+\gamma)}, \quad \tau \geq 0, \qquad [2]$$

as the events' influence is long-lasting and heavy-tailed. The memory parameter γ captures the speed at which the content is forgotten. A higher γ indicates that the content's importance is fast decreasing. The scalar κ describes the quality of the post. The shift parameter c captures the waiting time before users interact with the post. The exponent β warps the user follower count m. The follower count is known to follow power-law distribution $P(m) = (\alpha - 1)\, m^{-\alpha}$ with parameter $\alpha = 2.016$ (12). By integrating Eq. **2** over time and social influence, we obtain the potential harm as

$$n^* = \int_1^{\infty} \int_0^{\infty} P(m)\, \phi(\tau)\, d\tau\, dm = \kappa \frac{\alpha - 1}{\alpha - \beta - 1} \frac{1}{\gamma\, c^{\gamma}}, \qquad [3]$$

where $\alpha < \beta - 1$ and $\gamma > 0$. Social media content is described by a subcritical regime, meaning $n^* < 1$, as retweet cascades vanish in the long term.

Content half-life. The time required by a post to generate 50% of its expected direct offspring, $\tau_{1/2}$, is determined as

$$\int_0^{\tau_{1/2}} \phi(z; m)\, dz = \frac{1}{2} \int_0^{\infty} \phi(z; m)\, dz. \qquad [4]$$

We substitute Eq. **2** in Eq. **4** and obtain $\tau_{1/2} = c\,(2^{\frac{1}{\gamma}} - 1)$.

Content removal. Here, we outline the connection between the moderation time Δ and the harm reduction χ. When a post is moderated at deletion time Δ, its direct offspring generation rate drops to zero (see the *Bottom* panel of Fig. 1). The *moderated kernel* $\phi_{\Delta}(\tau; m)$ is

$$\phi_{\Delta}(\tau; m) = \begin{cases} \kappa\, m^{\beta} (\tau + c)^{-(1+\gamma)}, & 0 < \tau \leq \Delta, \\ 0, & \tau > \Delta. \end{cases} \qquad [5]$$

Hence, we formally define the *actual harm* as the expected number of direct offspring that a harmful post generates prior to its moderation at time Δ. We compute the actual harm similarly to Eq. **3**, by replacing $\phi(\tau; m)$ with $\phi_{\Delta}(\tau; m)$ (cf. Eq. **5**):

$$n^*_{\Delta} = \frac{\kappa}{\gamma} \frac{\alpha - 1}{\alpha - \beta - 1} \left(\frac{1}{c^{\gamma}} - \frac{1}{(c + \Delta)^{\gamma}} \right). \qquad [6]$$

We compute harm reduction as the percentage of avoided harmful offspring through the branching process. Assuming only one post in the expected event stream is harmful; this post will generate many offspring. Not all are harmful; however, we assume that all generated harmful posts are moderated in their turn through the DSA mechanism at time Δ. This, the resulting harm reduction χ is

$$\chi = 1 - \frac{\frac{1}{1 - n^*_{\Delta}}}{\frac{1}{1 - n^*}} = \frac{n^* - n^*_{\Delta}}{1 - n^*_{\Delta}}. \qquad [7]$$

Through the substitution of Eqs. **3** and **6** into Eq. **7**, we can reframe the equation to represent Δ as a function of χ,

$$\Delta = \max \left\{ 0, \left(\frac{1}{n^*\, c^{\gamma}} \frac{\chi(1 - n^*)}{1 - \chi} \right)^{-\frac{1}{\gamma}} - c \right\}. \qquad [8]$$

Finally, we compute the colormaps in Fig. 2 using Eqs. **7** and **8**.

Statistical inference. Given observed (real-world) data, the parameters of the Hawkes process specified in Eq. **1**–i.e., κ, β, and γ–are identified via MLE (*SI Appendix*). The exogenous effect μ is estimated from empirical observations, and we set the shift parameter c to thirty seconds. In Fig. 2, the resulting estimates are depicted as point clouds with kernel density estimate plots.

Data, Materials, and Software Availability. Code and dehydrated dataset, compliant with Twitter Terms of Service, is available at: https://github.com/behavioral-ds/harmful-content-moderation/ (13).

1. J. Burgess, N. K. Baym, *Twitter: A Biography* (NYU Press, New York, NY, 2020).
2. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
3. M.-A. Rizoiu *et al.*, "#DebateNight: The role and influence of socialbots on Twitter during the 1st 2016 U.S. presidential debate" in *Proceedings of the Twelfth International AAAI Conference on Web and Social Media* (AAAI, Palo Alto, CA, 2018), pp. 300–309.
4. T. Davidson, D. Warmsley, M. Macy, I. Weber, "Automated hate speech detection and the problem of offensive language" in *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (AAAI, Palo Alto, CA, 2017), pp. 512–515.
5. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act). *OJ L* **277**, 1–102 (2022).
6. M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, L. Xie, "SIR-Hawkes: Linking epidemic models and Hawkes processes to model diffusions in finite populations" in *Proceedings of the 2018 World Wide Web Conference* (International World Wide Web Conferences Steering Committee, Geneva, Switzerland, 2018), pp. 419–428.
7. S. M. Graffius, Lifespan (half-life) of social media posts: Update for 2023 (2023). https://dx.doi.org/10.13140/RG.2.2.19783.98722 (Accessed 31 July 2023).
8. P. M. Barrett, *Who Moderates the Social Media Giants?* (NYU Stern Center for Business & Human Rights, 2020).
9. O. Milman, #ClimateScam: Denialism claims flooding Twitter have scientists worried. *Guardian* (2022). https://bit.ly/guardian-climatescam-twitter (Accessed 31 July 2023).
10. D. L. Linvill, P. L. Warren, Troll factories: Manufacturing specialized disinformation on Twitter. *Polit. Commun.* **37**, 447–467 (2020).
11. A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90 (1971).
12. S. Mishra, M.-A. Rizoiu, L. Xie, "Feature driven and point process approaches for popularity prediction" in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management* (ACM, New York, NY, 2016), pp. 1069–1078.
13. P. J. Schneider, M.-A. Rizoiu, Materials repository for "The effectiveness of moderating harmfulonline content". Github. https://github.com/behavioral-ds/harmful-content-moderation. Deposited 13 July 2023.