Thèse n° 10 387

# Content Moderation in Online Platforms

Présentée le 28 juin 2024

Faculté informatique et communications
Laboratoire de science des données
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

**Manoel HORTA RIBEIRO**

Acceptée sur proposition du jury

Prof. K. Aberer, président du jury
Prof. R. West, directeur de thèse
Dr J. Hofman, rapporteur
Prof. A. Narayanan, rapporteur
Prof. D. Gatica-Perez, rapporteur

École
polytechnique
fédérale
de Lausanne

2024

To Jessica.

# Acknowledgements

I was lucky to have been advised by Professor Robert West, better known as Bob, but whom I've decided to call Beto since he started calling me Manolo. You are the researcher you want to be when "I grow up." You are playful but compassionate; ambitious with your research but selfless with your students; creative but attentive to detail. Thank you for your friendship, the insightful discussions, and all the care and support. I was also lucky to have a thoughtful, kind committee: Arvind, Daniel, Jake, Karl—thanks for the feedback and the great discussion!

My labmates have been a source of great joy in the past years. Kristina and Tiziano welcomed me to the lab as an intern and even took me to the Zoo when I was sad. Maxime taught me that French people are "built differently" regarding how much cheese they can eat. Many fond memories of my time in Switzerland involve you three: from snow-shoeing to the trip to the south of France (we will arrive in three hours!). Collaborating with Kristina was also one of the highlights of my Ph.D., thanks for being my "bullshit filter!"

But I feel like I have special moments with all of you, for which I'm grateful. Valentin and Marija conspired with me to fill Bob's room with 400 (?) balloons; Akhil conspired with me to run ADA (a much more challenging task); Saibo and Chris were my 'push-up' buddies; Masani 'pushed' me to walk a bit further to go to the better restaurant; Martin was my Seattle buddy, and entirely responsible for us getting lost at Mount Rainier; Niklas showed me how cool Zürich is; Candice taught me not eat my access card and not to lose my keys. My office mates Andreas (way tidier than me), Bhargav (less tidy than me), and Tim (tidier than me) entertained me daily over the years.

A big shout-out also goes to 'my boys.' Venia, who bullied me into writing a paper; Thorsten, who taught me so many exquisite German words (none of which I remember); Bhargav, whose Halloween costumes only slightly outdid mine. Tim, who increased my expenses with lunch by 5 francs a week and Peppe, who is yet to return my jacket as I write this. Thank you for all the fond memories, and I'm sorry for eating fries at twice the speed you all do. You're fantastic researchers, and I was fortunate to have collaborated with you. (I hope more is yet to come.)

Part of this thesis also came from master's students who collaborated with me within DLAB. Robin, Barbara, Marie(s), Deniz, Léopaul, Francis, Julia, Breno, Tiancheng, Jordi, Mihaela, Mateo, Maciej, Venia, Francesco, Paula, and Guosheng. I was lucky to help advise you.

Outside DLAB, my doctorate was filled with collaborations across industry and academia. My years were brightened by (often remote) interactions with Homa Hosseinmardi, Duncan Watts, Siddharth Suri, David Rothschild, Phillip Cozzolino, Teny Shapiro, Sanjay Kairam, Tiancheng Hu, Giuseppe Russo, Justin Cheng, Yiqing Hua, Mor Naaman, Shagun Jhaver, Savvas Zanettou, Jeremy "Jimmy" Blackburn, Emiliano DeCristofaro, Gianluca Stringhini, Denis Helic, Thorsten Ruprechter, Markus Strohmaier, Florian Lemmerich, Wagner Meira, and Virgilio Almeida. I rewrote this section a couple of times—I always end up making it too long by trying to write a quirky remark or memory about each of you. Eating too much 'churrasco' with Sid; playing Among Us with iDRAMA folks, long Zoom calls with Homa, Shagun, and Yiqing; and Sanjay neighbors' string quartet. Thanks for all the fish. A special thanks go to Sid and Duncan, who took the time to mentor me as I prepared for what's next.

Some people I'm grateful towards don't quite fit in other paragraphs. Nastia was my first Ph.D. friend, my housemate, and a fierce adversary in our 'who-can-eat-the-most-honey-melon' contest. Thanks for introducing me to Pietro, Volodia, and Slava. Victoire and Rebecca were good friends and travel buddies around Annency and Denmark, respectively. Chacha, Eleni, and Andalib made my visit to Seattle more fun. Gisa gave me a sneak peek of what it would be like to have a teenage sister and taught me how to enjoy the ocean properly. Mako was an impromptu mentor and hung out with me from east to west (coast).

During these five years, my friendships in Brazil held firm and were a crucial part of my late-year escapades from the cold. Meusbamigos has transcended TeamSpeak into barbecues and New Year's Eve retreats. Amaral, Carol, Muzinho, Luiza, Zezola, Sarinha, Brunolol, Paula, Vitor, Amanda, João, José, Paulinha, Marquinhos, Flavinha, Gabriel "N," Luli. Also, while I did not tattoo a giant Goose on my thighs, a shout-out goes to 'Gansões' Caio, Hermano, Laura, Faria, César, Deko, Marcelo, and Castanheira. Thank you all for being so wholesome.

My family has been a source of joy and infinite support, to the extent that all my accomplishments seem like a deterministic consequence of their love. My mom prepared me for life's adventures by reading The Odyssey and showing me not to cower in adversity. I owe her my curiosity and resilience. My dad has taught me not to get ahead of myself ('não se meta a gato mestre') and to strive for excellence in all I do. I owe him my ambition and integrity. Antônio has always been my greatest friend and has always uplifted me; if I went far, it was only because I followed in your footsteps.

My wife Jessica embarked on the journey of coming to Switzerland and played all the roles of a ship's crew: captain, navigator, pilot, and quartermaster. She has also made the last five years the happiest 1/6th of my life. Thank you for all the care, support, and love; and for being alongside me during the good and bad times, the trips to the Mediterranean, and the Pandemic. If I were to express how you helped me, this whole acknowledgment would turn into a love letter. But please know that building a life together gives meaning to everything.

*Lausanne, 13 Mars 2023* M. H. R.

# Abstract

A critical role of online platforms like Facebook, Wikipedia, YouTube, Amazon, Doordash, and Tinder is to moderate content. Interventions like banning users or deleting comments are carried out thousands of times daily and can potentially improve our online spaces. However, researchers, governments, and even platforms are ill-informed about the impact of content moderation.

In this thesis, we study the causal effect of various content moderation practices using observational studies and a large-scale field experiment. With online traces that range from users' daily activity on fringe websites to comments on online communities, we develop and apply observational and experimental study designs to understand how content moderation shapes subsequent user behavior and our online spaces.

The thesis is divided into two parts, each containing three studies. In the first part, we consider interventions targeting individuals or collectives in social media, conducting studies that assess the effect of banning influencers, online communities, and even entire online platforms. In the second part, we consider interventions targeting the creation of content in social media at different points in time. We study the effect of automatically deleting content after it has been posted, requiring content to be manually approved before posting, and helping users create content that is not rule-breaking.

Altogether, the scientific findings and methods presented advance our understanding of how content moderation shapes user behavior and online platforms and can inform the design of content moderation interventions, systems, and policies.

Key words: data science, human behavior, content moderation, causal inference, computational social science.

# Résumé

Un aspect crucial des plateformes en ligne telles que Facebook, YouTube, Amazon, Doordash et Tinder est leur rôle dans la modération du contenu. Des interventions telles que la suppression de commentaires ou l'interdiction d'utilisateurs sont effectuées à de multiples reprises chaque jour et ont le potentiel d'améliorer nos espaces en ligne. Cependant, les chercheurs, les gouvernements et même les plateformes sont souvent mal informés sur l'impact réel de cette modération.

Dans cette thèse, nous étudions l'effet causal de diverses pratiques de modération du contenu à travers des études observationnelles et une expérience à grande échelle. En utilisant des données en ligne allant des interactions quotidiennes sur des sites web marginaux aux discussions au sein des communautés en ligne, nous développons et appliquons des modèles d'analyses observationnelles et expérimentales pour mieux comprendre comment la modération du contenu influence le comportement des utilisateurs ainsi que nos espaces en ligne.

La thèse est organisée en deux parties, chacune comprenant trois études. Dans la première partie, nous examinons les interventions visant des individus ou des groupes spécifiques sur les réseaux sociaux, en menant des études qui évaluent l'impact de l'interdiction d'influenceurs, de communautés en ligne voire de plateformes entières. Dans la seconde partie, nous considérons les interventions ciblant la création de contenu sur les réseaux sociaux à différents stades. Nous étudions l'effet de la suppression automatique du contenu après sa publication, l'exigence d'une validation manuelle avant publication, ainsi que l'aide apportée aux utilisateurs pour créer un contenu respectant les règles établies.

Dans l'ensemble, les découvertes et méthodes scientifiques présentées font progresser notre compréhension de la façon dont la modération du contenu influence le comportement des utilisateurs et des plateformes en ligne et peuvent guider la conception d'interventions, de systèmes et de politiques de modération de contenu.

Mots clés : science des données, comportement humain, modération de contenu, inférence causale, sciences sociales computationnelles.

# Contents

# List of Figures

# List of Tables

# Introduction and Background Part I

# 1 Introduction

> The experimenting society will be one which will vigorously try out proposed solutions to recurrent problems, which will make hard-headed and multidimensional evaluations of the outcomes, and which will move on to try other alternatives when evaluation shows one reform to have been ineffective or harmful. We do not have such a society today.
>
> Donald Campbell

## 1.1 Motivation

Online platforms like Facebook, Wikipedia, YouTube, Amazon, Uber, DoorDash, AirBnb, and Tinder have changed the world and become embroidered into the social fabric [182, 3, 30]. It is hard to imagine how our lives would be in their absence: our economies, our relationships, and how we acquire knowledge have become deeply connected to these online platforms. The United Nations Conference on Trade and Development estimated that the global value of e-commerce sales reached almost 26 trillion dollars in 2018 [242]. Pew Research estimated that around 10% of partnered US adults met their match online dating in 2023 [30]. Wikipedia received over 7 billion monthly visits in 2023, satisfying users with the most diverse information needs [241].[I]

Thus, it is perhaps not surprising that online platforms are also strongly connected to some of the most significant societal challenges of the 21st century. E-commerce platforms are responsible for a sizeable chunk of greenhouse gas emissions [24], especially when customers return products [263]. Gig work platforms like Uber and DoorDash ignited discussions about

[I] https://go.epfl.ch/wikipedia_stats

workers' rights and precarious employment [22, 213]. Radicalization and terrorism have become an online-first phenomenon. Mainstream social media platforms like YouTube saw an influx of radical content that snowballed in popularity in the late 2010s [97], and fringe platforms like Gab and Parler were tightly associated with terrorist attacks and anti-democratic protest, respectively [271, 158].

But online platforms are not "immovable rocks" that we should accept as they are; they are sociotechnical systems where design choices, policies, and algorithms steer human behavior [76, 140, 85]. Their critical enterprise is to *curate content* [77, 76, 171]. Users on these platforms upload images, list products to sell, and make profiles. At the same time, platforms curate this content and serve it to other users on the platform. And there are some critical ways in which they do so: they decide what to recommend to users [78, 102], how to monetize content on the platform [104, 209], and, most important to this thesis, *they determine what is allowed on the platform* [76, 85].

Content curation practices have captured the imagination of journalists, politicians, and the general public. For example, tech CEOs testified in Senate Hearings in 2018, 2020, 2021, and 2024 [258, 177, 176, 175]; they were often asked to discuss content curation practices like recommender systems or content moderation; In a 2018 poll, 65% of self-described US conservatives thought social media platforms were censoring conservative ideas [12]. In that context, research informing content curation practices is *actionable*; it can propose concrete ways of changing online platforms and inform stakeholders. For example, deplatforming or banning individuals or collectives from our online ecosystem has been widely debated, as the intervention treads a thin line between preventing harm and censoring speech [21, 49, 48, 47]. How to weigh the benefits and harms of these practices? How do we assess whether they are effective?

Enter research on content curation. If we can understand the consequences of different content curation practices, we can improve them (or stop using them altogether). We can even devise and evaluate new ways of curating content. Recommender systems can be tuned, and moderation and monetization policies may be adjusted and tweaked.

Despite this promise, however, research on content curation practices has arguably failed to drive their development and adjustment. Compared to other (polarizing) topics like healthcare or labor, content curation practices in social media are disproportionally driven by anecdotal or observational evidence that describes problems, but not solutions [240, 97, 118]. This is (at least partly) because researching content curation practices is challenging. Content curation practices are opaque and carried by private companies at their discretion [19, 82, 212]; Researchers often lack access to data or the necessary experimentation infrastructure [186, 277, 136, 68]; Online platforms are highly dynamic, raising concerns about the temporal validity of findings [172]; And, in some cases, disentangling the effect of content curation practices is methodologically challenging, *e.g.*, see the literature on the effects of the YouTube algorithm on video consumption [97, 103, 102].

Studying content curation practices is challenging but can yield significant payoffs: given how widely used online platforms are [182, 3, 30], even marginal improvements to online environments are meaningful; and given that there's a wide appetite for regulating online platforms [134, 21], research can guide policy away from guesswork.

Motivated by these challenges and payoffs, this thesis proposes and assesses content moderation practices (a subset of content curation practices) as a way of *improving* online platforms, maximizing their benefits and minimizing their harms. We demonstrate that, although challenging, research on content curation practices is feasible and capable of providing meaningful insights.

We use the term content moderation to refer to what Grimmelmann [85] broadly defines as: "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[II] Some content curation practices are arguably *not* content moderation, *e.g.*, sharing ad revenue with users. Still, this line is blurry around the edges, *e.g.*, kicking out specific users from this revenue-sharing scheme arguably is [104]. Nonetheless, content moderation is a common practice among online platforms (even though platforms sometimes pretend to be only intermediaries) and is central to the experience they provide [75].

Our focus on this subset of curation practices is, above all, pragmatic. Content moderation involves well-defined interventions (*e.g.*, banning users and deleting posts), which are often easier to observe as external researchers. When Amazon AWS bans Parler or Twitter bans Donald Trump, the precise date of the ban is public knowledge and easily retrievable by researchers [99]. On the contrary, YouTube may change its recommender system over a year, slowly rolling out the changes to subsets of users, potentially without announcing it. In this case, it is hard for researchers to know which users are exposed to the new algorithm (or even which fraction of users), making analysis more challenging.

To ensure the research can translate into actionable insights to improve online platforms, we focus on obtaining *causal results.* Improving online platforms entails asking fundamentally causal questions. How would users behave if we did not delete their comments? Would user behavior improve if we introduced a new intervention? When we ban fringe platforms or communities, do users reduce the consumption of extreme content? These questions require us to go beyond mere observation and ask *what if.* Fortunately, we are in the wake of a causal revolution [188], and a growing range of methods allows us to draw causal conclusions. Methods used here come mainly from the statistics and econometrics literature: difference-in-differences, propensity score matching, and regression discontinuity designs approximate an experiment with observational data, enabling us to obtain causal conclusions in the absence of experimental data [93, 216, 110, 218].

[II] It is worth noting that Grimmelman uses the concept of community in a broad sense: "A community can be as small as the handful of people on a private mailing list or as large as the Internet itself. Communities can overlap, as anyone on both Twitter and Facebook knows. Communities can also nest: the comments section at Instapundit is a meaningful community, and so is the conservative blogosphere."

But what data? In this thesis, we consider comprehensive *digital traces* that capture human behavior generated indirectly and directly by online platforms.[III] Comments on Facebook; news crawled from Reddit; time-series of Wikipedia page views. Although eclectic, this data shares some commonalities, as discussed by Salganik [228]. It is *big* and *always-on*, allowing the study of unexpected events, the detection of small effects, and the study of the heterogeneity of the effect. This data is also often *inacessible*, which required us to find ways to get the data needed to answer questions of interest, or to figure out what meaningful questions could be answered with the data available. This journey entailed industry partnerships with Facebook and Reddit, applying a special algorithm to improve Google Trends data, and creating custom crawlers to extract data from incels.is and thedonald.win.

## 1.2 Contributions and thesis overview

This thesis is organized into eight chapters roughly divided into two parts. Part I is composed of the introduction and background. Part II examines the effects of moderation interventions targeting individuals or collectives on social media. Part III examines interventions targeting content creation in online platforms. Part IV presents a final discussion and conclusion.

Altogether, this thesis advances toward maximizing the benefits and minimizing the harms of online platforms by gathering causal evidence on the effect of content moderation practices on human behavior. Individually, the chapters of this thesis contribute to social computing and computational social science, as summarized in the following chapter-by-chapter outline.

### 1.2.1 Banning communities (Chapter 3)

*Originally published at The 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing [100]*

When toxic online communities on mainstream platforms face moderation measures, such as bans, they may migrate to other platforms with more lax policies or set up their dedicated websites. Previous work suggests that community-level moderation effectively mitigates the harm caused by the moderated communities within mainstream platforms. It is, however, unclear whether these results also hold when considering the broader Web ecosystem. Do toxic communities continue to grow in terms of their user base and activity on the new platforms? Do their members become more toxic and ideologically radicalized?

We conduct a large-scale observational study of how problematic online communities progress following community-level moderation measures. We analyze over 6 million posts made by more than 138 thousand from r/The_Donald and r/Incels, two communities that were banned from Reddit and the standalone websites these communities subsequently migrated to. We

[III] To get a sense of how comprehensive, see privacy notes of Google (https://go.epfl.ch/edit/alias/privacy-google), Amazon (https://go.epfl.ch/privacy-amazon), and AirBnB (https://go.epfl.ch/privacy-airbnb).

extract activity-related signals, such as the number of posts, active users, and newcomers, as well as content-related signals, such as algorithmically derived “toxicity scores,” that aim to identify behaviors indicative of user radicalization. Employing quasi-experimental setups, including matching and regression discontinuity analysis, we study these signals from a community-level perspective, analyzing how daily activity and overall content changed, and from a user-level perspective, examining how the behavior of individual users changed following platform migrations.

Our key findings are two-fold. First, analyzing activity levels and the inflow of newcomers to the communities, we find that the moderation measures significantly reduced the overall number of active users, newcomers, and posts in the new communities compared to the original ones. However, individually, users post more often on alternative platforms. A closer look at the users we matched before vs. after the migration suggests that this increase in relative activity is due more to self-selection rather than behavior change. Users who migrated were more active in the original platform, and their activity dropped on a user level. Second, analyzing changes in the content being posted in the communities following the migration, we find evidence that users in the r/The_Donald community became more toxic, negative, and hostile when talking about “objects of fixation” (*e.g.*, democrats, leftists). Changes in the usage of third-person plural (*e.g.*, “they”) and first-person plural (*e.g.*, “we”) pronouns also indicate an increase in ingroup identification and othering language. For the r/Incel community, we find that changes tend to be statistically non-significant.

Our analysis suggests that community-level moderation measures decrease the capacity of toxic communities to retain their activity levels and attract new members, but this may come at the expense of making these communities more toxic and ideologically radical. Therefore, as platforms moderate, they should consider their impact not only on their websites and services but also on the Web as a whole.

### 1.2.2 Banning platforms (Chapter 4)

*Originally published at PNAS Nexus [99]*

Online platforms have banned influencers, communities, and even entire websites to reduce content deemed harmful. Deplatformed users often migrate to alternative platforms, which raises concerns about the effectiveness of deplatforming. Here, we study the deplatforming of Parler, a fringe social media platform, between 2021 January 11 and 2021 February 25, in the aftermath of the US Capitol riot.

Using two large panels that capture longitudinal user-level activity across mainstream and fringe social media ($N$=112,705), we analyze the overall trends in fringe platform consumption. Further, using a matched sample of users that consumed fringe social media platforms, we study the effect of deplatforming ($N$=996) on subsequent fringe social media use with a difference-in-differences analysis.

Our key findings are two-fold. First, we find that other fringe social media, such as Gab and Rumble, prospered after Parler's was deplatformed, as the overall activity on fringe social media increased while Parler was offline. Second, we find that deplatforming increased the probability of daily activity across other fringe social media in early 2021 by 10.9 percentage points (pp) (95% CI [5.9 pp, 15.9 pp]) on desktop devices, and by 15.9 pp (95% CI [10.2 pp, 21.7 pp]) on mobile devices, without decreasing activity on fringe social media in general (including Parler).

Our results indicate that deplatforming Parler was ineffective at reducing the consumption of the type of content that was deplatformed. Web stakeholders may benefit from this insight by reconsidering platform-level interventions and by, *e.g.*, promoting simultaneous action against multiple fringe social media platforms or acting proactively rather than reactively during periods of political unrest.

### 1.2.3 Banning influencers (Chapter 5)

*Submitted to The 27th ACM Conference on Computer-Supported Cooperative Work and Social Computing*

From politicians to podcast hosts, online platforms have systematically banned ("deplatformed") influential users for breaking platform guidelines. Previous inquiries on the effectiveness of this intervention are inconclusive because 1) they consider only a few deplatforming events; 2) they consider only overt engagement traces (*e.g.*, likes and posts) but not passive engagement (*e.g.*, views); 3) they do not consider all the potential places users impacted by the deplatforming event might migrate to. We address these limitations in a longitudinal, quasi-experimental study of 165 deplatforming events targeted at 101 influencers.

We collect deplatforming events from Reddit posts and then manually curate the data, ensuring the correctness of a large dataset of deplatforming events. Then, we link these events to Google Trends and Wikipedia page views, platform-agnostic measures of online attention that capture the general public's interest in specific influencers. Through a difference-in-differences approach, we find that deplatforming reduces online attention toward influencers.

After 12 months, we estimate that online attention toward deplatformed influencers is reduced by −63% (95% CI [−75%,−46%]) on Google and by −43% (95% CI [−57%,−24%]) on Wikipedia. As we study over a hundred deplatforming events, we can analyze in which cases deplatforming is more or less impactful, revealing nuances about the intervention. Notably, we find that both permanent and temporary deplatforming reduce online attention toward influencers.

We contribute to an emerging literature mapping of the effectiveness of content moderation interventions and help guide platform governance practices. Our results offer empirical evidence that sanctioning influencers significantly reduces online attention directed at them, and that platforms and other stakeholders should consider temporary bans to prevent harm caused by influencers' online presence, particularly those receiving widespread online attention.

### 1.2.4 Removing content (Chapter 6)

*Originally published at The ACM Web Conference 2023 [95]*

Online social media platforms use automated moderation systems to remove or reduce the visibility of rule-breaking content. While previous work has documented the importance of manual content moderation, the effects of automated content moderation remain largely unknown. Moderation interventions may increase compliance with community guidelines, *e.g.*, as deleted comments may prevent a conversation from derailing, or, reversely, backfire and increase rule breaking, *e.g.*, because sanctioned users perceive the decision as unfair.

Here, in a large study of Facebook comments ($N$=412M), we used a fuzzy regression discontinuity design to measure the impact of automated content moderation on subsequent rule-breaking behavior (number of comments hidden/deleted) and engagement (number of additional comments posted).

We find that deleting comments reduced rule-breaking behavior in the thread where the comment was originally posted. Deleting comments also reduced rule-breaking among users whose comments were deleted, *i.e.*, other comments in the thread or that the user subsequently posted were hidden and deleted less often after the intervention. At the thread level, deleting rule-breaking comments significantly decreased rule-breaking behavior in threads with 20 or fewer comments before the intervention, even among other participants in the thread. This effect was statistically insignificant for threads with more than 20 comments. Deletion at the user level led to a decrease in subsequent rule breaking and posting activity. However, while the reduction in rule breaking persisted with time, the decrease in posting activity waned.

Our results indicate that automatically deleting rule-breaking content, as currently applied on Facebook, decreases the subsequent creation of content that goes against community guidelines. Our results suggest two ways that this may happen. First, users whose comments are deleted are less likely to produce subsequent rule-breaking content. Second, other users are also less likely to create rule-breaking comments in the thread where the content was deleted.

### 1.2.5 Pre-approving content (Chapter 7)

*Originally published The 16th International AAAI Conference on Web and Social Media [96]*

In many online communities, community leaders (*e.g.*, moderators, and administrators) can proactively filter undesired content by requiring posts to be approved before publication. However, although many communities adopt post approvals, little research has been done on their impact on community behavior. On the one hand, well-moderated spaces are more attractive to users and can improve the quality of users' contributions. However, overenforcement of rules can discourage participation, and moderation creates more work for leaders.

Here, we analyze 233,402 Facebook Groups from March to July 2021, comparing communities that enabled post approvals (PA-ON; *N*=8,767) to communities that did not change any moderation-related settings (PA-OFF; *N*=224,635). First, we examine the factors that led to a community adopting post approvals by comparing the activity in PA-ON and PA-OFF communities four weeks before the former enabled the setting. Second, we study how the setting shapes subsequent user activity and moderation in the group in a matched study (*N*=14,682) considering PA-ON and PA-OFF communities that were similar before the former enabled the setting.

We find that communities that adopted post approvals tended to do so following sudden increases in user activity (*e.g.*, comments) and moderation (*e.g.*, reported posts). This adoption of post approvals led to fewer but higher-quality posts. Though fewer posts were shared after adoption, not only did community members write more comments, use more reactions, and spend more time on the posts that were shared, they also reported these posts less. Further, post approvals did not significantly increase the average time leaders spent in the group, though groups that enabled the setting tended to appoint more leaders. Last, the impact of post approvals varied with both group size and how the setting was used; for example, group size mediates whether leaders spent more or less time in the group after adopting the setting.

Our findings suggest that post approvals substantially change how online communities work and that the setting creates communities centered around fewer, higher-quality posts. These insights may guide improvements to community-level moderation processes and the quasi-experimental approach we adopted can be easily extended to analyze other opt-in features provided by social media platforms.

### 1.2.6 Guiding content creation (Chapter 8)

*Submitted to The 27th ACM Conference on Computer-Supported Cooperative Work and Social Computing*

Content moderation in online communities is often a delicate balance between maintaining content quality and fostering user participation. In this chapter, we introduce *Post Guidance*, a novel approach to community moderation that proactively guides users' contributions using rules that trigger interventions as users draft a post to be submitted; *e.g.*, rules can surface messages to users, prevent post submissions, or flag posted content for review.

We evaluate a version of Post Guidance implemented on Reddit, which enables the creation of rules based on both post content and account characteristics, via a large randomized experiment, capturing activity from 97,616 posters in 33 subreddits over 63 days. Using behavioral logs of the 28 days after enrollment, we calculated aggregated measures of (1) the users' posting activity, (2) moderation received on those posts, (3) community engagement with those posts, and (4) the users' overall engagement in the subreddit. This randomized setup allows us to analyze the causal effects of Post Guidance using simple regression analysis.

We find that Post Guidance (1) increased the number of “successful posts” (posts not removed after 72 hours), (2) decreased moderators’ workload in terms of manually-reviewed reports, (3) increased contribution quality, as measured by community engagement, and (4) had no impact on posters’ own subsequent activity, within communities adopting the feature. Post Guidance on Reddit was similarly effective for community veterans and newcomers, with greater benefits in communities that used the feature more extensively.

This study finds Post Guidance to be effective as a scaleable and flexible content moderation paradigm, with the potential to improve user-generated content across the Web. Though this approach is currently implemented for posts on Reddit, it would easily adapt to various contribution types and platforms, *e.g.*, comments on social media, direct messages in instant messaging platforms, and wiki contributions in collaborative encyclopedias. Future work could extend this new paradigm “breadth-wise,” *i.e.*, adapting it to other parts of the Web, or “depth-wise,” *i.e.*, increasing the possibilities for rules and nudges and quality of the feedback.

## 1.3 Publication list

The following publications are associated with this thesis. Note that other work published during my doctorate can be found in the curriculum vitae at the end of the thesis.

- Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-24.
- Horta Ribeiro, M., Jhaver, S., Reignier-Tayar, M., & West, R. (2024). Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them. arXiv preprint arXiv:2401.01253.
- Horta Ribeiro, M., Hosseinmardi, H., West, R., & Watts, D. J. (2023). Deplatforming did not decrease Parler users’ activity on fringe social media. PNAS nexus, 2(3), pgad035.
- Horta Ribeiro, M., Cheng, J., & West, R. (2023, April). Automated content moderation increases adherence to community guidelines. In Proceedings of the ACM web conference 2023 (pp. 2666-2676).
- Horta Ribeiro, M., Cheng, J., & West, R. (2022, May). Post approvals in online communities. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 16, pp. 335-346).

# 2 Background

Online content moderation is a widely studied topic in computer science and beyond. Here, we provide a broad background on content moderation and set the scope for the thesis' contributions. We discuss antisocial behavior online, content moderation paradigms, and content moderation interventions, *i.e.*, concrete practices employed by online platforms.

## 2.1 Online antisocial behavior

In 2021, Pew Research estimated that four in ten Americans have experienced online harassment [32]. And it would be unfair to say this is a novel phenomenon—antisocial behavior, *i.e.*, those that harm others, has been present on social media since its early days [54]. Online antisocial behavior (sometimes also referred to as cyber-deviant behavior [111]) comes in all shapes and forms [262]: trolling, hate speech, stalking, doxing, cyber-bullying, SWATing, Zoomboombing. Perhaps unsurprisingly, it has been taxonomized in different ways [249, 262, 288, 127]; although the line between different categories of antisocial online behavior (*e.g.*, toxicity vs. cyber-bullying) is often blurry. Last, online antisocial behaviors are linked to some of the critical challenges brought forth by online platforms, including online radicalization, the spread of mis/disinformation, and political polarization [101, 150, 42, 141, 152].

### 2.1.1 Characterizing online antisocial behavior

Motivated by the effects of antisocial behavior on people's lives [5, 284, 72], a vast body of research has characterized antisocial behavior across platforms, languages, and contexts [290, 262, 225, 185]. Cheng et al. [38] study antisocial behavior in the online communities formed around three large news sites (CNN, Breitbart, and IGN). Saha et al. [226] characterized messages that attempt to instill fear among the population about a specific minority in Indian WhatsApp groups. Barlett et al. [13] study cross-cultural differences in cyberbullying behavior, comparing samples from the United States and Japan. Overall, this research has found antisocial behavior widespread on the Web but heterogeneous across contexts.

### 2.1.2 Detecting online antisocial behavior

Closely related to the literature *characterizing online antisocial behavior* is research developing quantitative methods, including machine learning models, to detect various types of antisocial behavior, including cyberbullying [200], hate speech [124], trolling [131], and online harassment [247]. Broadly, the methods employed fall under one of two categories. They either *(a)* count words related to the antisocial behavior of interest (*e.g.*, using HateBase [92]); or *(b)* deploy machine-learning based method (*e.g.*, Google's Perspective API [192]). Unfortunately, measuring some antisocial behaviors, *e.g.*, hate speech, through text is difficult due to their contextual nature [53], which has led researchers to try to incorporate contextual cues in its classification [98].

Many commonly used classifiers generate scores that are subsequently used to determine when interventions are appropriate. For example, Google Perspective's flagship classifier [192] outputs a "toxicity" score for short texts that reflect "rude, disrespectful or unreasonable comments that are likely to make someone leave a discussion." This score has been used to proactively intervene upon potentially rule-breaking content, for instance, on Coral, an open-source commenting platform used in over 100 newsrooms, including the Washington Post and Der Spieger [195]. Lastly, automated measures of antisocial behavior, particularly online toxicity, have been widely used by researchers to study the effect of content moderation interventions in online platforms [265, 221].

### 2.1.3 Causes of online antisocial behavior

The perception of a lack of civility online led to a rich literature exploring the mechanisms causing antisocial behavior online [248, 39, 128]. Central to this research agenda is the idea of an "online disinhibition effect," *i.e.*, that differences between how we communicate online and offline would translate to differences in how we behave across both settings [248].

We highlight two mechanisms that have been widely studied. First, a large body of work has examined whether anonymity increases antisocial behavior [57, 126, 135, 4]. *e.g.*, already in 1984, before the advent of the Web, Kiesler et al. showed that participants had higher degrees of disinhibition, reflected in flaming behavior when anonymously communicating online [126]. Second, another line of research has examined whether anti-social behavior is "contagious," *i.e.*, studying a user's likelihood to produce trolling or uncivil content after being exposed to similar content. In both cases, findings have been mixed: some papers have found that rule-breaking behavior can spread from comment to comment [39, 128], while others have found null results [91, 217]. Likewise, some papers have found that anonymity significantly increases antisocial behavior online [126, 290, 229], while others found null results or negative effects [57, 4]. Disagreements among these past studies may be explained by the fact many of these studies were conducted in labs (which can limit or call into question their ecological validity) and their relatively small scale (which would prevent studies from capturing the heterogeneity of the effect).

## 2.2 Content moderation

Online platforms like Facebook, Amazon, and Tinder have content moderation policies [75], as do the online communities within them [64]. These policies are often (but not always) paired with moderation *practices*, *i.e.*, how the policies are carried out in practice. Altogether, policies and practices prevent antisocial behavior and their subsequent harms [117, 61] and allow communities and platforms to thrive [219, 237]. In what follows, we discuss different ways online platforms and communities have organized and practiced content moderation.

### 2.2.1 Platform-level and community-level moderation

Major online platforms have dedicated teams of paid content moderators spread throughout the globe, enforcing the platforms' policies and guidelines upon user-generated content [75]. Different aspects of "commercial" content moderation have been extensively criticized by scholars [75, 211, 84]. "Commercial content moderation" would be often performed by workers who 1) lack the cultural context associated with specific communities necessary to interpret content [211], 2) must to engage with gruesome, cruel, and hateful content for hours on end [75]; and 3) have precarious work conditions; they are poorly paid, have unstable earnings, and rely on (often algorithmic) reputation systems to maintain a steady stream of income [84]

In contrast, on platforms such as Reddit, Facebook Groups, or Discord, users create their own online communities, and each community often represents its microcosm with different goals and idiosyncrasies. On Reddit, for example, while some norms and expectations are shared across communities, previous research has found a long tail of explicit and implicit norms peculiar to some corners of the website [64, 33]. Given the highly contextual nature of the rules, a one-size-fits-all moderation approach is unfeasible, and, therefore, the work of enforcing community-specific rules is performed by volunteer moderators whose powers concern specific communities [60]. Recent work suggests that volunteer self-organized community moderation is a step into better-governed online spaces, as decisions would be more contextual and legitimate [234]. From that viewpoint, scholars have argued that the extensive [142] moderation work carried out by volunteers carries immense commercial [143] and civic value [155], as it enables meaningful discourse online.

### 2.2.2 Automated content moderation

Whether moderation is done on the platform or the community level, automated tools have played an important role in allowing moderators to handle rule-breaking behavior. For example, in mainstream social media platforms like Facebook, Youtube, Twitter/X,and TikTok, most moderation interventions are automated [163, 81, 264, 267]. As previously discussed, these systems typically rely on machine learning classifiers and often struggle with the contextual and multimodal nature of user-generated content [28, 98].

Automated moderation in community-centered platforms typically enables community moderators to set rules about what content should be flagged or removed [55, 116, 293]. Reddit's 'AutoModerator,' for example, empowers moderators to automate content moderation as they see fit. Given a configuration, the AutoModerator enacts moderation decisions like removing or flagging content based on the title or content of the submission or attributes associated with its creator, *e.g.*, its karma. Previous research has found that the AutoModerator reduces the workload and the emotional labor required of content moderation, but it is often too brittle and fails to capture nuances [116].

### 2.2.3 Visibility as content moderation

Many services enable online communities to control the visibility of content in a distributed fashion by allowing users and algorithms to uprank and downrank content. Platforms like Facebook and Instagram use sophisticated ranking algorithms to generate feeds, often using implicit engagement metrics, *e.g.*, predictions about how much time you will spend interacting with the post [165]. These rankings often explicitly try to downrank harmful content, *e.g.*, Instagram notes that: "If our systems detect that a post may contain bullying, hate speech or may incite violence, we'll show it lower on Feeds and Stories of that person's followers [109]".

In contrast, community-driven platforms such as Reddit, Stack Overflow, and Slashdot often rely on explicit user feedback (*e.g.*, upvotes) and the recency of posts [152] to determine what content is shown most saliently to users (this is often done with simple, deterministic algorithms, like Reddit's "Hot" algorithm). This approach effectively filters content that is accepted and appreciated by a community [133]; but previous research has found that it may propagate misinformation [74] and (further) marginalize minorities [52, 152].

### 2.2.4 Monetization as content moderation

Platforms, particularly social media platforms, also moderate content by deciding what is and is not monetizable [104, 149]. The last decades have witnessed the professionalization of the social media influencer [123], individuals (and at times organizations) who make a living out of creating social media content. Many platforms provide explicit monetization schemes to reward popular influencers, defining clear boundaries for monetizing content or creators. For example, YouTube decides who and what gets monetized on its platform [294] and often resorts to demonetization as a moderation strategy for problematic content on the platform [257, 180].

Nonetheless, creators are not entirely dependent on the monetization schemes offered by the platforms they publish their content at; previous research has documented how content creators use *off-platform* monetization strategies [104], e.g., selling merchandising, or establishing a direct line for fans to support their content through platforms like *patreon.com*.

## 2.3 Content moderation interventions

A growing body of work has measured the causal effect of existing content moderation practices through experimental and quasi-experimental research designs. At times, researchers have also proposed new content moderation practices to improve online platforms.

### 2.3.1 Community or platform-level interventions

Some work has focused on large-scale, platform-wide moderation interventions, *e.g.*, banning popular influencers from Twitter or YouTube [115, 202, 129], limiting access or banning toxic online communities [265, 266, 34, 221, 222, 220], and even banning entire fringe social media platforms like Parler [132]. Overall, this work suggests platform-level interventions reduce activity and the capacity of impacted communities or interest groups to attract new members [115, 202, 129]. But this may come at the cost of communities creating smaller, more radical, and polarized communities [265, 266, 6] that are oftentimes less public-facing [270]; *e.g.*, Urman and Katz (2022) found that far-right Telegram groups experienced explosive growth coinciding with the banning of far-right actors on mainstream social media platforms [270].

However, the migration of communities from mainstream to fringe platforms is not the only issue with “deplatforming” as a moderation strategy. Russo et al. (2023) show that deplatforming creates dual-citizen, coactive users who participate in both mainstream and fringe platforms [221]. This finding is particularly relevant to mainstream platforms as it shows that antisocial behavior happening elsewhere can bleed back in, resonating with a body of work indicating the disproportional influence of fringe platforms in shaping our information ecosystem [298, 297].

### 2.3.2 Fine-grained content moderation interventions

Although platform-level moderation interventions can shape our information ecosystem, most content moderation efforts concern micro-level decisions about millions of posts, comments, images, and videos. In that direction, previous work has analyzed the effect of removing content on various platforms [246, 119], banning users [6, 169], and labeling content that might be misleading or polarizing [18, 43, 71, 145, 295, 194, 161, 59]. Altogether, this body of work suggests that the effectiveness of various moderation strategies in shaping subsequent user behavior is mediated by their target and how the intervention is carried out. For example, Jhaver et al. (2019) [114] suggest that users are more likely to post again after having their content removed if they perceive the decision as fair. Likewise, Martel and Rand (2023), in reviewing the warning labels literature, indicate that they are particularly beneficial for decreasing the belief and spread of politically agreeable content but less efficient for content produced by generative artificial intelligence [151].

### 2.3.3 Proactive content moderation

So far, the moderation interventions discussed are reactive: they are enforced *after* users or communities have broken platform rules. Yet, a growing body of research has focused on preventing users from breaking the rules or moderating content before it is widely shared [88, 300]. In that direction, we highlight previous work by Chang et al. (2022), who, in an experiment, found that users became more civil when aided by a tool that indicated when conversations might be going awry [36]. Closely related to *proactive moderation* are access or participation controls [127]. For instance, Seering, Kraut and Dabbish (2017) have studied how streamers use "chat modes" on Twitch that alter how spectators can participate, *e.g.*, allowing only emotes to be sent [236]. They have found that streamers use to shape audience behavior, *e.g.*, preventing spam.

### 2.3.4 User-centric content moderation

Moderation intervention may burden moderators (being "moderator-centric", *i.e.*, moderators are the ones removing posts) or users in the community ("user-centric") In the existing literature, the most noteworthy user-centric interventions entail making norms salient. Bringing attention to the rules and expectations in online spaces has consistently improved user behavior in online platforms—no matter whether done reactively or proactively [269, 154, 122, 114]. For example, Katsaros et al. (2022) show that asking users to post a tweet containing offensive content to reconsider decreased the creation of offensive content [122]. Other work shows that even norms are made salient *after* content is removed, *e.g.*, when "removal explanations" are provided to users, there is an improvement in subsequent user behavior [112]. Last, recent work has shown that providing explanations about removals in public [113] or highlighting may improve user behavior online by itself [278], suggesting that bystanders learn community norms and etiquette by example.

# Moderating entities Part II

# 3 Banning communities

## 3.1 Introduction

The term "content moderation" is commonly associated with the process of screening the appropriateness of user-generated content, as well as imposing penalties on *users* who break the rules [211]. However, social networking platforms sometimes host entire *communities* that systematically defy regulations. There, host platforms oftentimes ban or limit the functionalities of one or several online communities. This has happened, for example, when Reddit decided that not all communities were welcome on the platform [280] and banned subreddits like `r/FatPeopleHate` and `r/transfags`. Also, more recently, following the 2021 storming of the United States Capitol, groups supporting far-right ideologies and the QAnon movement have been banned across different mainstream social media platforms [67].

The extent to which platforms should be the judges, juries, and executioners of these interventions is a topic of debate and has prompted experiments of governance models with societal participation [44]. There, platforms outsource some of their policy decisions—*e.g.*, should we ban an online movement from our platform?—to a panel of experts (*e.g.*, journalists, politicians, lawyers) representing the public interest [8]. Nonetheless, we still ought to answer a preceding question: is community-level moderation effective? We argue why this is not obvious, visually, in Fig. 3.1, depicting possible decisions (*a* and *b*) that users associated with a recently banned toxic[I] community may take. The users may *(a)* continue to be active on the same platform and participate in other groups and communities there, or *(b)* abandon the platform altogether and migrate to a different platform. In both scenarios, community-level moderation could have unintended consequences. In scenario *a*, the moderation measure could set loose an army of trolls across the platform, creating issues in other communities or *new* problematic communities [34]. In scenario *b*, the ban could unintentionally strengthen an alternative platform (*e.g.*, 4chan) where problematic content goes largely unmoderated [153], and where harms inflicted by the toxic community could be higher.

[I] We use 'toxic' as an umbrella term to refer to socially undesirable content: sexist, racist, homophobic, or transphobic posts, targeted harassment, and conspiracy theories that target racial or political groups.

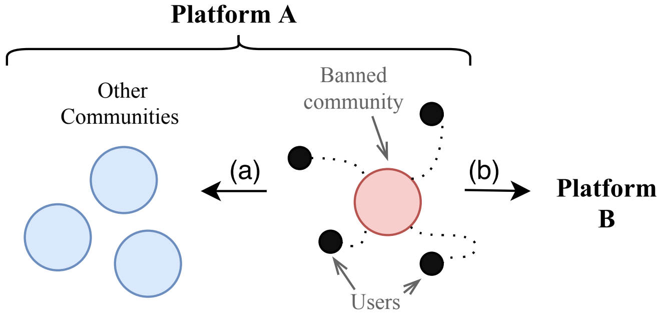


Figure 3.1: Motivation. As a result of community-level bans, users from affected communities may choose to *(a)* participate in other communities on the same platform or *(b)* migrate to an alternative, possibly fringe platform where their behavior is considered acceptable. Scenario *a*, on which most prior work has focused, is more amenable to data-driven analysis. The present paper, on the contrary, focuses on the harder-to-analyze scenario *b*.

Previous work has addressed the *"within platform"* concern. Chandrasekharan et al. [34] and Saleem and Ruths [227] studied what happened following Reddit's 2015 bans, finding that users who remained on the platform drastically decreased their usage of hate speech and that counter-actions taken by users from the banned subreddits were promptly neutralized. More broadly, Rajadesingan et al. [201] showed that, when "toxic users" migrate to healthy communities, they reduce their toxicity levels. Nevertheless, the concern that migrations to an *alternative* platform would strengthen the toxic communities or make them more ideologically radical is still largely unexplored. Existing work suggests that, in the wake of community-level moderation, users actively seek out, and migrate to, alternative websites where their speech will not be censored [239, 174]. However, partly due to the data collection challenges posed by cross-platform studies, quantitative work on the consequences of community-level moderation *across platforms* has remained at the simulation level [120].

**Present work.** This chapter presents an observational study of the efficacy of community-level moderation across platforms. We examine two popular communities that were initially created and grew on Reddit, r/The_Donald, and r/Incels. Faced with sanctions from the platform, they created their own standalone websites—thedonald.win and incels.co—and encouraged their Reddit user base to mass migrate to the new websites. To assess whether community-level moderation measures were effective in reducing the negative impact of these communities (which we refer to as *TD* and *Incels*, respectively), we study how they progressed following their platform migrations. More specifically, we ask:

**RQ1** Have the communities retained their activity levels and their capacity to attract new members following the migration to a new platform?

**RQ2** Have the communities become more toxic or ideologically radical following the migration to a new platform?

Both dimensions are crucial to assess whether community-level moderation measures were truly effective. If the communities simply “changed addresses” and grew larger and more toxic on the new platforms, the moderation measures may have actually increased their capacity to harm society as well as their own members; *e.g.*, outside of Reddit, these communities might orchestrate online harassment campaigns more effectively or disseminate more hate speech.

**Materials and methods.** To study how migrations affect communities, we leverage over 6 million posts made by more than 138 thousand users pooled across the platforms before (Reddit) and after (standalone websites) the migration event. We extract activity-related signals, such as the number of posts, active users, and newcomers, as well as content-related signals, such as algorithmically derived “toxicity scores,” that aim to identify behaviors indicative of user radicalization, such as fixation and group identification [46]. Employing quasi-experimental setups, including matching and regression discontinuity analysis, we study these signals from a *community-level perspective*, analyzing how daily activity and overall content changed, and from a *user-level perspective*, examining how the behavior of individual users changed following platform migrations.

**Summary of findings.** Analyzing activity levels and the inflow of newcomers to the communities (**RQ1**), we find that the moderation measures significantly reduced the overall number of active users, newcomers, and posts in the new communities compared to the original ones. However, individually, users post more often on alternative platforms. A closer look at the users whom we managed to match before *vs.* after the migration suggests that this increase in *relative activity* is due more to self-selection rather than behavior change. Users who migrated were more active in the original platform, and their activity dropped on a user level.

Analyzing changes in the content being posted in the communities following the migration (**RQ2**), we find evidence that users in the TD community became more toxic, negative, and hostile when talking about “objects of fixation” (*e.g.*, democrats, leftists). Changes in the usage of third-person plural (*e.g.*, “they”) and first-person plural (*e.g.*, “we”) pronouns also indicate an increase in ingroup identification and othering language. For the Incel community, we find that changes tend to be statistically non-significant.

**Implications.** Our analysis suggests that community-level moderation measures decrease the capacity of toxic communities to retain their activity levels and attract new members, but that this may come at the expense of making these communities more toxic and ideologically radical. Therefore, as platforms moderate, they should consider their impact not only on their own websites and services but in the context of the Web as a whole. Toxic communities respect no platform boundary, and thus, platforms should consider being more proactive in identifying and sanctioning toxic communities before they have the critical mass to migrate to a standalone website. Overall, we expect that our nuanced analysis will aid stakeholders to take moderation decisions and make moderation policies in an evidence-based fashion.

## 3.2 Background

### 3.2.1 Community-level moderation on Reddit

Reddit employs two community-wide moderation measures: *quarantining* and *banning*. When a community is quarantined, it stops appearing in Reddit's search results and front page. Moreover, users who attempt to access quarantined subreddits (directly through their URLs) are met with a splash page warning them of the shocking or offensive content contained inside. In contrast, banning a community makes it inaccessible and removes all its prior posts. Quarantining frequently precedes banning, so in practice, it serves as a warning for the subreddit to reform itself.

The history of community-level moderation in Reddit dates back to 2015 when Reddit banned five subreddits for infringing their anti-harassment policy [280]. Newell et al. [174] studied how these bans led users to migrate towards alternative platforms (*e.g.*, Voat). Using a mix of self-reported statements and large-scale data analysis, they identified reasons why users left Reddit and found that alternative platforms struggled to attain the same diversity of communities as Reddit. The effects of these bans *within* Reddit were also extensively studied [227, 34], as previously discussed. Overall, findings from these studies suggest that the bans worked for Reddit: they led to sustained reduced interaction of users with the Reddit platform; users who stayed became less toxic after they migrated to other communities within Reddit; and counter-actions taken by users (*e.g.*, creating alternative subreddits) were not effective.

### 3.2.2 Communities of interest

**TD.** The r/The_Donald subreddit (TD) was created on 27 June 2015 to support the then-presidential candidate Donald Trump in his bid for the 2016 U.S. Presidential election. The discussion board, linked with the rise of the Alt-right movement at large, has been denounced as racist, sexist, and islamophobic [148]. Its members often engaged in "political trolling," harassing Trump's opponents, promoting satirical hashtags, and creating memes with pro-Trump and anti-Clinton propaganda [66]. TD is also known for spreading unsubstantiated conspiracy theories like Pizzagate [181] and the Seth Rich murder conspiracy [254].

We depict key events in TD's history in Fig. 3.2. The subreddit was quarantined in mid-June 2019 for violent comments, and on 26 February 2020, Reddit administrators removed a number of TD's moderators, and the community was placed under a "restricted mode," which removed the ability of most of its users to post. Months after the subreddit became inactive, it was banned in late June 2020. While these moderation measures were taking place, TD users were actively organizing a "plan B." In 2017, its members were already considering migrating to alternative platforms [197], and in 2019, after getting quarantined, moderators created a backup site, thedonald.win, that was promoted in the subreddit using stickied posts [198] (*i.e.*, always shown among the first in the feed for the community). TD users continued using

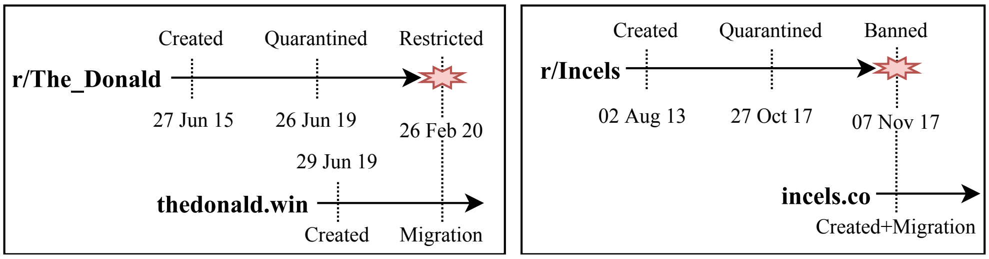


Figure 3.2: Timelines. We depict the dates of creation, quarantining, and banning for the two communities studied here.

the subreddit until the community became “restricted.” Then, they largely flocked to the alternative website [279]. Note that, although TD was eventually banned, we focus here on its “restriction,” since it was this measure that halted user participation and ignited the community migration.

**Incels.** The r/Incels subreddit was created in August 2013. Short for involuntary celibates, it was a community built around “The Black Pill,” the idea that looks play a disproportionate role in finding a relationship and that men who do not conform to beauty standards are doomed to rejection and loneliness [144, 79]. Incels rose to the mainstream due to their association with mass murderers [94] and their obsession with plastic surgery [253]. The community has been linked to a broader set of movements [144, 101] referred to as the “Manosphere,” which espouses anti-feminist ideals and sees a “crisis in masculinity.” In this worldview, men and not women are systematically oppressed by modern society. Lately, specialists have also suggested that these communities may play an important role in radicalizing disenfranchised men and producing ideological echo chambers that promote violent rhetoric [94].

The r/Incels subreddit grew swiftly in early 2017, reaching over 3,000 daily posts [101]. Shortly after, in late October 2017, it was quarantined and then banned two weeks later [256]. In an interview for a podcast [121], one of the subreddits’ former core members, *seargentincel*, mentions that he had already discussed moving the community outside of Reddit with moderators. According to him, when the subreddit was banned, he created the standalone website incels.co, and former r/Incels members quickly organized the migration in Discord channels. Again, we provide exact dates for relevant events in Fig. 3.2.

**Choice of communities.** We study these two communities for two main reasons. First, due to their importance: they have a large number of members and have impacted society at large, *e.g.*, w.r.t. conspiracy theories [254] and real-world violence [94]. Second, these are communities whose migrations were backed by community leaders and that migrated to other public websites. Had the members of these communities spread to a loosely connected network of private channels (*e.g.*, on Telegram), there would be several additional technical and ethical research challenges. (Viewing this choice of the community from 2024, when this thesis is being written, I would argue that this choice was also motivated by data access and by my personal knowledge of these communities due to prior work [101, 97].)

Table 3.1: Overview of our datasets.

| Platform | Community | Submissions | Comments | Users |
|---|---|---|---|---|
| Reddit | /r/Incels | 17,403 | 340,650 | 18,088 |
| | /r/The_Donald | 251,090 | 2,703,615 | 80,002 |
| Websites | Incels.co | 25,138 | 385,765 | 2,270 |
| | thedonald.win | 280,156 | 2,390,641 | 38,510 |

## 3.3 Materials and methods

### 3.3.1 Data collection

We collect data from both Reddit (for the period *before* migrations) and standalone websites (for the period *after*).

**Reddit.** We use Pushshift [15] to collect Reddit data, a service that performs large-scale Reddit crawls. We collect all submissions and comments made on r/The_Donald and r/Incels, starting from 120 days before the moderation measure and until its date. Specifically, for r/Incels, we collect data between 10 July 2017 and 7 November 2017; for r/The_Donald, between 29 October 2019 and 26 February 2020. Overall, we collect around 3 million comments in 270K submissions (or “threads”) from both subreddits (see Table 3.1).

**Standalone websites.** We additionally implement and use custom Web crawlers to collect data from the standalone websites (incels.co and thedonald.win). For each, we collect all submissions and comments posted for a period of 120 days after the community-level moderation measure. Specifically for incels.co, we collect data between 7 November 2017 and 6 March 2018; for thedonald.win, between 26 February 2020 and 24 June 2020. Overall, we collect over 2.5 million comments and submissions from thedonald.win and over 400K comments and submissions from incels.co. In the rest of the chapter, to ease presentation, we refer to both submissions and comments as “posts.”

### 3.3.2 User analysis

We describe our methods for matching users across platforms and analyzing newcomers.

**Matched Users.** To better understand changes at the user-level, we also carry out analyses with matched users, finding pairs of users with the exact same username on both Reddit and the standalone websites. We consider that these users are the same individuals in the two platforms, an assumption backed by anecdotal evidence from within the communities (thedonald.win even had a feature to reserve your Reddit username [259]) and by previous research [174]. Allowing for upper/lower-case differences, using this method, we were able to match 8,651 users between r/The_Donald and thedonald.win (around 20% of the user base of the latter) and 286 users between r/Incels and incels.co (around 13%).

Table 3.2: Fixation dictionaries.

| | |
|---|---|
| Incels | female(s) normie(s) chad(s) virgin whore(s) girl(s) rope gf girlfriend women beta cunt suicide pussy woman bitch(es) cuck(s) feminism |
| TD | trans commie dem(s) democrat(s) deep communist diversity leftist communism antifa socialist left socialism libs gender |

**Newcomers.** We estimate the inflow of newcomers in each community, considering both the pre- and post-migration periods, by counting the daily number of posts with usernames that were never observed before. For instance, if a user $X$ posted in thedonald.win on the 1st of March of 2020, and no user with such username posted before in either thedonald.win or r/The_Donald, we would consider him or her a newcomer. Note that here, if we used only the data from 120 days before and after the migration, we would observe a spike in newcomers at the beginning of the study period. To prevent that, for each community, we additionally download all history available in Pushshift to act as a buffer.

### 3.3.3 Content analysis

To understand the impact of platform migration on the content being produced by the communities, we use text-based signals associated with toxicity and user radicalization [87, 153].

**Fixation dictionary.** We generate a *fixation dictionary* for each of the communities, selecting terms related to their "objects of fixation." More specifically, we: (1) select terms that are more likely to occur in the communities of interest as compared to Reddit in general, and (2) manually curate these terms, selecting those that are related to these communities' objects of fixation (*e.g.*, *women* and *feminism* for Incels). To obtain the list of terms, we extract words from the communities of interest and from a 1% random sample of Reddit for a period of one month (immediately prior to the study period previously described). We exclude bot-related messages (*e.g.*, auto-moderation), stop-words, and words that occurred fewer than 50 times, and calculate the log-ratio between the frequency of a keyword in the communities being studied and on Reddit in general. From this, we obtain, for each community, the 250 terms that have the highest relative occurrence. Then, to build the fixation dictionary, three researchers (all familiar with the communities at hand) discussed each term and came to an agreement on whether or not that term was an object of fixation. Table 3.2 reports the terms in our fixation dictionary for each community; be advised that the terminology in this table is offensive.

**Toxicity score.** To analyze content toxicity, we use Google's Perspective API [192], an API consisting of machine learning models trained on manually annotated corpora of text. More specifically, we employ the "Severe Toxicity" model which allows us to assess how likely (on a scale between 0 and 1) a post is to be "rude, disrespectful, or unreasonable and is likely to make you leave a discussion." This model is also trained specifically to not classify benign usage of foul language as toxic.

**LIWC.** We measure changes in word choice using the Linguistic Inquiry and Word Count (LIWC) tool [251]. LIWC consists of various dictionaries (in total 4.5K words) to classify words into over 70 categories, including general characteristics of posts (*e.g.*, word count), linguistic components (*e.g.*, adverbs), psychological processes (*e.g.*, cognitive processes), and non-psychological processes (*e.g.*, pronouns). In this work, we study changes for the following (aggregated) LIWC 2015 categories: (1) Negative Emotions: sum of the *Anger*, *Anxiety*, and *Sadness* LIWC categories; (2) Hostility: sum of *Anger*, *Swear*, and *Sexual* LIWC categories. (3) Pronouns: we focus on the usage of third-person plural (*e.g.*, "they"), and first person plural (*e.g.*, "we") pronouns.

**Mapping signals to warning behaviors.** These different signals, as well as their combinations, have been described as warning behaviors of ideological radicalization. We focus on two warning behaviors described by Cohen et al. [46]: (1) *Fixation:* a pathological preoccupation with a person or cause that is increasingly expressed with negative and angry undertones; and (2) *Group Identification:* strong identification and moral commitment to the ingroup and distancing from the outgroup. To study changes in *fixation*, we analyze our fixation dictionaries along with *Toxicity scores* and the word categories *Negative Emotions* and *Hostility*. To study changes in *group identification*, we study changes in the usage of pronouns as measured by LIWC. These choices were motivated, as discussed earlier, by previous work by Grover and Mark [87].

### 3.3.4 Ethics and reproducibility

In this work, we only used data publicly posted on the Web and did not (1) interact with online users in any way, nor (2) simulate any logged-in activity on Reddit or the other platforms. When we matched users on Reddit and the fringe platforms, we did not attempt to gain any information about users' personal identities. Anonymized reproducibility data and code are available at https://doi.org/10.5281/zenodo.5171068 We stress that the data is provided without the usernames or the actual text posted (i.e., only the signals extracted). We believe that this makes de-anonymization harder than crawling the standalone websites and downloading existing Reddit dumps. These steps follow previous work studying toxic communities on Reddit [201], and we believe they minimize the potential harms associated while ensuring the study is reproducible. Additionally, we note that we only do *exact matching* on publicly available data while not singling out any individual user, and thus, we believe we are not infringing on reasonable privacy expectations.

## 3.4 Results

### 3.4.1 Changes in activity levels

In this section, we measure how the community-level moderation measures changed posting activity levels and the capacity of the two communities to attract newcomers (**RQ1**). We do so from two different perspectives. First, we aggregate our data on a daily basis, inspecting *community-level* changes in the number of posts, active users, and newcomers. Next, we zoom in to the *user-level* and examine how individual users' behavior changed post-migration.

#### Community-level trends

Fig. 3.3 shows the daily number of newcomers, posts, and active users in each community before and after the migrations for both the TD and the Incel community. Note that we consider data from both the subreddit and the fringe platform users migrated towards (*e.g.*, r/incels and incels.co).

To better understand the overall trends, we perform a regression discontinuity analysis for each statistic in each community. We employ a linear model:

$$y_t = \alpha_0 + \beta_0 t + \alpha i_t + \beta i_t t, \tag{3.1}$$

where $t$ is the date, which takes values between $-120$ and $+120$ and equals 0 in the day of the moderation measure; $y_t$ is statistic we are modeling; and $i_t$ is an indicator variable equal to 1 for days following the moderation measure (*i.e.*, $t > 0$), and 0 otherwise.

Our model assumes that daily activity levels (for the different metrics) can be approximated by a line (defined by coefficients $\alpha_0$ and $\beta_0$), which, post-migration, can change both its intercept ($\alpha$) and its slope ($\beta$). We analyze these changes to understand the impact of platform migrations on the communities at hand.

We exclude data from a "grace period" of 15 days before and after the moderation measure.[II] This accounts for the bursty behavior on user activity metrics on the days around the migration. For example, for newcomers, many of the users who migrated to the new website (thedonald.win or incels.co) choose new usernames, which creates a spike in the metric. However, considering this initial spike makes it harder to capture the overall trend of newcomers on the website, and the grace period addresses that. Additionally, there were a few days on which the Pushshift ingest had problems or where there was a large volume of spam-like content. The values for the statistics on these dates are depicted as gray crosses in Fig. 3.3 and were not considered to fit the models. Both coefficients and 95% CI for each parameter in the regressions are shown in Table 3.3.

[II] We stress that our results are robust to changes in this parameter: we have experimented with different window sizes (*e.g.*, 7 and 21 days), obtaining largely the same results

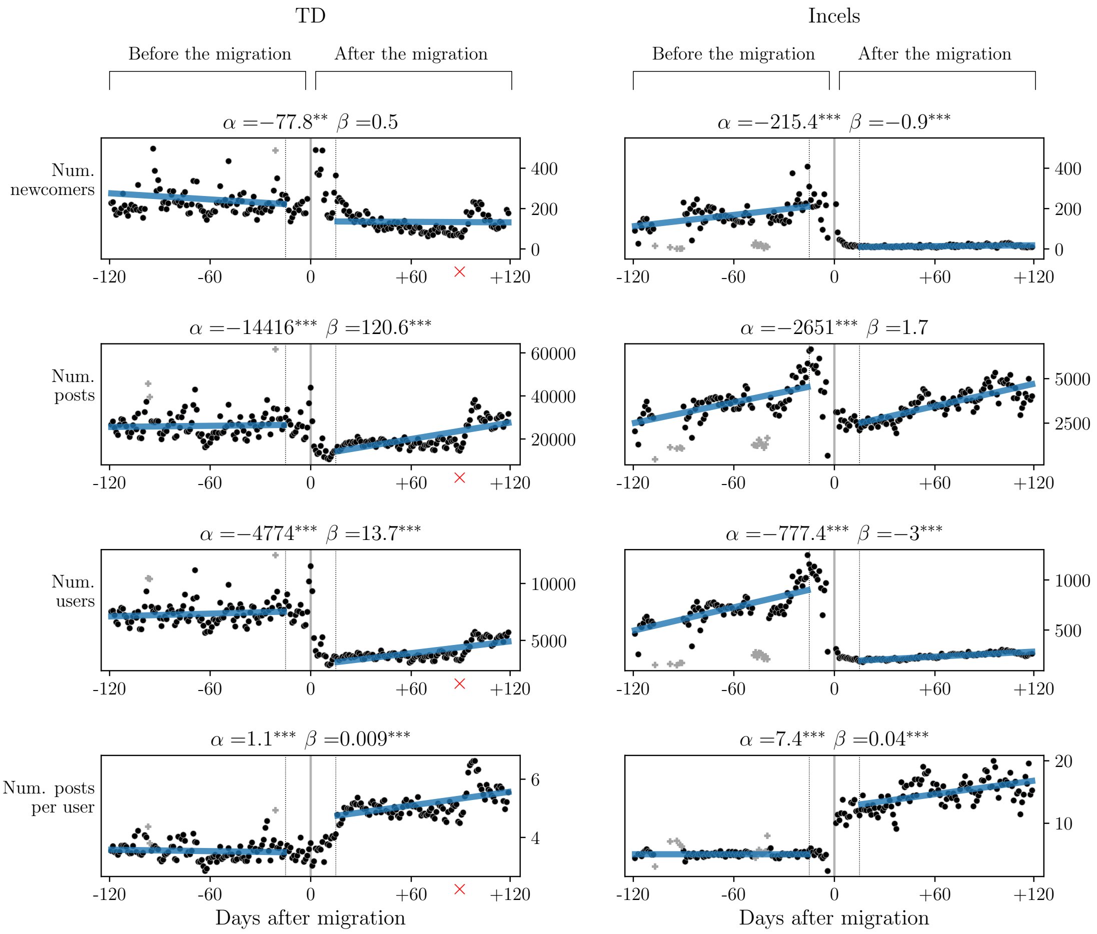


Figure 3.3: Activity levels. Daily activity statistics for the TD community (left) and the Incel community (right) 120 days before and after migrations. Dots represent the daily average for each statistic, and the blue lines depict the model fitted in the regression discontinuity analysis. The migration date and the grace period around it (used in the model) are depicted as solid and dashed gray lines, respectively. Gray crosses represent days where the Pushshift ingest had issues, or where there was a large volume of spam-like content. On top of each subplot, we report the coefficients associated with the moderation measure in the model ($\alpha$ and $\beta$). Coefficients for which $p < 0.001$, 0.01, and 0.05 are marked with ***, **, and *, respectively. For the TD community, we mark the killing of George Floyd (on 25 May 2020), with a red cross (×) close to the x-axis.

**Newcomers.** The first row of Fig. 3.3 shows the number of daily newcomers in each community (as described in Sec. 3.3.2). We find that, for both communities, there was a significant *decrease* in the influx of newcomers following the migration. The TD community saw a significant decrease of around 78 daily newcomers ($\alpha = -77.8$). This represents a percent change of around −30% of the *M*ean *V*alue *B*efore the community-level *I*ntervention (referred to as *MVBI* henceforth), *i.e.*, the drop represents roughly 30% of the average daily value in the pre-migration period. The decrease was even more substantial for the Incel community, which experienced around 215 fewer newcomers a day ($\alpha = -215.4$), roughly −150% of the *MVBI* (note that the drop was, therefore, *bigger* than the pre-migration average). Furthermore, the Incel community had a significant increasing trend before the migration ($\beta_0 = 0.9$), which was weakened in the post-migration period ($\beta = -0.9$).

**Posts and users.** The second and third rows of Fig. 3.3 show that both the total number of daily posts and daily posting users dropped significantly post-migration. TD experienced a decrease of around 14.4k daily posts ($\alpha = -14416$, −55% of the *MVBI*) and of around 4.7k daily active users ($\alpha = -4774$, −65% of the *MVBI*). In both cases, the slope became steeper after the migration, with a significant increase of around 121 new posts a day ($\beta = 120.6$), and around 14 additional active users a day ($\beta = 13.7$). A possible explanation for this increase is that the killing of George Floyd (25 May 2020) and the demonstrations that ensued may have boosted participation on the platform, since the date coincides with a sharp rise in both statistics. Repeating the regression analysis excluding the period after 24 May 2020, we find non-significant *decreases* in the slope ($\beta$) for both statistics, which strengthens this hypothesis. We further discuss this confounder in Sec. 3.5. For the Incel community, there were significant decreases of around 2.6k posts a day ($\alpha = -2651$, −73% of the *MVBI*), and of around 777 daily active users ($\alpha = -777.4$, −116% of the *MVBI*). Looking at the trends for the number of active users, we find a significant positive trend across the whole period ($\beta_0 = 3.9$, see Table 3.3) but the slope decreases significantly after the migration ($\beta = -3$).

**Posts per user.** The fourth row of Fig. 3.3 shows the daily average of the posts per user ratio. Here, we find that the moderation measure *significantly increased* relative activity. The TD community showed an increase in the number of daily posts per user of around 1 extra posts per user ($\alpha = 1.1$, 31% of the *MVBI*); for Incels, the increase was of around 7 extra posts per user ($\alpha = 7.4$, 123% of the *MVBI*). In both cases there is also a significant increase in the trend ($\beta = 0.009$ for TD, and $\beta = 0.04$ for Incel). This adds nuance to the overall scenario: although the activity in the communities is reduced, *relative* to the number of users, it increases.

### User-level trends

The analyses done so far paint a comprehensive picture of the changes in activity due to migration at the community-level. Yet, they do not disentangle the effects happening at the user-level. We found that *relative* activity increased (*i.e.*, fewer users posted more often), but the underlying mechanism for this change is still unclear. Users' activity may have indeed

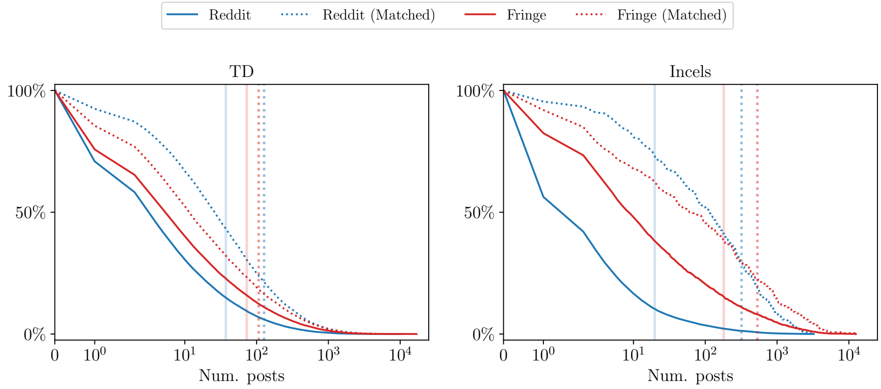


Figure 3.4: CCDFs of posts per user. For each community, we depict the complementary cumulative distribution function (CCDF) of the number of posts per user for: (1) all users who posted in the 120 days *before* (solid blue) and *after* the moderation measure (solid red), and (2) users we managed to match based on username while they were still on Reddit (dashed blue) and on the fringe platform (dashed red). The plot also depicts the mean value for each one of these populations as vertical lines in the same color/style scheme. Recall that the CCDF maps every value in the $x$-axis to the percentage of values in a sample that are bigger than $x$ (in the $y$-axis).

increased *after* the migration, *i.e.*, individually, each user who migrated might post more often on the fringe website, but the increase could also be due to self-selection: users who migrated following the moderation measure might have been more active to begin with.

Understanding the reason behind this (relative) activity increase is important to evaluate the efficacy of the moderation measure. If the increase occurred because users became more active, the subset of users "ignited" by the moderation measure could cause even greater harm in the new platform. However, if the increase was only due to self-selection, we might consider the measure successful in decreasing the activity and reach of the communities. To better understand the mechanism behind this activity increase, we perform an additional set of analyses inspecting what changed at the *user level* post-migration. To do so, we analyze the set of users before and after the migration, and additionally, the set of *matched users* described in Sec. 3.3.2.

**Comparing posts-per-user distributions.** We begin by comparing the distribution of posts from matched users and the general population of users both before and after the migration. Fig. 3.4 depicts the complementary cumulative distribution function (CCDF) of the number of posts for all users (solid line) and for matched users (dashed line) both in the fringe communities (in red) and on Reddit (in blue).

Considering all users (solid lines), the CCDFs confirm our previous analysis, showing that users are more active on the fringe websites since the red solid line is consistently above the blue solid line. This is also captured by the mean of user activity in fringe communities, which

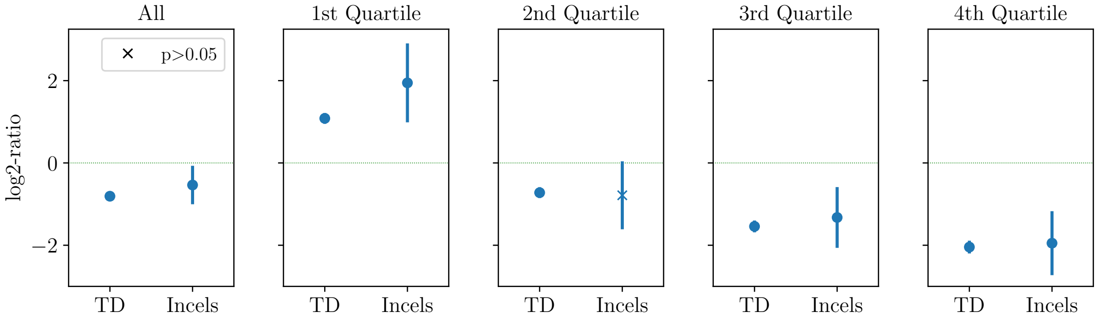


Figure 3.5: User-level change in number of posts. Mean log-ratios between the number of posts before and after the migration for each user. In the first column, the mean is calculated for all users, while for the last four, we stratify users according to their level of pre-migration activity. The horizontal line depicts the scenario where the number of posts remained the same (log-ratio = 0). Error bars represent 95% CIs.

is of 73 posts per user (95% CI: [70, 76.3][III]) for the TD community, and of 180.6 (95% CI: [155.1, 209.3]) for Incels. These values are significantly higher than in Reddit, where there are, on average, 37.3 posts per user (95% CI: [36.2, 38.5]) in the TD community, and 19.8 (95% CI: [18.2, 21.5]) in the Incel community.

Comparing the number of posts per user on Reddit in general (blue line) with users we managed to match (dashed blue line), we find that matched users are more active than users in general. In Reddit, matched users had an average of 127.3 posts (95% CI: [120.2, 135]) in the TD community, and 319.9 posts (95% CI: [264.7, 381.8]) in the Incel community, significantly higher than the average user in Reddit in each community (reported above).

**Matched comparisons.** The above analysis suggests that users who migrated were more active than average on Reddit, which could lead to an increase in *relative* activity due to self-selection. To further investigate this, we compare, for each matched user, the change in number of posts before and after the migration. More specifically, we analyze the log-ratio of posts before *vs.* after the migration for each matched user, defined as $\log_2 \frac{\text{\# posts after}}{\text{\# posts before}}$. Note that this metric provides an intuitive interpretation of the change in activity for a user: if the numbers of posts before and after the migration are the same, the log-ratio will be 0; if the user posted twice as much, it will be 1; and if the user posted half as much, $-1$.

In Fig. 3.5, we depict the mean value of the log-ratios for all users in the first column and, in the following four columns, for users stratified by their activity in the pre-migration period. We divide users into quartiles according to how much they posted in the pre-migration period[IV] and then report the mean for each quartile.

[III] Confidence intervals were calculated through bootstrapping.

[IV] For the TD community, the quartile ranges for the number of posts before the migration were $Q1 = [1,7)$, $Q2 = [7,27)$, $Q3 = [27,101)$, $Q4 = [101,\infty)$; for Incels, $Q1 = [1,19)$, $Q2 = [19,116)$, $Q3 = [116,398)$, $Q4 = [398,\infty)$.

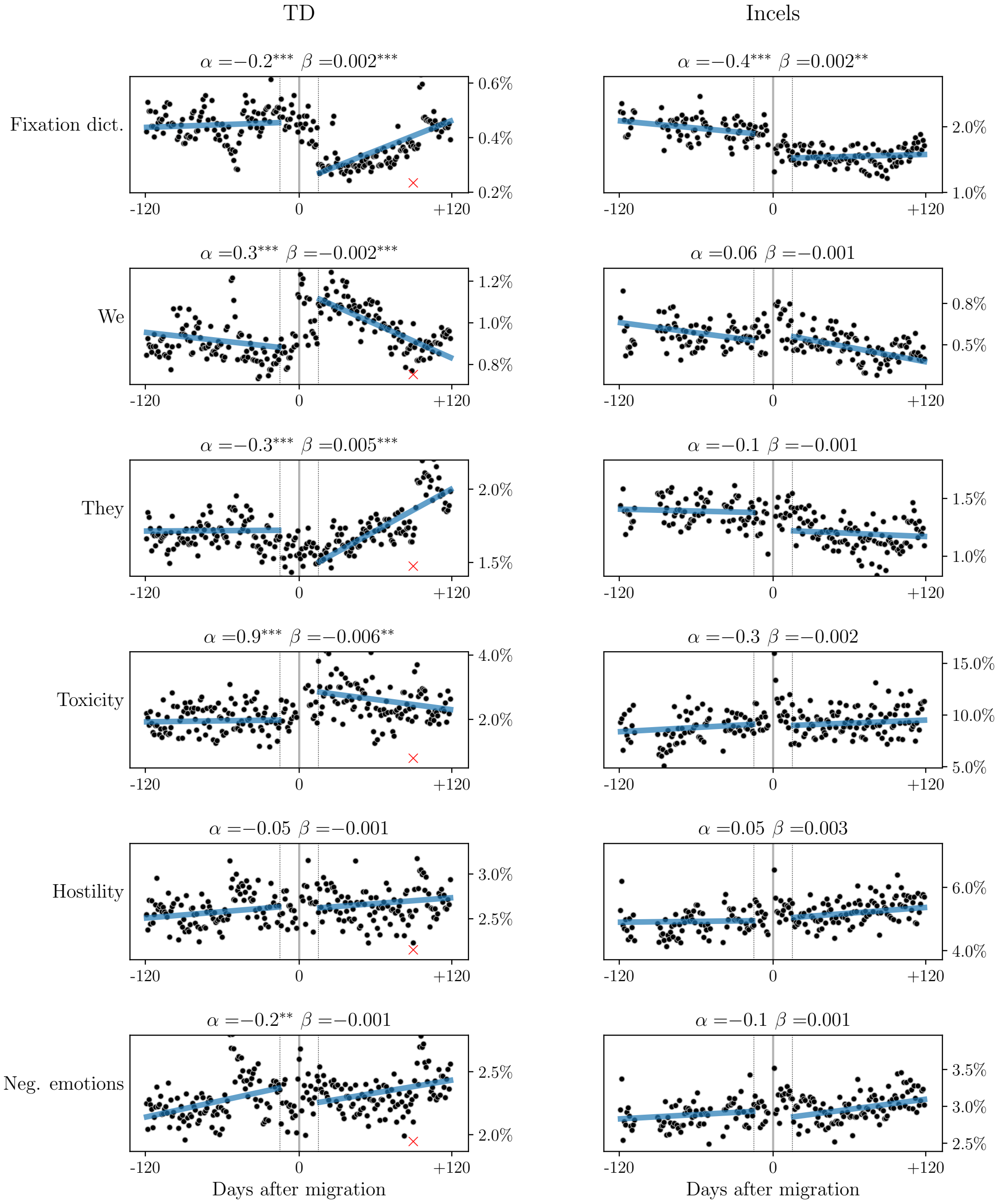


Figure 3.6: Content signals. Daily content-related statistics for the TD community (left) and the Incel community (right) 120 days before and after migrations. For the *Fixation Dictionary* and the LIWC-related metrics, black dots depict, for each day, the percentage of words belonging to each word category. For *Toxicity*, they depict the daily percentage of posts with toxicity scores higher than 0.80. For *Toxicity*, *Negative Emotions*, and *Hostility*, we limit our analysis to posts that contain at least one word in our fixation dictionaries. We again show the output of our model as solid blue lines, and the coefficients related to the moderation measure ($\alpha$ and $\beta$) on top of each plot (marking those for which $p < 0.001$, 0.01, and 0.05 with ***, **, and *, respectively). For the TD community, we mark the killing of George Floyd, with a red cross (×) close to the x-axis.

Considering the complete set of matched users (first column of Fig. 3.5), we find that the mean activity log-ratios are significantly smaller than zero for both communities: −0.81 (95% CI: [−0.86, −0.75]) for the TD community and −0.53 (95% CI: [−0.96, −0.10]) for the Incel community. This result provides further evidence for the self-selection hypothesis: not only did we find the group of matched users to be more active, but, within this group, activity has *decreased.*

Analyzing the users stratified by their activity (in the last four columns of Fig. 3.5), we find that this decrease in activity is stronger for users who were the most active in the pre-migration period. The mean log-ratios for each quartile in TD are, respectively, $\mu_{Q1} = 1.1$, $\mu_{Q2} = -0.7$, $\mu_{Q3} = -1.5$, and $\mu_{Q4} = -2.0$. This shows that users in the least active quartile (Q1) became around twice ($2^{1.1}$) as active, while those in the most active quartile (Q4) decreased their activity to around one-quarter ($2^{-2.0}$). For the Incel community, we observe a similar pattern, with mean log-ratios of $\mu_{Q1} = 1.9$, $\mu_{Q2} = -0.8$, $\mu_{Q3} = -1.3$, and $\mu_{Q4} = -1.9$. Overall, these findings mitigate the concern that a core group of extremely dedicated users was “ignited” by the migration.

### 3.4.2 Take-aways

Our analysis suggests that community-level moderation measures significantly hamper activity and growth in the communities we studied. After the moderation measure, the number of newcomers, active users, and posts in both communities substantially decreased. Yet, this tells only part of the story: we also find an increase in the *relative* activity for both communities: per user, substantially more daily posts occurred on the fringe websites.

A closer look into *user-level* indicates that this relative increase in activity is due to self-selection rather than an increase in user activity post-migration. Not only do we find that users we managed to match were more active on Reddit before the migration, but they even *reduced* their overall activity after they went to the new platform.

### 3.4.3 Changes in content

In this section, we use the signals described in Sec. 3.3.3 to analyze whether the communities and their users became more toxic and ideologically radical following the migrations. We again analyze community- and user-level trends separately.

#### Community-level trends

To study community-level trends, we use a regression discontinuity design similar to Equation (3.1); however, we add an extra term to control for changes in length associated with the migration.[V] The model now takes on this form:

$$y_t = \alpha_0 + \beta_0 t + \alpha i_t + \beta i_t t + \gamma l_t, \tag{3.2}$$

where $l_t$ represents the median length of posts (in characters) on day $t$. We add this covariate to ensure that changes in the intercept ($\alpha$) and the slope ($\beta$) following the intervention are not confounded by changes in the way people post on the new platform (e.g., longer posts). Note that a consequence of this added term is that when we plot the number of posts (on the $y$-axis) per day (on the $x$-axis), we no longer get a straight line since changes in the median length (which varies with time) may impact the outcome of the regression. Thus, for the plots, we fix the value of the length as the average value through the entire period in order to isolate the effect of the intervention and simulate that there is no length change. For descriptions of the other coefficients, see Equation (3.1). Again, all coefficients for the regression analysis, along with confidence intervals, are shown in Table 3.3.

**Fixation dictionary.** We begin by inspecting the prevalence of the fixation dictionary terms over time, as depicted in the first row of Fig. 3.6. For the TD community, we observe a significant drop of $\alpha = -0.2$ percentage points in the usage of terms in the fixation dictionary (−44% of the *MVBI*). For the Incel community, following the intervention, we also see a decrease of around $\alpha = -0.4$ percentage points in the usage of words in the fixation dictionary (−21% of the *MVBI*). In both cases, we observe a positive increase in the trend after the intervention ($\beta = 0.002$ for both communities).

**Fixation-related signals.** Next, we study changes in *Toxicity*, *Negative Emotions*, and *Hostility*. We limit this analysis to the set of posts containing at least one word in the fixation dictionary (see Table 3.2) since we are particularly interested in how the communities are talking about their objects of fixation. We consider a comment to be toxic if it has a toxicity score above 80% and calculate, for each day, the fraction of toxic posts. This threshold has been used as a default in other work [296] and production-ready applications that use the API [195]. For the other LIWC-based metrics, we calculate the proportion of words in the specific dictionaries used per day.

The fourth row in Fig. 3.6 shows the changes in the percentage of toxic posts for both communities. For the Incel community, we find no significant change following the interventions. For TD, there is a significant increase right after the intervention of around $\alpha = 0.9$ more toxic posts containing the fixation dictionary (42% of the *MVBI*). However, we see a significant decreasing trend of around $\beta = -0.006$ fewer toxic posts containing words in the fixation dictionary per day. This decrease in the overall trend does not necessarily mean that the average percentage of toxic posts will return to the pre-migration levels. After the sharp increase in toxicity following the moderation measure, the daily toxicity levels may settle at a new baseline higher than pre-migration values.

[V] We find significant changes in the average post length pre- *vs.* post-migration: 131.7 *vs.* 118.0 for Incels, and 129.2 *vs.* 141.2 for TD.

The fifth and sixth rows in Fig. 3.6 depict changes in *Hostility* and *Negative Emotions*, respectively. We find that in most cases these two metrics experience a *decrease* in the intercept following community level interventions, although effects are not always significant ($p > 0.05$).

**Pronoun usage.** In the second and third rows of Fig. 3.6, we report the usage of two types of personal pronouns: first-person plural pronouns (*e.g.*, "we," "us," "our") and third-person plural pronouns (*e.g.*, "they," "their"). For the Incel community, we see no significant change in the usage of either type of pronoun following the migration. For TD, however, there are interesting changes in their usage. For first person plural pronouns, following the intervention, we find a significant increase in usage of around $\alpha = 0.3$ percentage points (33% of the *MVBI*), and a significant *decrease* in the slope, $\beta = -0.002$. For third-person plural pronouns, we find the opposite. Following the intervention, we find a significant *decrease* of $\alpha = -0.3$ percentage points (−18% of the *MVBI*), followed by a significant increase in the trend, $\beta = 0.005$.

First-person plural pronouns capture group identification and third-person plural pronouns have been associated with extremism [191, 87, 46]. Thus, for the TD community, the intervention seems to have transiently increased group identification immediately after the ban, and later, attention seems to have shifted to the outgroup. The reduced focus on the outgroup following the community intervention could also be related to the way words in the *fixation dictionary* were used after migration. There too, we observe a similar pattern: a sharp drop followed by a gradual increase in usage.

Overall, these findings suggest that the community migrations heterogeneously impacted the communities at hand. While not much changed for the Incel community, we find that for TD, there were significant increases in signals related to both the fixation warning behavior (*Toxicity*) and the group identification warning behavior (both first- and third-person plural pronouns).Again, here a potential confound is the death of George Floyd on 25 May 2020,[VI] which impacted user activity (see Fig. 3.3) and coincides with increases in some of the metrics studied (*e.g.*, third-person plural pronouns). By repeating the analysis for TD excluding the period after 24 May 2020, we still find that these changes hold.

### User-level trends

Similar to our content-level analysis, the reasons behind the increase in some of the signals related to online radicalization are important. Here, again, it could be that the subset of users who migrated to the fringe platform was more radical to begin with *or* that the users became more radical after the migration. Thus, it is crucial to analyze changes at the user level. Fortunately, the sample of matched users gives us the opportunity to control for self-selection since we can measure, *e.g.*, the percentage of toxic posts before *vs.* after the migration for the same group of matched users.

[VI] As well as the wave of protests and the nation-wide debate around police brutality that ensued.

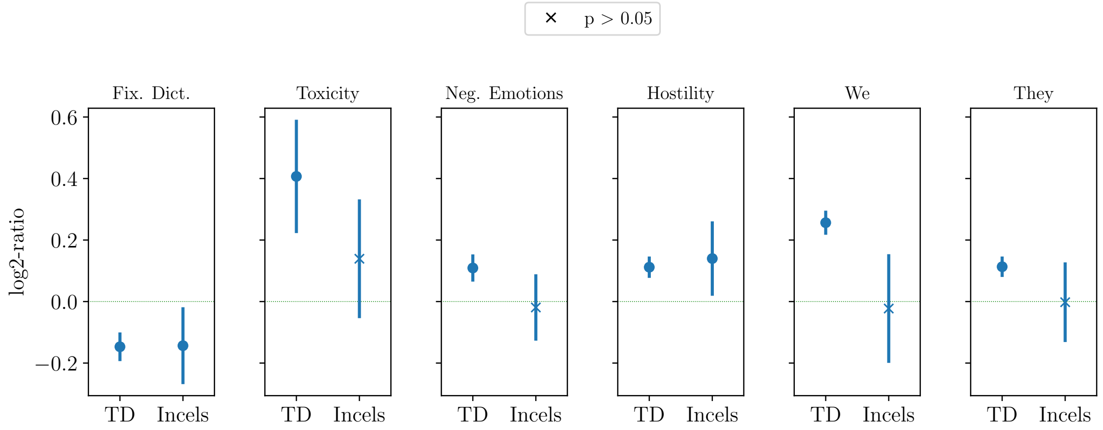


Figure 3.7: User-level change in content. We depict the mean user-level log-ratio for each of the content-related signals studied. A green horizontal line depicts the scenario of no change (log-ratio = 0). Error bars represent 95% CIs.

**Matched comparison.** To disentangle self-selection from user-level increases following the migration, we compare changes in each of the signals for the set of matched users. We calculate, for each user, the fraction of toxic posts (*Toxicity* higher than 0.8) and the percentage of words used in each of the defined categories (*Hostility*, *We*, *etc.*) both before and after the migration. Then, similar to Fig. 3.5, we compare the log-ratio between the signals associated with each user *before* and *after* the migration. However, here, calculating the log-ratio may involve dividing by 0, *e.g.*, for a user who posted no toxic posts before the migration and 2 after. Thus, for each individual signal, we limit our analysis to users with positive values for that signal before and after the migration. Therefore, when comparing the changes in toxic posts, we consider only users with at least one toxic post before and one toxic post after the migration. Similarly, for the LIWC-related signals, we consider only users who used words in the given category at least once before and at least once after the migration. We report the mean log-ratio across matched users for each signal in Fig. 3.7.

For the TD community, we again observe significant increases for *Toxicity* ($\mu = 0.41$, which represents an increase of around 32% since $2^{0.41} \approx 1.32$), *We* ($\mu = 0.26$, 20% increase), and *They* ($\mu = 0.11$, 8% increase). This suggests that the increases previously observed were not caused merely by self-selection. For the Incel community, there were non-significant increases in *Toxicity* (0.14, 10% increase) and small non-significant decreases in the usage of both pronoun-related categories. For both communities, we again significant decreases in the usage of words in the fixation dictionary ($\mu = 0.15$ for TD and $\mu = 0.14$ for Incels, around 11% increase in both cases). We also find significant increases in signals that we did not observe in the community-level analysis. Namely, for both communities we find significant increases in *Hostility* (8% increase for TD and 10% for Incels), and for TD, we find a significant increase in *Negative Emotions* (8% increase).

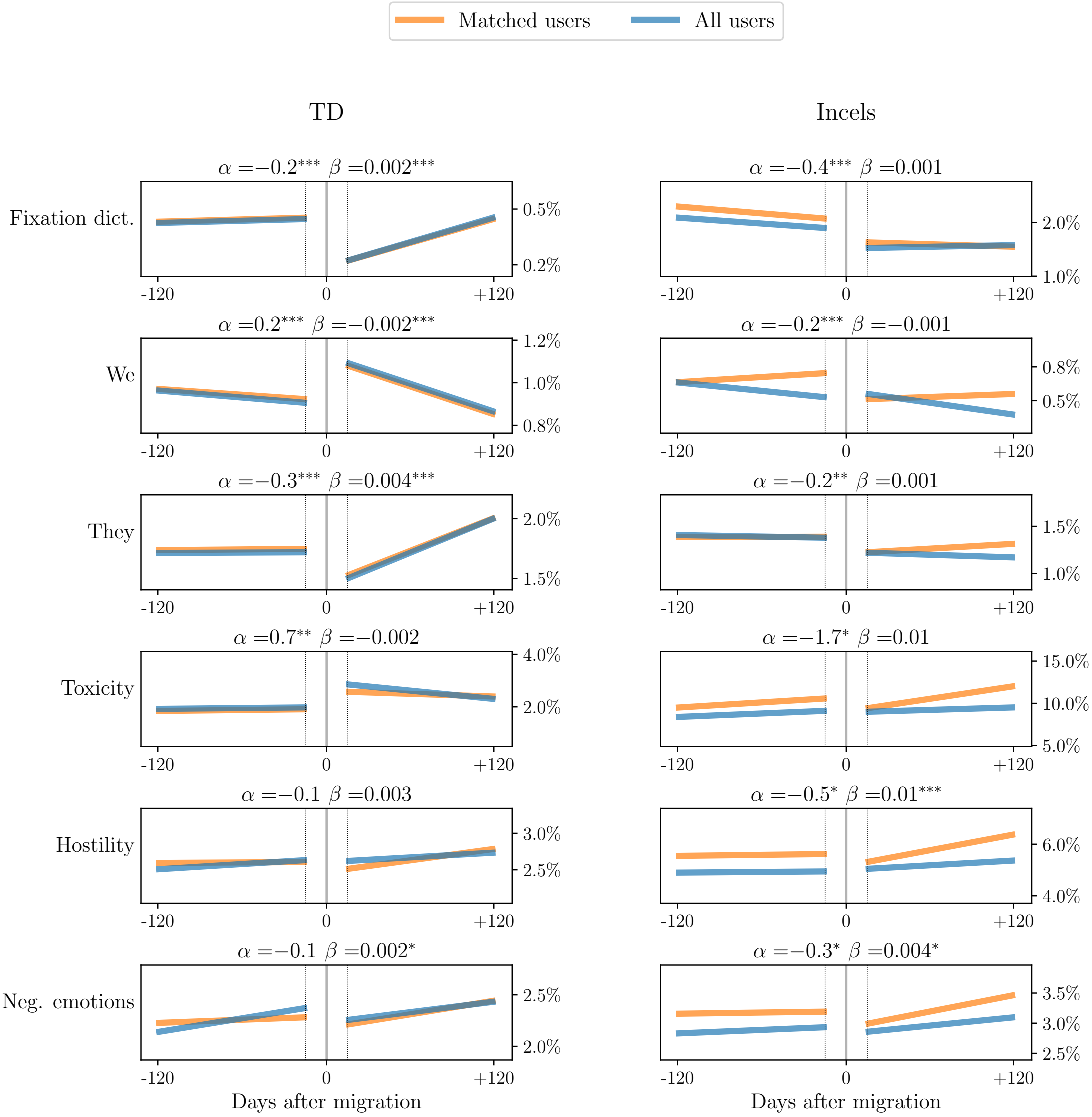


Figure 3.8: Daily content signals for matched users. We repeat the same analysis from Sec. 3.4.3 considering the sample of matched users. We show the regression lines considering all users (in blue) and only matched users (in orange). Above each plot, we show the coefficients related to the moderation ($\alpha$ and $\beta$) for the model considering only matched users. For additional details, see Fig. 3.6.

Table 3.3: Coefficients for all regression discontinuity analyses done throughout the chapter, including 95% confidence intervals. Coefficients for which $p < 0.001$, 0.01, and 0.05 are marked with ***, **, and *, respectively. The value $[10^{-3}]$ at the beginning of a cell indicates that the value of the cell, as well as the confidence intervals presented, should be multiplied by $10^{-3}$. This may cause slight differences in the numbers in this table and the ones presented in the plots, since here we present the results at higher precision. Note that this table contains the regression results for three different analysis carried out throughout the chapter and depicted in Fig. 3.3, Fig. 3.6, and Fig. 3.8. For presentation reasons, we omit the confidence intervals for the intercept across the whole period ($\alpha_0$), which is significant ($p < 0.001$) across all of the models.

(a) Community-level activity (Fig. 3.3)

| Venue | Statistic | $\alpha_0$ | $\beta_0$ | $\alpha$ | $\beta$ | $R^2$ |
|---|---|---|---|---|---|---|
| TD | #newcomers | $214.7^{***}$ | $-0.5\,(-1.3, 0.2)$ | $-77.8^{**}\,(-127.9, -27.7)$ | $0.5\,(-0.5, 1.4)$ | 0.24 |
|  | #posts | $26650^{***}$ | $8.8\,(-17.9, 35.5)$ | $-14416^{***}\,(-16947, -11886)$ | $120.6^{***}\,(83.9, 157.2)$ | 0.41 |
|  | #users | $7593^{***}$ | $4\,(-0.5, 8.5)$ | $-4774^{***}\,(-5184, -4365)$ | $13.7^{***}\,(8.1, 19.3)$ | 0.85 |
|  | #posts/#users | $3.5^{***}$ | $-0.001\,(-0.003, 0.001)$ | $1.1^{***}\,(0.9, 1.3)$ | $0.009^{***}\,(0.006, 0.01)$ | 0.85 |
| Incels | #newcomers | $224.9^{***}$ | $0.9^{***}\,(0.5, 1.4)$ | $-215.4^{***}\,(-250.3, -180.5)$ | $-0.9^{***}\,(-1.3, -0.4)$ | 0.84 |
|  | #posts | $4840^{***}$ | $19.5^{***}\,(14.1, 25)$ | $-2651^{***}\,(-3098, -2203)$ | $1.7\,(-4.8, 8.2)$ | 0.55 |
|  | #users | $960.1^{***}$ | $3.9^{***}\,(2.9, 4.9)$ | $-777.4^{***}\,(-850.2, -704.6)$ | $-3^{***}\,(-4, -2.1)$ | 0.93 |
|  | #posts/#users | $5^{***}$ | $-0.001\,(-0.003, 0.003)$ | $7.4^{***}\,(6.6, 8.3)$ | $0.04^{***}\,(0.02, 0.05)$ | 0.92 |

(b) Community-level content (Fig. 3.6)

| Venue | Statistic | $\alpha_0$ | $\beta_0$ | $\alpha$ | $\beta$ | $R^2$ |
|---|---|---|---|---|---|---|
| TD | Fix. Dict | $0.6^{***}$ | $[10^{-3}]0.2\,(-0.2, 0.5)$ | $-0.2^{***}\,(-0.3, -0.2)$ | $[10^{-3}]1.7^{***}\,(1.2, 2.2)$ | 0.53 |
|  | Toxicity | $3.1^{***}$ | $[10^{-3}]0.5\,(-2, 3.1)$ | $0.9^{***}\,(0.6, 1.3)$ | $[10^{-3}]-5.9^{**}\,(-10.1, -1.7)$ | 0.3 |
|  | Neg. Emotion | $2.7^{***}$ | $[10^{-3}]2.2^{***}\,(1.1, 3.3)$ | $-0.2^{**}\,(-0.3, -0.06)$ | $[10^{-3}]-0.5\,(-2.1, 1.1)$ | 0.14 |
|  | Hostility | $3.3^{***}$ | $[10^{-3}]1.2\,(-0.04, 2.4)$ | $-0.05\,(-0.2, 0.09)$ | $[10^{-3}]-0.1\,(-2, 1.8)$ | 0.08 |
|  | We | $0.6^{***}$ | $[10^{-3}]-0.7^{**}\,(-1.2, -0.2)$ | $0.3^{***}\,(0.2, 0.3)$ | $[10^{-3}]-2.1^{***}\,(-2.7, -1.4)$ | 0.46 |
|  | They | $1.3^{***}$ | $[10^{-3}]0.05\,(-0.5, 0.6)$ | $-0.3^{***}\,(-0.4, -0.2)$ | $[10^{-3}]4.8^{***}\,(3.8, 5.7)$ | 0.55 |
| Incels | Fix. Dict | $1.9^{***}$ | $[10^{-3}]-1.8^{*}\,(-3.3, -0.4)$ | $-0.4^{***}\,(-0.5, -0.2)$ | $[10^{-3}]2.4^{**}\,(0.9, 3.8)$ | 0.69 |
|  | Toxicity | $11.4^{***}$ | $[10^{-3}]6.9\,(-3.1, 16.8)$ | $-0.3\,(-1.2, 0.6)$ | $[10^{-3}]-1.9\,(-13.7, 9.9)$ | 0.13 |
|  | Neg. Emotion | $3.2^{***}$ | $[10^{-3}]1\,(-0.7, 2.7)$ | $-0.1\,(-0.3, 0.03)$ | $[10^{-3}]1.3\,(-0.4, 3.1)$ | 0.25 |
|  | Hostility | $6.3^{***}$ | $[10^{-3}]0.4\,(-3.1, 4)$ | $0.05\,(-0.3, 0.4)$ | $[10^{-3}]2.6\,(-0.8, 6.1)$ | 0.38 |
|  | We | $0.6^{***}$ | $[10^{-3}]-1\,(-2.5, 0.4)$ | $0.06\,(-0.05, 0.2)$ | $[10^{-3}]-0.4\,(-1.3, 0.5)$ | 0.3 |
|  | They | $1.2^{***}$ | $[10^{-3}]-0.3\,(-2.6, 2.1)$ | $-0.1\,(-0.3, 0.04)$ | $[10^{-3}]-0.2\,(-1.6, 1.2)$ | 0.44 |

(c) Community-level content for matched users (Fig. 3.8)

| Venue | Statistic | $\alpha_0$ | $\beta_0$ | $\alpha$ | $\beta$ | $R^2$ |
|---|---|---|---|---|---|---|
| TD | Fix. Dict | $0.6^{***}\,(0.5, 0.8)$ | $[10^{-3}]0.2\,(-0.2, 0.5)$ | $-0.2^{***}\,(-0.3, -0.2)$ | $[10^{-3}]1.6^{***}\,(1.1, 2.1)$ | 0.54 |
|  | Toxicity | $3.2^{***}\,(2.1, 4.3)$ | $[10^{-3}]0.6\,(-2.8, 4)$ | $0.7^{**}\,(0.2, 1.2)$ | $[10^{-3}]-2.4\,(-9, 4.3)$ | 0.13 |
|  | Neg. Emotion | $2.5^{***}\,(2.2, 2.8)$ | $[10^{-3}]0.5\,(-0.5, 1.6)$ | $-0.1\,(-0.2, 0.006)$ | $[10^{-3}]1.7^{*}\,(0.06, 3.3)$ | 0.08 |
|  | Hostility | $3.1^{***}\,(2.7, 3.5)$ | $[10^{-3}]0.1\,(-1.4, 1.6)$ | $-0.1\,(-0.3, 0.04)$ | $[10^{-3}]2.5\,(-0.001, 5)$ | 0.06 |
|  | We | $0.5^{***}\,(0.2, 0.8)$ | $[10^{-3}]-0.6\,(-1.2, 0.007)$ | $0.2^{***}\,(0.2, 0.3)$ | $[10^{-3}]-2.2^{***}\,(-2.9, -1.4)$ | 0.37 |
|  | They | $1.6^{***}\,(1.3, 1.8)$ | $[10^{-3}]0.1\,(-0.6, 0.8)$ | $-0.3^{***}\,(-0.4, -0.2)$ | $[10^{-3}]4.5^{***}\,(3.4, 5.6)$ | 0.44 |
| Incels | Fix. Dict | $1.9^{***}\,(1.6, 2.2)$ | $[10^{-3}]-2.1^{*}\,(-3.9, -0.4)$ | $-0.4^{***}\,(-0.6, -0.2)$ | $[10^{-3}]1.4\,(-0.9, 3.6)$ | 0.68 |
|  | Toxicity | $11^{***}\,(7.9, 14.1)$ | $0.01\,(-0.008, 0.03)$ | $-1.7^{*}\,(-3.2, -0.1)$ | $0.01\,(-0.007, 0.04)$ | 0.08 |
|  | Neg. Emotion | $3^{***}\,(2.6, 3.4)$ | $[10^{-3}]0.3\,(-3, 3.7)$ | $-0.3^{*}\,(-0.5, -0.04)$ | $[10^{-3}]4.2^{*}\,(0.6, 7.8)$ | 0.1 |
|  | Hostility | $6^{***}\,(5.3, 6.7)$ | $[10^{-3}]0.7\,(-3.9, 5.3)$ | $-0.5^{*}\,(-0.9, -0.06)$ | $[10^{-3}]9.5^{***}\,(4, 15.1)$ | 0.2 |
|  | We | $0.4^{***}\,(0.3, 0.6)$ | $[10^{-3}]0.6\,(-0.6, 1.8)$ | $-0.2^{***}\,(-0.3, -0.1)$ | $[10^{-3}]-0.3\,(-1.5, 1)$ | 0.44 |
|  | They | $0.9^{***}\,(0.7, 1.2)$ | $[10^{-3}]0.04\,(-1.7, 1.8)$ | $-0.2^{**}\,(-0.3, -0.04)$ | $[10^{-3}]0.8\,(-1.2, 2.8)$ | 0.28 |

**Regression discontinuity analysis.** The previous analysis indicates that there were significant changes in the radicalization-related signals for the matched sample, some of which we did not observe in the community-level analysis. To better understand the matched sample, and the differences between the results at the community-level and the user-level, we repeat the regression discontinuity analysis done for the signals of interest using only posts from the matched user sample. We use exactly the same model as in Equation (3.2), changing only the data: *there* we used all posts by all users, *here* we use all posts by matched users. In Fig. 3.8, we plot the regression lines for the analysis done with all users in blue and for matched users in orange. Coefficients along with confidence intervals are again presented in Table 3.3.

For several signals, the results in this reduced sample are very similar to the previous analysis. For example, for TD, we have almost exactly the same coefficients for the usage of the fixation dictionary ($\alpha = -0.2$, and $\beta = 0.002$) and of third-person plural pronouns ($\alpha = -0.3$, and $\beta = 0.004$). Yet, for some of the signals, we do find significant differences following the community migrations. More specifically, for TD community, following the migration, we find significant increases in the trends for Negative Emotions ($\beta = 0.002$) and we find no significant *decrease* in the trend for the Toxicity signal (which used to be the case). Additionally, for Incels, we find significant increases in the trend for *Negative Emotions* ($\beta = 0.004$) and *Hostility* ($\beta = 0.01$).

Overall, this analysis confirms the results previously discussed in Fig. 3.7 and suggests that the community-level intervention impacted users in the matched sample. This is different from what we observed when looking at activity levels. There, when we zoomed in on matched users, we found that they had *decreased* their activity (even though the number of posts per user grew). Here, on the contrary, we find that these users seem to have become more radical.

### 3.4.4 Take-aways

Altogether, our analysis shows that, for TD, community-level interventions and the migrations that ensued are associated with significant increases in radicalization-related signals. A closer look at the matched user sample indicates that these increases were not merely due to self-selection, since we also observe significant user-level increases. Furthermore, analyzing the matched sample, we find that the migration may have impacted these users more substantially since their differences are more substantial.

A second important result of our content-level analysis is that communities were heterogeneously impacted. When comparing how the *activity* in the two communities changed (Sec. 3.4.1), we found the same patterns overall; whereas, when comparing how the *content* changed, we found rather distinct behaviors across the two communities. Unlike the TD community, for Incels, there were often *decreases* in signals related to radicalization following community migration.

## 3.5 Discussion

Our work paints a nuanced portrait of the benefits and possible backlashes of community-level interventions. On the one hand, we found that the interventions were effective in decreasing activity and the capacity of the community to attract newcomers. Moreover, we found evidence that *relative* increase in activity (i.e., fewer users posting more) is likely due to self-selection: the users who migrated to the new community were more active to begin with. On the other hand, we found significant increases in radicalization-related signals for one of the communities studied (TD), even when controlling for self-selection. In fact, these increases were even more substantial for the set of matched users studied.

An interesting angle to consider the changes observed in communities pre- *vs.* pos- migration, is through the lens of characteristics and affordances of large online platforms such as Reddit, YouTube and Facebook. According to Gillespie [75], online platforms differ from traditional media outlets in that they provide the means of distribution, but not the content (which is user generated). Moreover, a key attribute of online platforms is that they moderate and gatekeep content, despite best efforts to present themselves as neutral "facilitators."

In that context, the migration of online communities from mainstream platforms to fringe, alternative websites provokes shifts associated with how content is distributed and moderated, two important roles of online platforms. The decrease in activity observed after communities migrated emphasizes the *power of the distribution* of online platforms such as Reddit. Since Reddit has thousands of highly popular subreddits, toxic communities inhabiting the platform are easily discoverable, and consuming the content they produce is convenient. Using a similar line of reasoning, the increase in toxicity observed when members of r/The_Donald migrated out of the subreddit can be associated with the *power of the moderation* of online platforms. Toxicity may be understood as a proxy for content that would likely clash with Reddit's content policy[VII]. Therefore, the rise in toxicity following the ban can be understood as a consequence of the removal of platform moderation.

Overall, our results strengthen the hypothesis that a handful of mainstream platforms are largely responsible for our online information ecosystem [75]. Besides determining what kinds of content flourishes [173], platforms allow communities to exploit their affordances to recruit new members, and are able to influence the content being posted in toxic communities. In the remainder of this section, we discuss the implications of these results for platforms and future research, as well as the limitations of our study.

### 3.5.1 Limitations and future work

**Communities.** Our work focuses on two communities: TD and Incels. However, Reddit has sanctioned many other communities that may have migrated to new fringe websites.

[VII] https://www.reddit.com/r/reddit.com/wiki/revisions/contentpolicy

The implications of such sanctions for migration may differ based on the specifics of each community. That said, the communities we study are among the most prominently sanctioned subreddits, and our analysis provides early insight into the consequences of such sanctions. In the future, similar analysis on other sanctioned communities would help disentangle how contextual factors including community size, topic, and the design of the alternative platform may affect migration patterns.

**Migrations and dispersion.** We consider the effects of migration to only one fringe website per each of the sanctioned communities we study. In both cases, the migrations to the websites we analyze were officially endorsed by the subreddits' moderators, and, for r/The_Donald, the subreddit promoted the migration to the new site while it could. However, users may have migrated to other platforms as well. For example, on Reddit, after r/Incels was banned, an old subreddit called r/Braincels reportedly became popular (until eventually being banned too). Also more broadly, some community-level interventions may not result in "successful" coordinated migration. Rather, users can be dispersed through a variety of other platforms (*e.g.*, Gab, 4chan, Parler, *etc.*). Studying what happens in these cases is an important direction to completely understand the impact of deplatforming communities. For example, one could try to measure the activity boost experienced in each of these platforms whenever a toxic community in a mainstream platform (*e.g.* r/The_Donald) gets banned. A challenge here would be to obtain data for a variety of fringe communities and to control for other confounders, such as geopolitical events.

**Confounders.** The responsiveness of these communities to real-world events creates confounders. This is particularly true for the TD community, where we found significant changes in the content- and activity-related signals in reaction to the killing of George Floyd. While our quasi-experimental research design controls for linear trends, sudden bursts in content-related signals can partially impact our results. Controlling for these trends is hard since the reaction of these communities to real world changes is inherently linked to the harms they pose to society. However, in our specific case, we find that the effects observed held even when limiting the period of the regression discontinuity analysis to before the event (i.e., George Floyd's killing). Another possible set of confounders are changes to rules and moderation actions that could have changed pre- *vs.* post-intervention. Although we did not explicitly incorporate these changes into our analysis, we carefully analyzed the set of rules before (Incels: [196], TD: [199]) and after (Incels: [108], TD: [260]) the migration and did not find any substantial changes.

**Matched Users.** Another limitation of the work at hand is that user-level analyses are made on a set of users matched according to their usernames. These users tend to be more active than the average user (*cf.* Fig. 3.4) and may differ from users who migrated and did not change their username. Although important, we argue that this bias is not impactful to the external validity of our results. The main purpose of looking at matched users is to distinguish between behavior change and self-selection. When studying changes in activity, analyzing matched users provides us with the useful insight that, although the number of posts per user increases after the migration (*cf.* Fig. 3.3), on a user-level, this is not the case (*cf.* Fig. 3.5). For the sample bias to be an issue here, reverting or weakening the results, it would be necessary that users who migrated and did not keep the same username became more active after the ban while those who kept the same username did not, which is unlikely. When studying changes in the content, we find that community-level trends on the set of matched users are very similar to community-level changes considering all users (*cf.* Fig. 3.8), weakening concerns that there would be a strong difference between the nature of the content posted by these users. An interesting direction to further understand these matched users (and explore user-level trends) would be to additionally analyze users with known usernames pre- *vs.* post- ban in other subreddits.

**Mapping signals to externalities.** Our analysis relies on user activity and signals derived from user-generated content to analyze online toxic communities. Our main result suggests that community-level interventions may involve a trade-off: less activity at the expense of a more radical community elsewhere. Yet, the relationship between these activity- and content-related signals from toxic online communities and their real-world harms is still fuzzy. It is unclear, for instance, whether a reduction of 50% in posting activity where each user is 10% more "toxic" is desirable or not. While such a fine-grained assessment of the consequences of a moderation intervention is out of the scope of this chapter, further study of the causal links between toxicity, user activity, and real-world harm is an important research direction to improve the quality of moderation decisions.

### 3.5.2 Implications for online platforms

Our analysis of migration dynamics highlights that community-wide moderation interventions do not happen in a vacuum. When platforms sanction an entire community, as opposed to taking user-level actions, communities may migrate *en gros* to a different platform. Platforms face difficult decisions: they need to consider the effects of community-wide sanctions not only on their own backyard but also on other online and offline spaces. Our results suggest that there may be a trade-off associated with this decision: banning a community from a mainstream platform may come at the expense of a smaller but more extreme community elsewhere. However, this takeaway should be handled with nuance since our work is limited to two communities, and the increase in toxicity was observed only in one of them.

Nevertheless, a practical implication that follows from our results is that, given that a community eventually gets banned, the time the said community was allowed to flourish in a mainstream platform may increase its potential for harm post-banning. The reasoning is simple: since the deplatforming halts community growth, and the earlier the community is banned, the fewer members a possible spin-off community would have. In that context, if banning is a commonly used practice in a given platform, it is advantageous to employ the measure proactively rather than reactively.

# 4 Banning platforms

## 4.1 Introduction

To address the rise of false information, hateful speech, and conspiracy theories, online platforms have banned influencers, communities, and even entire websites associated with content deemed harmful [301]. Such actions, broadly referred to as “deplatforming,” often result in the migration of affected users to other, more permissive platforms [70]. Because individual platforms typically lack access to other platforms’ data, the overall effectiveness of deplatforming is hard to evaluate conclusively. Reflecting this difficulty, previous work has focused on the effects of deplatforming users and communities within a single platform [115] or between a pair of platforms where one was the object of enforcement and the other was set up explicitly as a substitute [6, 100]. In general, these studies have found that deplatforming leads to a decrease in overall harmfulness, albeit an increase in the harmfulness of users who remain active [115, 6, 100]. The designs of existing analyses have, however, two important limitations. First, they do not consider the full range of websites users might migrate to after deplatforming, especially less public-facing platforms such as Telegram [270]. Second, they rely on active engagement (*e.g.*, tweeting), which may underestimate the passive consumption of harmful content (*e.g.*, views that do not lead to tweets). These limitations make it hard to obtain comprehensive and reliable estimates of the effects of deplatforming.

In this chapter, we address these limitations in the context of a high-profile deplatforming event: the suspension of the US social networking service Parler from Amazon’s Web hosting services on January 11, 2021, following the US Capitol attack. At the time, Parler was associated with conspiracy theorists and far-right extremists [7] and had around 2.3 million daily active users [301]. During the shutdown, Parler users were reported to have migrated to other fringe social media such as Rumble, Gab, and Telegram. To capture the total consumption of fringe content across various social media platforms (see Sec. 4.2 for a complete list), we analyzed two large panels from the Nielsen Company encompassing US desktop ($N_{\text{Desktop}} = 76{,}677$) and mobile ($N_{\text{Mobile}} = 36{,}028$) users from August 2020 to June 2021 (adjusted to be representative of US desktop and mobile users). This data’s user-level, longitudinal nature allows us to identify

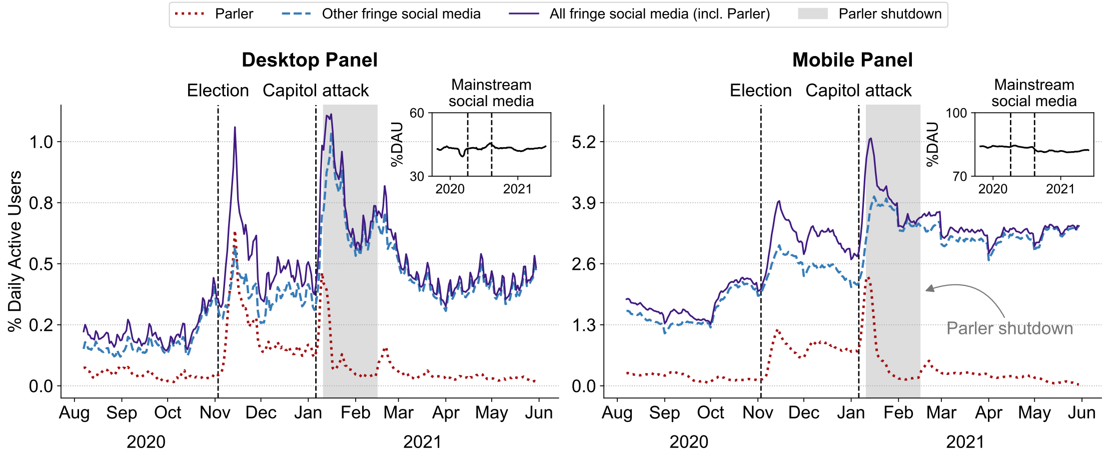


Figure 4.1: Daily percentage of US desktop (left) and mobile (right) users who were active on Parler (red), other fringe social media websites (blue), and across all fringe social media (including Parler; purple), smoothed with a 7-day moving average. Users were weighted so that the panel is representative of the US population (see Appendix A.5). Insets show user activity on mainstream social media sites (*e.g.*, Facebook, Twitter, and YouTube) for comparison. Parler's number of active users is non-zero during the shutdown period because some users still navigate toward the defunct domain. After the deplatforming of Parler, other fringe platforms (*e.g.*, Gab) prospered and the overall consumption of fringe social media increased.

the causal effect of deplatforming on Parler users. In sum, we find that Parler's temporary removal effectively stopped on-platform activity but caused a surge in activity on other fringe sites that roughly compensated for the drop-off, resulting in a negligible overall effect.

## 4.2 Materials and methods

**Data.** Our data is drawn from panels maintained by the Nielsen Company, where individuals agree to have their media and Internet consumption habits tracked in exchange for payment. Specifically, we use two such panels: a desktop panel and a mobile panel. The panels have rotating membership, meaning that tracked individuals join and leave the panel over time. In both panels, tracking software is installed on the user's devices (their computer for the desktop panel and their phone for the mobile panel).

**Labeling social media platforms.** To analyze user activity on the fringe and mainstream social media across the panels, we create lists of apps and domains and then detect panelists accessing websites/apps in the list. We consider fringe social media platforms to be broadly what Freelon et al. [70] define as "Alt-tech" platforms: social media websites that have become popular among groups espousing extreme or fringe opinions due to moderation policies that are less stringent than those of mainstream social media such as Facebook or YouTube. Specifically, we consider domains and apps associated with Locals, Gab, Rumble, BitChute, 8kun, Telegram, 4chan, MeWe, DLive, Minds, and Parler. We consider mainstream social media platforms to be those included in Pew's 2021 survey about social media use [31].

**Difference-in-differences (DiD).** We estimate the effect $\delta$ of deplatforming Parler on users' social media usage with a DiD model:

$$Y_{it} = \gamma P_t + \lambda T_i + \delta P_t T_i + \epsilon_{it}, \tag{4.1}$$

where the daily usage $Y_{it}$ of user $i$ on day $t$ is determined by whether day $t$ came after the deplatforming of Parler ($P_t \in \{0,1\}$) and whether the user was an active consumer of Parler before the intervention ($T_i \in \{0,1\}$).

We make our results more robust by estimating our DiD model using weights generated by coarsened exact matching and clustering standard errors at the user level. We matched users on sociodemographic characteristics and their pre-intervention activity (using Scott binning). We achieved exact matches for the sociodemographic features and low standardized mean differences for pre-intervention activity (0.028 for desktop; 0.011 for mobile). To obtain such matching, we discard 346 units in the mobile panel (out of 942) and 112 units in the desktop panel (out of 512), obtaining matched samples of size $N_{\text{Desktop}} = 400$ and $N_{\text{Mobile}} = 596$.

## 4.3 Results

Fig. 4.1 shows the estimated percentage of daily active users (*i.e.*, who visited the websites or mobile apps) of Parler and other fringe social media considering US desktop (left panel) and mobile (right) users (see *cf.* Appendix A.5). User activity on Parler and other fringe social media grew sharply in the aftermath of the 2020 US Presidential Election and peaked following the US Capitol attack on January 6, 2021. For example, Parler's percentage of daily active users among all panelists increased nearly six-fold between the highest value in October (desktop: 0.12%; mobile: 0.25%; Fig. 4.1 is smoothed by a 7-day moving average) and the week following the election (desktop: 0.83%; mobile: 1.54%), and increased again after January 6 (desktop: 0.96%; mobile: 3.5%). On January 11, Parler was deplatformed by Amazon, leading to a decrease in daily active users through February 25, when Parler was relaunched on another hosting service (daily activity was not precisely zero post-shutdown as some users still navigated to the defunct domain). Over this same period, however, the percentage of daily active users of fringe social media platforms other than Parler surged both on desktop and mobile to such an extent that user activity across all fringe platforms was higher during the shutdown period than before. By comparison, activity on mainstream social media (e.g., Facebook) remained roughly constant over this interval (see Fig. 4.1 insets).

This last result seems to suggest that deplatforming Parler backfired, causing users to migrate to other venues with similar (fringe) content. However, the increase could also have been driven by non-Parler users whose behavior was affected by other factors (*e.g.*, increased media attention to fringe content around January 6). In other words, aggregate data of the sort shown

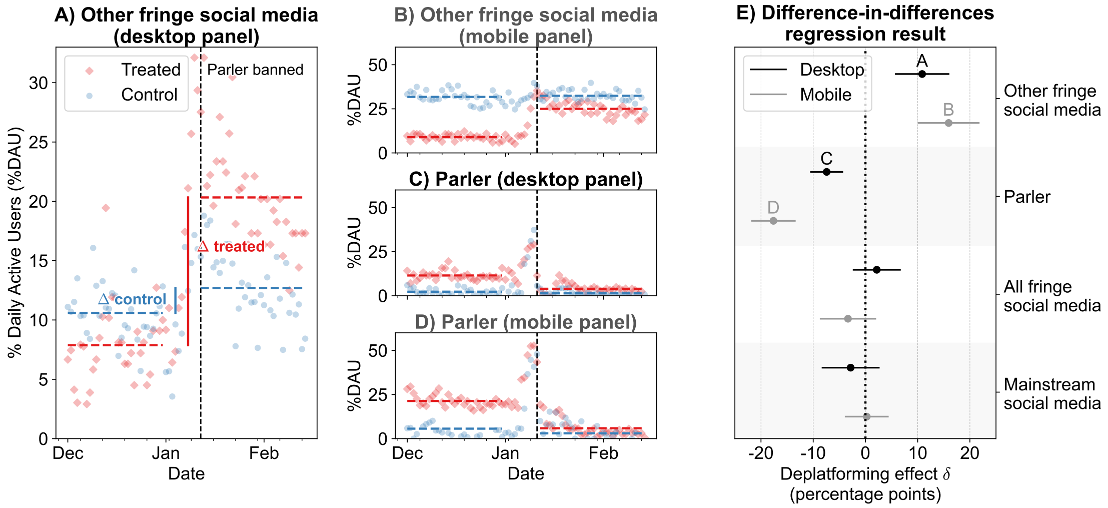


Figure 4.2: **(A–D)** Considering treated users (active on Parler pre-deplatforming; red) and control users (active on other fringe social media but not Parler, pre-deplatforming; blue), we show the percentage of daily active users across different types of platforms for the desktop and mobile panels between December 1, 2020, and February 25, 2021. **A** illustrates the difference-in-differences model (DiD). **(E)** DiD regression results for the desktop (black) and mobile (gray) panels. Coefficients from scenarios depicted in **A–D** are annotated with the corresponding letters in **E**. Error bars represent 95% confidence intervals. The DiD coefficients indicate that deplatforming increased the probability of daily activity across other fringe social media and did not significantly decrease activity on fringe social media in general.

in Fig. 4.1 cannot identify the causal effect of deplatforming on Parler users. To overcome this hurdle, we apply a difference-in-differences (DiD) approach to longitudinal user-level activity in the desktop and mobile panels. We consider the month of December 2020 as our pre-intervention period and the days between January 11, 2021, and February 25, 2021, as the post-intervention period. Our DiD approach compares two matched groups of panelists (*cf.* Sec. 4.2), illustrated in Fig. 4.2A: "treated" users (red) who spent over 3 minutes on Parler in December 2020 ($N^{\text{Treated}}_{\text{Desktop}} = 135$; $N^{\text{Treated}}_{\text{Mobile}} = 209$) and "control" users (blue) who spent over 3 minutes on other fringe social media platforms and less than 3 minutes on Parler over the same period ($N^{\text{Control}}_{\text{Desktop}} = 265$; $N^{\text{Control}}_{\text{Mobile}} = 387$). Our model calculates the difference in probability of daily activity (*i.e.*, the chance of a user visiting a website or set of websites on a given day) between the pre- and post-intervention periods for both treatment and control groups (Δtreated and Δcontrol in Fig. 4.2A). Under the identifying assumption that these differences would remain constant in the absence of the intervention (here, the deplatforming of Parler), we can estimate its causal effect through the difference in differences ($\delta = \Delta\text{treated} - \Delta\text{control}$; *cf.* Sec. 4.2 for details). Fig. 4.2A–D shows the same information for all four combinations of Parler vs. other fringe social media and desktop vs. mobile, where parallel pre-intervention trends across all four panels suggest that the identifying assumption is credible (as does placebo testing; *cf.* Appendix A.5).

Fig. 4.2E depicts the results of the DiD regression analysis (*cf.* Sec. 4.2). We observe a significant decrease in the probability of daily activity on Parler itself for both the desktop ($-7.4$ percentage points; 95% CI $[-10.4, -4.4]$) and mobile ($-17.6$; 95% CI $[-21.7, -13.5]$) panels. Consistent with Fig. 4.1, however, there was also a significant increase in the time spent by active Parler users on *other* fringe social media for both panels (desktop: 10.9 percentage points; 95% CI $[5.8, 15.9]$; mobile: 15.9; 95% CI $[10.2, 21.7]$). In sum, the net effect of deplatforming on Parler users over *all* fringe social media (Parler as well as others) was small and not statistically significant. As a sanity check, we run the same model comparing the user activity on mainstream social media websites (*e.g.*, YouTube, Twitter; *cf.* Appendix A.5), which we would not expect to be affected by the deplatforming, again finding small and statistically insignificant effects. We obtain qualitatively similar results using another outcome, the total time spent on the different sets of platforms, *cf.* Appendix A.5.

## 4.4 Discussion

These results indicate that deplatforming Parler was ineffective at reducing the consumption of the type of content that was deplatformed. It *increased* user activity on other fringe social media platforms and did not significantly decrease the total user activity on all fringe social media platforms taken together. This finding is aligned with previous research suggesting that online hate groups are resilient to uncoordinated interventions that affect only part of their ecosystem [120]. Web stakeholders may benefit from this insight by reconsidering platform-level interventions and by, *e.g.*, promoting simultaneous action against multiple fringe social media platforms or acting proactively rather than reactively during periods of political unrest. Our analysis is also relevant to researchers studying content moderation in general, as it demonstrates the value of analyzing passive engagement across multiple platforms. Measurements capturing overt engagement (*e.g.*, posts, comments) on specific platforms [115] or pairs of platforms [100, 6] tend to underestimate user activity on fringe social media after deplatforming, as users move to a variety of other platforms (some, like Telegram, less public-facing).

Our findings are limited to the deplatforming of an entire social media (Parler) during a period of exceptional political unrest. We note, however, that our scenario is similar to other instances of deplatforming; *e.g.*, 8kun, and Gab, here analyzed as part of the "other fringe social media" category, were themselves temporarily deplatformed after being associated with mass shootings [158]. Deplatforming policies applied to individual actors (*e.g.*, Twitter users, YouTube channel owners) or groups (*e.g.*, subreddits) may have different effects from those observed here. We also note that the validity of our causal results is predicated on the identifying assumptions of our DiD model. Fringe platforms are an important part of an ecosystem of fringe communities and personalities who exert influence over the media [70, 16] and large mainstream social media [298]. They are tied to hate crimes [158] and anti-democratic riots [271], and were pivotal to the so-called "infodemic" during the 2020 coronavirus pandemic [299]. In this context, we hope our findings will help inform policy responses to such websites.

# 5 Banning influencers

## 5.1 Introduction

To *deplatform* is “to remove and ban (a registered user) from a mass communication medium (such as a social networking or blogging website)” [162]. The term has gained notoriety in recent years as a roster of provocateurs, extremists, and conspiracy theorists (often from the far right) were banned from online platforms like YouTube [208], Twitter [243], and Facebook [255]. The practice has also become a contentious political issue. On the one hand, deplatforming advocates suggest that it could minimize the consequences of users’ exposure to inappropriate content on online platforms, often pointing at the negative role that platforms have played in the rise of science denial [283], political extremism [14], *etc.*. On the other hand, opponents worry about the lack of accountability and transparency in platforms’ deplatforming decisions and raise concern that this practice may silence dissenting but valid viewpoints [147].

Importantly, deplatformed individuals are reactive and, following sanctions, migrate to alternative platforms [99, 214]. Websites like Gab, Rumble, and Truth Social have capitalized on the banning of prominent influencers (as well as the people outraged by it), advertising themselves as ‘censorship-free’ and welcoming of extreme viewpoints [69]. Zuckermann and Niccoluci [301] argue these migrations help solidify the infrastructure of an “alt-tech” information ecosystem, where the visibility of extreme content is reduced, but extreme viewpoints resonate louder.

Empirical evidence on whether deplatforming works is inconclusive. Previous work has explored this question in various settings: from studying how fans of controversial figures like Alex Jones behaved on Twitter after their bans [115] to examining how entire communities re-organized themselves after being banned from Reddit [99]. This literature (described in Sec. 5.2) has found that deplatforming reduces the number of posts [99], tweets [115], or videos [202] related to the deplatformed individuals or communities, albeit at the cost of an increase in the harmfulness of the remaining content [99, 6, 169] often hosted on alt-tech platforms or standalone websites.

However, three important limitations of the existing literature threaten the validity and generalizability of its findings. Previous work typically (1) considers only a handful of deplatforming events (often a single one) [99, 115, 169, 265, 266, 221, 99, 261]; (2) considers only overt engagement traces like posts, but not passive engagement traces like impressions [221, 99, 115, 169, 265, 266]; (3) does not consider all the potential places users might migrate to [202, 221, 270, 99, 115, 169, 265, 266, 261]. Given these limitations, work capturing the heterogeneity of the effect (*i.e.*, how the intervention works against different targets) and considering the impact of deplatforming across the Web is needed to provide adequate policy guidance on the matter [99].

**Present work.** In this chapter, we present a longitudinal analysis of 165 deplatforming events concerning 101 influential users on social media, carefully sourced and annotated at both the entity level (*i.e.*, who was the influencer—a politician or a media personality?) and at the event-level (*e.g.*, why was the influencer deplatformed—for harassment or for spreading false information?). We consider online attention toward these influencers, considering two easily accessible, freely available data sources: Wikipedia pageviews and Google search interest, as measured by Google Trends [282].

We obtain our data from a comprehensive data collection pipeline. First, we use a semi-supervised approach to collect deplatforming events from Reddit posts. We match each deplatformed entity with a Google knowledge graph identifier that allows us to obtain online attention data for these entities on Wikipedia and Google Trends. Second, we ensure the completeness and correctness of the deplatforming events for the entities considered by retrieving news sources associated with each entity and individually attributing a news piece confirming each platforming event. Finally, we manually label each deplatforming event, specifying, *e.g.*, the reason behind the intervention. We make our dataset, the largest collection of deplatforming events (and associated attention traces) to date, and the code to reproduce our analyses publicly available.[I]

We use a stacked difference-in-differences regression to disentangle the causal effects of deplatforming from the reason behind the intervention, following Cengiz et al. [29]. This approach allows us to compare treated units (deplatformed influencers) with 'yet-to-be-treated' control units (influencers who will be deplatformed in the future) and to identify the causal effect of deplatforming under the parallel-trends assumption, *i.e.*, that in the absence of treatment, the differences between the two groups would remain constant. Further, we address the limitations of previous research as we study passive engagement traces that are agnostic to specific social media platforms and consider orders of magnitude more deplatforming events than past research. This has two implications. First, we can better answer the question: does deplatforming reduce online attention toward influencers? Second, we can study whether different individuals are affected differently; *e.g.*, do influencers banned for different reasons respond differently to being deplatformed?

[I] https://github.com/epfl-dlab/deplatforming_influencers

**Summary of results.** We find that deplatforming reduces online attention toward influencers. Specifically, after 12 months, we estimate that online attention toward deplatformed influencers is reduced by −63% (95% CI [−75%,−46%]) on Google and by −43% (95% CI [−57%,−24%]) on Wikipedia. We also find that the effect of deplatforming varies according to the characteristics of the deplatformed entity. For example, sorting entities into two groups according to the attention they received 12 months before deplatforming (top 1/3rd of influencers with highest attention *vs.* bottom 2/3rds), we find that the effect of deplatforming for high-attention entities was roughly 60% lower relative to low-attention entities (−58%; 95% CI [−73%,−38%]). Studying the heterogeneity of deplatforming, we also find that both permanent and temporary deplatforming significantly reduced online attention and that users banned for spreading misinformation seem to have their subsequent online attention reduced further than those banned for other reasons.

**Implications.** Overall, we contribute to an emerging literature mapping the effectiveness of content moderation interventions and helping to guide platform governance practices (*e.g.*, [246, 119, 43, 59]). Our results offer empirical evidence that sanctioning influencers significantly reduces online attention directed at them. Further, we find similar effects for both temporary and permanent deplatforming. Therefore, platforms and other stakeholders should consider temporary bans to prevent harm caused by influencers' online presence, particularly those receiving widespread online attention.

## 5.2 Background

**Reddit.** Reddit is a community-oriented social media platform centered around "subreddits" where users can contribute with posts and comments. We use Reddit to obtain information about deplatforming events because the platform is commonly used to discuss internet-related events [138, 215]. Since Reddit users are mostly Western and English-speaking [215], this choice biases our data to deplatforming events relevant to English-speaking countries. Yet, it is worth noting that content moderation, in general, is biased toward Western and English-speaking countries [238].

**Google Trends.** Google Trends is a freely accessible tool that allows anyone to measure the popularity of search queries on Google.com, the world's largest search engine. Queries can be specified as plain text or as identifiers from Google's knowledge graph [20], *e.g.*, Alex Jones corresponds to the identifier /m/01_6j_. The research community has extensively used Google Trends in diverse scenarios, from nowcasting economic indicators [41] to estimating the prevalence of diseases [179]. One issue with using Google Trends is that it yields rounded and normalized results, i.e., the output is quantized to integer precision and scaled such that the maximum value of every time series equals 100. To address this issue, we leverage Google Trends Anchor Bank (G-TAB) [282], a method for calibrating Google Trends time series such that time series for an arbitrary number of Google queries can be expressed on a common scale with high resolution.

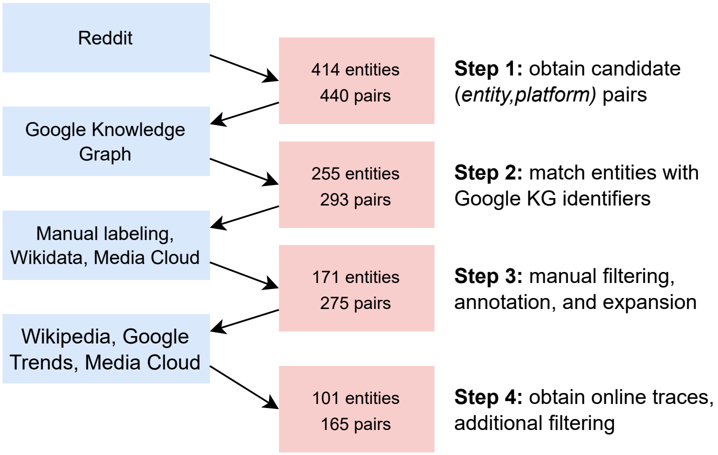

Reddit
Google Knowledge Graph
Manual labeling, Wikidata, Media Cloud
Wikipedia, Google Trends, Media Cloud
414 entities 440 pairs
255 entities 293 pairs
171 entities 275 pairs
101 entities 165 pairs
Step 1: obtain candidate (entity,platform) pairs
Step 2: match entities with Google KG identifiers
Step 3: manual filtering, annotation, and expansion
Step 4: obtain online traces, additional filtering

Figure 5.1: Overview of our data collection and curation pipeline. Starting from Reddit, we obtain 440 ⟨entity, platform⟩ pairs, each corresponding to a deplatforming event (*Step 1*). Then, we link entities to Google Knowledge Graph identifiers, which are subsequently linked to GKG-ids (*Step 2*) and manually filter, annotate, and expand the data (*Step 3*). Last, we obtain online traces corresponding to each entity from Wikipedia, Google Trends, and Media Cloud, performing additional filtering to ascertain the quality of the online traces (*Step 4*).

**Wikipedia.** Wikipedia is the world's largest encyclopedia and one of the most visited sites on the Web. People turn to Wikipedia for various information needs, from keeping up with current events to randomly surfing across the Web due to boredom [241]. The Wikimedia Foundation makes aggregate statistics of Wikipedia page views publicly available, and the research community has largely used this data as a metric of public interest: from measuring public interest in biodiversity and conservation [168] to forecasting election results [292]. It is worth noting that viewing Google and Wikipedia as completely unrelated data sources is ill-informed. Previous research has shown that both platforms are interdependent, with the search engine heavily relying on Wikipedia to provide factual content [273, 159] and the online encyclopedia receiving a substantial amount of its visitors from Google [193, 159].

**Relationship between prior and present work.** Our work differs from previous research in the nature and the scale of online traces analyzed. Using Google Trends and Wikipedia Pageviews, publicly available signals not typically used in content moderation research, we study the effect of deplatforming in online traces capturing *passive engagement* across over a hundred deplatforming events. This allows us to assess the causal impact of deplatforming holistically but also to study the heterogeneity of the effect of the intervention (*i.e.*, when does deplatforming work best?). Also, we focus on the deplatforming of influencers, whereas much of previous research has studied the deplatforming of fringe communities [222, 221, 34, 266, 265] or even of entire social media platforms [99].

## 5.3 Materials and Methods

We summarize our data collection methodology in Fig. 5.1 and detail the steps below.

#### Step 1: Obtaining candidate ⟨entity, platform⟩ pairs

To obtain information about "who was deplatformed where," we used a semi-automatic approach using Reddit data. In Reddit, a large community-oriented social network, it is common for users to share news pieces and commentary on internet-related events, including when individuals, media outlets, or even entire online communities are banned from social media platforms. Considering all Reddit posts until the end of August 2021, we extracted ⟨entity, platform⟩ pairs from the title of Reddit posts using a pattern-based bootstrapping methodology inspired by Quootstrap [187]. Starting from the seed patterns "⟨entity⟩ (was | were | has been | ∅) banned from ⟨platform⟩," we iterated these two phases:

1. **Pair extraction step:** Extract ⟨entity, platform⟩ pairs in the data that match previously discovered patterns.

2. **Pattern extraction step:** Discover new patterns expressing the previously extracted ⟨entity, platform⟩ pairs.

At the end of phase #2, we manually filtered the set of extracted ⟨entity, platform⟩ pairs as well as the set of extracted patterns, removing incorrect pairs, as well as patterns that occur infrequently (as, in practice, we found that these patterns generalize poorly across iterations).[II] Note that here, we considered 'incorrect' titles that did not indicate a deplatforming event, e.g., "Activists push for Twitter to Ban Donald Trump." We repeated this iterative process twice, obtaining 440 ⟨entity, platform⟩ pairs involving 414 entities and 18 distinct platforms. The final list of patterns used is listed in Table A.3 at the end of the chapter.

#### Step 2: Match entities with GKG-ids

We then managed to link 255 of the extracted entities to Google Knowledge Graph identifiers (henceforth referred to as GKG-ids; *e.g.*, Alex Jones corresponds to the identifier /m/01_6j_[III]) as well as Wikidata ids (Wiki-ids; *e.g.*, Alex Jones corresponds to the identifier Q319121). These identifiers helped us link entities in our dataset with their corresponding Google Trends and Wikipedia page view data, respectively.

[II] More precisely, we removed patterns that occurred fewer than 14 times. This threshold was determined by analyzing the distribution of pattern occurrences.

[III] Note that GKG-ids are a superset of the identifiers of Freebase, a collaborative knowledge base acquired by Google in 2010. For the relationship between Freebase and Wikipedia, refer to Tanon et al. [190]

Table 5.1: Examples of sources obtained for ⟨entity, platform⟩ pairs.

| Date | Entity | Platform | Source |
|---|---|---|---|
| 2018-04-16 | Richard B. Spencer | Facebook | BBC[a] |
| 2021-06-04 | Yair Netanyahu | Twitter | Times of Israel[b] |
| 2020-08-07 | Tommy Robinson | Instagram | Daily Caller[c] |

[a] https://www.bbc.com/news/technology-43784982

[b] https://www.timesofisrael.com/yair-netanyahu-temporarily-blocked-from-social-media-for-protest-call

[c] http://dailycaller.com/2018/08/07/instagram-tommy-robinson-ban

### Step 3: Manual filtering, annotation, and expansion

Next, we filtered, annotated, and expanded the data with the aid of Wikidata [276], Media Cloud [210], Wikipedia, and the search functionalities of Reddit and Google. This resulted in a dataset with 171 entities and 275 deplatforming events.

**Ensuring the completeness of deplatforming events.** We search for additional deplatforming events for the entities already in the dataset using Media Cloud, an open-source collection of news on the Web [210]. For each entity, we conducted a search using all curated patterns obtained in Step 1.[IV] With this procedure, we managed to retrieve 8036 stories about 124 of the entities. For each entity, we manually inspected the news headlines (opening the URL), looking for different deplatforming events that were not covered in the original dataset. We found nine additional deplatforming events corresponding to eight distinct entities. In many cases, those events were associated with entities that were banned from several distinct platforms, *e.g.*, Alex Jones (banned from six platforms) and David Duke (banned from two platforms). Overall, this step worked as a sanity check for the internal completeness of the data (*i.e.*, for the entities considered, are all meaningful deplatforming instances represented?) but also allowed us to further refine our data through the addition of the nine deplatforming events we found. We acknowledge that despite this rigorous effort, some salient deplatforming events might have escaped our data collection or might not have appeared in any news media stories at all. Therefore, we present our approach as our best effort at ensuring the completeness of prominent deplatforming events about relevant entities.

**Ensuring the correctness of deplatforming events.** One author of this chapter manually checked whether each ⟨entity, platform⟩ pair was correct (*i.e.*, was the entity banned?). They checked the links shared on Reddit for each pair to find a reliable source confirming the deplatforming event. They also searched for the entity on Google News and Wikipedia where necessary. Through this process, which took around 10 hours, they ensured that all deplatforming events were correct and attributed a deplatforming date to each event. We provide examples of sources and dates for different ⟨entity, platform⟩ pairs in Table 5.1, e.g., per BBC, Richard B. Spencer was banned from Facebook on the 16th of April, 2018.

[IV] For example, for Alex Jones, we used the query *"bans Alex Jones" OR "suspends Alex Jones" OR "suspended Alex Jones" OR "banning Alex Jones" OR "Alex Jones locked out of" OR "Alex Jones blocked on" (...) has been banned from".*

Table 5.2: Entity-level labels assigned to person entities along with examples. Size here corresponds to the number of deplatforming events with each label in the dataset obtained at the end of *Step 3*.

| **Label** (size) | **Examples** |
|---|---|
| Politician ($n = 31$) | David Duke, Donald Trump, Marjorie Taylor Greene, Ron Paul |
| Internet personality ($n = 51$) | Blaire White Carl Benjamin Paul Joseph Watson Stefan Molyneux |
| Media personality ($n = 65$) | Candace Owens, Graham Linehan, Katie Hopkins, Tila Tequila |
| Fringe movement ($n = 83$) | James Allsup, Gavin McInnes, Nicholas J. Fuentes, David Duke |

**Entity-level labels.** We used Wikidata and retrieved whether each entity is an instance of either a “Person” (Human Q5; 146 entities in our data) or an “Organization” (Organization Q43229; 25 entities in our data). We also considered organizations websites (Q35127; *e.g.*, SciHub) and political movements (Q2738074; *e.g.* The Boogaloo movement) aligned with Wikidata’s definition [285] of an organization as a “social entity established to meet needs or pursue goals.” Then we annotated, for each *person* in the dataset, whether they are 1) a politician; 2) a media personality; 3) an Internet personality; or 4) associated with fringe movements and/or ideologies. Considering the first sentences in each person’s article, these labels were assigned, which, according to Wikipedia’s Style Manual [286], should tell the nonspecialist reader what or who the subject is. Note that labels are non-exclusive, e.g., Alex Jones has both the “media personality” and the “fringe movement” labels. To validate our labels, two authors of this chapter manually verified the four labels for 39 entities, reading their Wikipedia entries of each, reaching a consensus through discussion in cases of disagreement. They agreed on 96.7% (151/156) of the labels assigned in this fashion, which we considered adequate for subsequent analyses (done with the labels assigned by a single coder). Table 5.2 depicts the (non-exclusive) entity-level labels assigned to person entities.

**Event-level labels.** We developed a taxonomy for the reason an entity was suspended from the online platforms by examining Meta’s community standards [164], Twitter Rules [268], and the headlines of all news sources obtained. Our final taxonomy, shown in Table 5.3 along with examples and relevant parts of the policy documents, consists of three categories, selected based on their prominence in platform policies and frequent occurrence in deplatforming-related news: A) *Hate, Harassment, Incitement to Violence*; B) *Misinformation, Platform Manipulation*; and C) *Other/Unknown*, a category encompassing less common reasons (*e.g.*, lewd content, copyright infringement), as well as cases where the reason for the suspension was not clear. Further, we classified each deplatforming event as either “permanent” or “temporary.” We considered the bans *permanent* when the social media pages for the entity in question were online at the time of annotation (October 2021) and temporary otherwise. Note this was before the wave of “re-platforming” events started in late 2022 on Twitter after Elon Musk’s acquisition of the platform. To validate our labels, another author manually verified the two labels for 30 entities, reaching 96.5% agreement (56/58) with the original labels, which we considered adequate for subsequent analyses (again, with the labels from a single coder).

Table 5.3: Categorization of deplatforming reasons used in this chapter. We consider three categories (first column), pointing associated policies (second column), and example headlines (third column). Note that the associated policies can be found in Meta and Twitter respective websites [268, 164] and that the headlines given as examples are from the sources extracted for each deplatforming instance. Size here corresponds to the number of deplatforming events with each label in the dataset obtained at the end of *Step 3*.

| **Reason** (size) | **Associated Policies** | **Example Headlines** |
|---|---|---|
| Hate, Harassment, Incitement to Violence<br>($n = 166$) | **Meta**: Hate Speech, Violence and Incitement, and Dangerous Individuals and Organizations Policies<br>**Twitter**: Abusive Behavior, Hateful Conduct and Violent Organizations Policies | Twitter suspends Azealia Banks for transphobic tweets<br>YouTube removes 3 prominent white supremacist channels |
| Misinformation and Platform Manipulation<br>($n = 36$) | **Meta**: False News, Manipulated Media and Inauthentic Behavior Policy.<br>**Twitter**: Platform manipulation and spam, Civic integrity, Synthetic and manipulated media policy, COVID-19 misleading information policy | Charlie Kirk: Trump supporter has Twitter account locked for spreading misinformation about mail-in voting<br>Facebook suspends Cambridge Analytica (...) |
| Other/Unknown<br>($n = 73$) | This label was assigned to unclear cases or those that did not fit the above labels. | Courtney Stodden, 17, Banned from Facebook for “Sexy” Shots<br>Facebook blocks "Atheist Republic" on government directive |

**A dataset of deplatforming events.** We obtain a carefully curated, comprehensive dataset of 275 deplatforming events between 2010 and August 2021. Subsequently, we filter this data to answer the research questions at hand, e.g., keeping only entities for which we obtain meaningful online traces. We also make this intermediate data available. The research community can use this data to explore a wider variety of relevant questions about the dynamics of social media deplatforming, sanctioned influencers, and the linkages between social media and news media.

### Step 4: Obtain online traces, additional filtering

Finally, we retrieved online attention data from Google Trends and Wikipedia and further filtered and processed the data to answer our specific research questions.

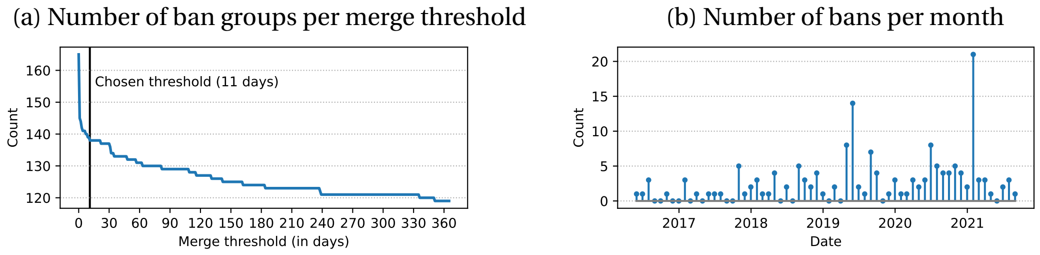
**Data crawling.** To obtain Wikipedia pageviews, we used a well-established API[V] offered by Wikimedia. To obtain search interest, we use Google Trends Anchor Bank (G-TAB), a method for calibrating Google Trends data that allows us to obtain queries on a universal scale and to minimize errors stemming from Google's rounding to the nearest integer [282]. We considered only data between 1 April 2015 and 1 September 2022. As discussed in the following paragraph, we only consider bans after 2016 (the vast majority), and these dates correspond to the period from the first [last] ban minus [plus] 1 year. We search for entities using the GKG-ids (which are matched one-to-one to every Wikipedia page).

**Additional filtering.** Starting from the 171 entities from *Step 3*, we excluded 23 entities that we labeled as Organizations (as we focus on individuals) and 35 entities for which data was unavailable (or noisy[VI]) for one of the two considered data sources. Further, to simplify our data analysis, we filtered eight entities that were banned before 2016 and 4 entities that were banned from platforms other than YouTube, Facebook, Instagram, or Twitter.

**Ban groups.** Several platforms often ban individuals within a couple of days for the same reason, *e.g.*, on 6 August, 2018, YouTube, Facebook, and Instagram banned Alex Jones simultaneously for his promotion of conspiracy theories [275]. Then, exactly one month later, on 6 September, Twitter also banned Jones due to his behavior on the platform [45]. To simplify our analyses, we introduce the concept of a "ban group," a series of highly connected bans that are close in time. For example, we determined that Jones's first three bans (YouTube, Facebook, and Instagram) should be in the same ban group, while the Twitter ban should not. Using the elbow method (*cf.* Fig. 5.2a), we determined that all bans imposed within 11 days are in the same "ban group." Manually inspecting the ban groups, we find that all bans grouped this way are highly related, *i.e.*, relating to the same incident.

[V] Wikimedia's pageview API, https://wikitech.wikimedia.org/wiki/Analytics/AQS/Pageviews

[VI] Even using G-TAB, the signal provided by Google for entities with low attention remains unreliable, as Google adds random noise to the queries. This noise is not noteworthy for most queries, but for entities that receive low attention, it causes substantial fluctuations in value. In this context, we filtered entities with average attention smaller than 0.001 units.



(a) Number of ban groups per merge threshold

(b) Number of bans per month

(c) Event-level labels

| **Label** | **Category** | *n* |
|---|---|---|
| Reason | Hate/Harassment/Incitement | 116 |
| | Manipulation/Misinformation | 23 |
| | Other/Unknown | 26 |
| Temporary | True | 55 |
| | False | 110 |

(d) Entity-level labels

| **Label** | **Category** | *n* |
|---|---|---|
| Type | Politician | 28 |
| | Media personality | 33 |
| | Internet personality | 40 |
| | Fringe movements | 68 |

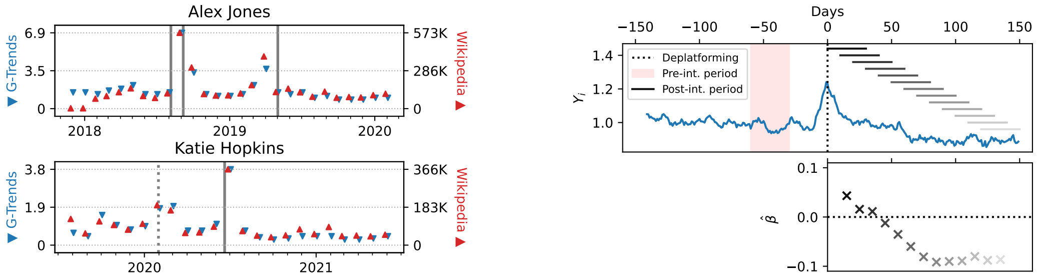
Figure 5.2: Dataset details. *(a)* We show how the number of ban groups varies when changing the *"merge threshold," i.e.*, the maximum number of days between two bans in the same ban group. We also show the chosen threshold. *(b)* We depict the number of bans per month in the final dataset. *(c)/(d)* Number of deplatforming events associated with each entity- and event-level labels in the final dataset.

**A dataset of deplatforming events and online attention data.** In sum, at the end of this step, we obtain a dataset containing 101 person entities involved in 165 deplatforming events between January 2016 and September 2021, each linked to online attention traces from two different sources. Online traces are almost complete, with less than 5% of data points missing around the 12 months of each deplatforming event considered (due to them being unavailable on either Google Trends or Wikipedia). We plot the distribution of bans over the considered period in Fig. 5.2a and provide the number of entity- and event-level labels in Table 5.2c and Table 5.2d, respectively.

## 5.4 Results

### 5.4.1 Changes in Online Attention Following Deplatforming

We show the data collected for two entities, Alex Jones and Katie Hopkins, in Fig. 5.3a, illustrating the challenges associated with estimating the causal effect of deplatforming on online attention. Deplatforming is often preceded by a controversial event that boosts online attention toward that entity, and the intervention itself triggers subsequent attention toward the entity. In other words, the controversial event may explain an increase in online attention without a causal effect of deplatforming on online attention. This can be seen in the figure showing the sharp increases in online attention right before and right after the deplatforming



(a)

(b)

Figure 5.3: Illustration of our descriptive approach. (a) Monthly online attention traces for Alex Jones and Katie Hopkins (Google Trends in blue, left $y$-axis; Wikipedia pageviews in red, right $y$-axis). Vertical lines correspond to "ban groups," *e.g.*, in August 2018, several platforms banned Jones,and all those events are merged into the same ban group. Dashed lines indicate that the suspension was temporary, *e.g.*, Katie Hopkins was suspended temporarily from Twitter in late January 2020. (b) Synthetic data depicting our fixed-effects approach, considering the online attention received by an entity $i$ ($Y_i$, $y$-axis top plot) deplatformed on day 0 ($x$-axis), we estimate the change in online attention post-deplatforming ($\beta$) as the average difference in attention ($Y_i$) between the post- and pre-intervention periods. When estimating Eq. (5.1), we vary the post-intervention period considered, as indicated by horizontal lines, to obtain time-varying estimates (bottom).

event occurred.

Thus, before dwelling upon the causal question, we first focus on describing the changes using extensions of this simple fixed-effects model:

$$Y_{i,t} = \beta \cdot 1\{D_{i,t} = 1\} + \alpha_i + u_{i,t}, \tag{5.1}$$

where $Y_{i,t}$ is the natural logarithm of the attention toward entity $i$ at time $t$, $1\{D_{i,t} = 1\}$ is an indicator that equals 1 if $i$ was already deplatformed at time $t$ (and 0 otherwise), $\alpha_i$ is an entity-level fixed effect, and $u_{i,j}$ is the error term. Note that when estimating the model with least squares, $\beta$ is the average entity-level change in attention pre- *vs.* post-deplatforming.

Importantly, we do not fit this model using all of our data, but instead, consider a fixed pre-intervention period and fit the regression model repeatedly considering different post-intervention periods (Gelman's "secret weapon" [73]). We illustrate this approach with synthetic data in Fig. 5.3b. Given a deplatforming event that takes place on day 0, we consider the attention the entity received between days −60 and −30 as the pre-intervention period[VII] and estimate the model repeatedly for different post-intervention periods, indicated by horizontal lines in grayscale. Fig. 5.3b (bottom) shows the estimated coefficients $\hat{\beta}$ using different post-intervention periods. This approach helps us estimate the effects of deplatforming at varying temporal distances.

[VII] We do not use the period right before the deplatforming event as a pre-intervention to prevent considering attention traces contaminated by a controversial event that has led to the entity's deplatforming.

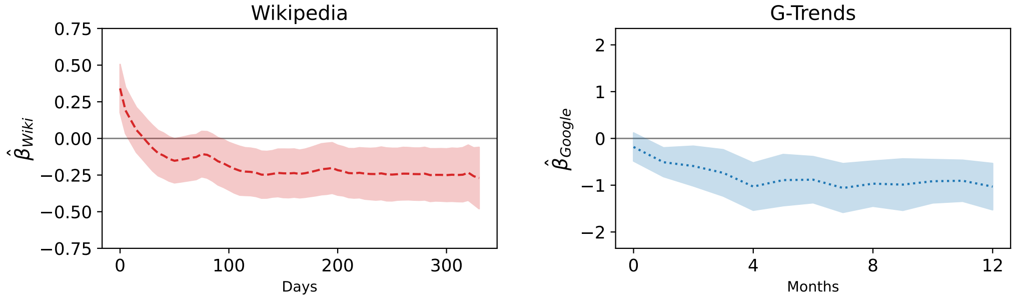


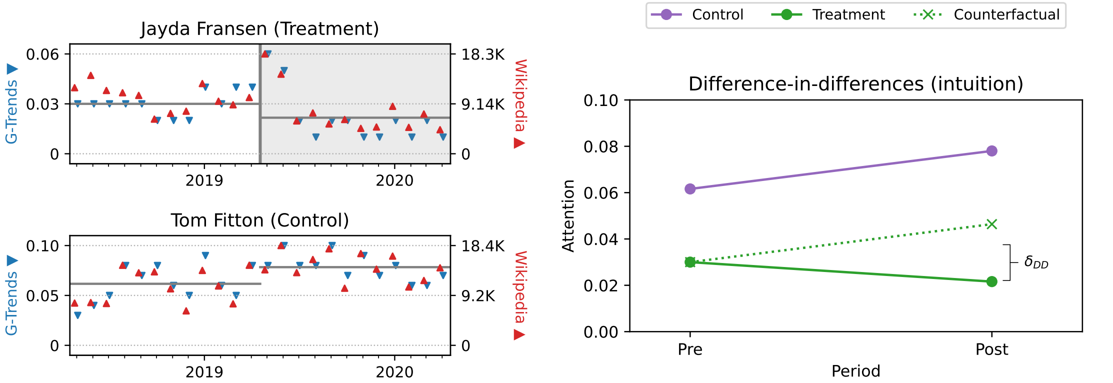
Figure 5.4: Results from fixed-effects model. We show the estimated parameters of our fixed-effects models varying the start date of the 30-day post-intervention period considered for Wikipedia pageviews (left) and Google Trends (right). Specifically, we show the estimated $\beta$ for the model depicted in Eq.(5.1), which captures the average difference in attention pre- *vs.* post-deplatforming for the last deplatforming event of each entity in our datasets.

One practical challenge with naively using the model in Eq. (5.1) to estimate changes in online attention post-deplatforming is that, as shown in Fig. 5.3a, entities can be deplatformed multiple times. Thus, the post-deplatforming periods used in estimating the model can coincide with subsequent deplatforming events. To prevent this issue, we estimate the model using the last deplatforming event and discarding entities who happened to have another deplatforming event in the pre-intervention period.[VIII] Further, note that observations within each entity are likely to be serially correlated, which makes the estimates of Var($\hat{\beta}$) inconsistent under the standard OLS estimator. We hence estimate standard errors using robust standard errors clustered at the entity level [26] to prevent this, obtaining consistent, albeit conservative, estimates of Var($\hat{\beta}$).

We depict the estimates of $\beta$ in Fig. 5.4, considering the logarithmic number of Wikipedia pageviews (left) and the logarithmic G-TAB estimate (right) as the outcomes $Y$. We use the setup illustrated in Fig. 5.3b, considering days −60 to −30 (relative to the last deplatforming event of each entity, *i.e.*, day 0, when deplatforming happened) as the pre-intervention period, and varying the post-intervention period in the 360 days (or 12 months) that followed the last deplatforming event, always considering a 30-day window. For instance, when considering the estimate corresponding to day 100 in the plot, we are thus considering the model fit with attention data from dates $t \in [-60, -30) \cup [100, 130)$.[IX] Note that taking logarithms of the outcome prior to estimating the model in Eq. (5.1) diminishes the influence of entities that receive a lot of attention and allows us to interpret the estimated coefficients $\hat{\beta}$ as multiplicative; *e.g.*, if $\hat{\beta} = 0.1$, this means that deplatforming led to roughly a 10% increase in attention received ($e^{0.1} - 1 \simeq 10.5\%$).

[VIII] We also estimated the model using only the *first* deplatforming event. This analysis yields similar results but is much noisier, as the number of units used to estimate the model varies with the period considered.

[IX] For Google Trends, data is aggregated in the monthly granularity, and thus we use month -1 as the pre-deplatforming period and estimate the model with data for each month after the moderation decision took place.



(a) (b)

Figure 5.5: Illustration of our difference-in-differences (DiD) approach. We compare deplatformed and yet-to-be-deplatformed units, *e.g.*, Tom Fitton *vs.* Jayda Fransen in (a), and estimate the effect of deplatforming with the difference-in-differences estimator (b). Note that in (a), horizontal lines represent the average attention received by the entities pre- and post-deplatforming. These values are also represented in (b) with circles. The DiD estimator $\delta_{DD}$ constructs a counterfactual estimate of how the outcome of the treatment group would have progressed had the group not received the treatment (× in the plot), contrasting it with how the outcome actually changed between the pre and post periods.

Analyzing Fig. 5.4, we can conclude that deplatforming was correlated with significant decreases in attention considering both Google Trends and Wikipedia page views. For Wikipedia, in particular, the attention received by deplatformed entities is initially positive compared with the pre-intervention period, becoming negative as time passes and remaining negative long after.

### 5.4.2 The causal effect of deplatforming on online attention

In this section, we estimate a lower bound of the causal effect of deplatforming on online attention. One notable shortcoming of the “pre *vs.* post” analyses performed in Sec. 5.4.1 is that they do not account for potential trends in outcomes. If online attention toward all entities decreases, we could find a negative decrease in post-deplatforming attention even without an effect. We address this issue with a difference-in-differences (DiD) approach (further explained below). The other substantial shortcoming is that, as mentioned in Sec. 5.4.1, the event that caused deplatforming may also impact online attention, acting as a confounder. While the DiD methodology used here does not, by itself, solve this second issue, we argue that, as this confounding event is only likely to *increase* the subsequent online attention, this makes the estimate obtained by the DiD estimator a lower bound (see Sec. 5.5 for details).

Our difference-in-differences approach compares deplatformed units and yet-to-be deplatformed units,e.g., in Fig. 5.5a, we show the online attention received by Jayda Fransen and Tom Fitton; given that Tom was deplatformed many months after Jayda, he will be used as a "control" when considering Jayda's deplatforming. The gist of the identification strategy is the parallel-trends assumption: if, in the absence of treatment, the difference between the "treatment" and "control" groups is constant over time, then we can estimate the causal impact of deplatforming with the difference-in-differences estimator. We illustrate this in Fig. 5.5b, considering Jayda and Tom's time series. We calculate the pre *vs.* post difference in online attention for the control group and assume that the treatment group would behave similarly, obtaining a counterfactual estimate for the treatment group, indicated by an orange cross and a dashed orange line. Then, we estimate the causal effect as the difference between how much the treatment group actually changed and the counterfactual estimate ($\hat{\delta}_{DD}$ in Fig. 5.5b).

While the intuition behind parallel trends is the clearest in the simple "2 by 2" difference-in-difference illustrated in Fig. 5.5b, observational studies often opt for a "leads-and-lags" difference-in-differences specification [29]:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell < -3} \delta_\ell D_{i,t}^{\ell} + \sum_{\ell \geq -2} \delta_\ell D_{i,t}^{\ell} + \epsilon_{i,t}, \tag{5.2}$$

where $Y_{i,t}$ is the outcome associated with unit $i$ at time $t$, $\alpha_i$ and $\lambda_t$ are unit and time fixed effects, and $D_{i,t}^{\ell}$ is an indicator variable for unit $i$ being $\ell$ periods away from when it received the treatment. The advantage here is that the $\delta_\ell$ coefficients obtained by estimating this model capture the causal effect $\ell$ months after the intervention, enabling the analysis of how the effect progresses with time. Moreover, we also obtain coefficients associated with periods *prior* to the intervention, *e.g.*, $\delta_{-2}$. These coefficients allow for "pre-testing," i.e., checking if the parallel trends assumption holds for periods before the intervention, like $\ell = -2$. This sanity check increases the credibility of the parallel trends assumption for the periods when the intervention actually happened (which is impossible to test).

However, this estimator is biased when we have multiple treatment periods because, under the hood, the fixed effect estimator erroneously compares newly treated units relative to already treated units [80]. To address that, we use the "stacked" difference-in-differences approach proposed by Cengiz et al. [29]. Their approach consists of considering an event window (here, we use 12 months before and after the ban) and then creating "sub-datasets," each containing a treated unit and control units that have not yet been treated during the event window. Using the same "leads-and-lags" estimator in this stacked dataset yields an unbiased estimate, as the "forbidden comparisons" are pruned from the data.

**Effect of deplatforming.** We show the estimated effects of deplatforming in Fig. 5.6 for both Wikipedia pageviews (left) and Google Trends (right). Again, we consider logarithmic outcomes in the difference-in-differences model to interpret the effects multiplicatively. However, for simplicity's sake, here we show the percentage changes on the $y$-axis (*i.e.*, $e^{\hat{\beta}} - 1$ for a logarithmic effect $\hat{\beta}$) rather than the logarithmic effects.

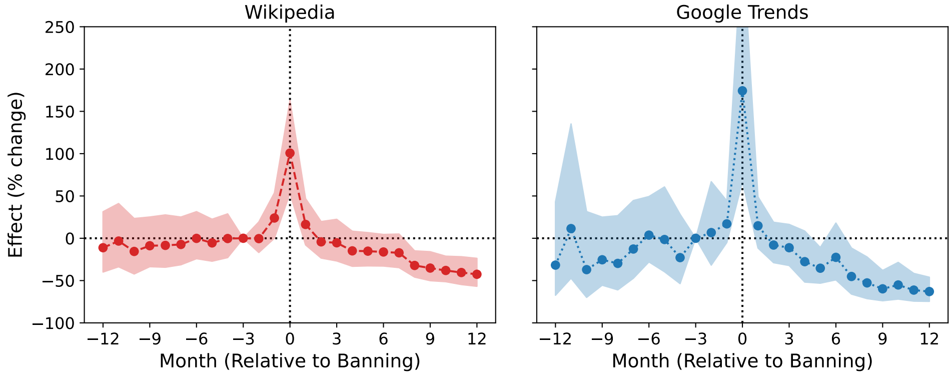


Figure 5.6: Results from our difference-in-differences (DiD) approach. We show the estimated reduction in online attention received by influencers due to deplatforming on Wikipedia (left) and Google Trends (right). We show the effect estimated across different months. Overall, these results indicate that deplatforming reduces online attention toward influencers.

Results with this quasi-experimental methodology are very similar to those obtained with the simple fixed-effects model. There is a sudden increase in online attention in the month when entities are deplatformed (Month = 0 in Fig. 5.6), followed by a continuous decrease in the attention that treated entities receive relative to control units. After 12 months, we estimate that online attention towards deplatformed influencers is reduced by −63% (95% CI [−75%,−46%]) on Google and by −43% (95% CI [−57%,−24%]) on Wikipedia.

We also show the estimated effect in the 12 months *before* deplatforming, which allows us to assess the plausibility of the parallel trends hypothesis. We find that trends remain parallel until right before the deplatforming event (Month = −1), when online attention spikes, suggesting that the event or events that triggered deplatforming also boosted online attention (see Section 5.5 for further discussion).

**Heterogeneity of the effect.** We adapt our difference-in-differences methodology to estimate whether the effect is heterogeneous across various dimensions. In particular, we decompose the effect by adding interactions with other variables and then analyze the coefficients associated with these interactions. For example, to measure whether permanent bans decrease subsequent online attention more than temporary bans, we consider the modified leads-and-lags specification:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{\ell<-2} \delta_\ell D^\ell_{i,t} + \sum_{\ell<-2} \gamma_\ell D^\ell_{i,t} P_i + \sum_{\ell\geq 0} \delta_\ell D^\ell_i + \sum_{\ell\geq 0} \gamma_\ell D^\ell_{i,t} P_i + \epsilon_{i,t}, \tag{5.3}$$

where $P_i$ is an indicator variable that equals 1 only when the ban is permanent. This allows us to measure if the effect is significantly smaller or bigger in this subset of deplatforming events by analyzing the coefficients $\gamma$. (Note that the effect on influencers banned temporarily is captured by $\delta$, whereas the effect on the permanently banned is captured by $\delta + \gamma$.)

Table 5.4: Heterogeneity of the effect. We study how the effect of deplatforming varies depending on the characteristics of the deplatforming event (e.g., whether it is temporary) and on the characteristics of the influencer being deplatformed (e.g., whether it is a politician). Note that this is not the *overall effect* for deplatforming events with these characteristics, but rather, the extent to which the effect of deplatforming events associated with these characteristics differs from the effect of deplatforming events that are not.

| Model | Variable | Source | Effect | *p*-value |
|---|---|---|---|---|
| 1 | Temporary | G-Trends | 13.2% 95% CI [-45.5%, 135.1%] | 0.739 |
| | | Wikipedia | -5.3% 95% CI [-36.5%, 41.3%] | 0.791 |
| 2 | High attention | G-Trends | -58.8% 95% CI [-72.6%, -37.9%] | <0.001 |
| | | Wikipedia | -33.7% 95% CI [-48.2%, -15.1%] | 0.001 |
| 3 | Hate | G-Trends | -19.5% 95% CI [-57.2%, 51.4%] | 0.501 |
| | | Wikipedia | -20.8% 95% CI [-46.6%, 17.5%] | 0.247 |
| | Manipulation | G-Trends | -59.8% 95% CI [-82.4%, -7.7%] | 0.032 |
| | | Wikipedia | -48.7% 95% CI [-67.5%, -19.2%] | 0.004 |
| 4 | Politician | G-Trends | -5.5% 95% CI [-63.5%, 144.7%] | 0.907 |
| | | Wikipedia | -0.9% 95% CI [-44.5%, 76.7%] | 0.975 |
| | Media pers. | G-Trends | 32.2% 95% CI [-44.7%, 215.8%] | 0.530 |
| | | Wikipedia | 22.8% 95% CI [-24.1%, 98.5%] | 0.403 |
| | Internet pers. | G-Trends | 25.1% 95% CI [-45.1%, 184.8%] | 0.594 |
| | | Wikipedia | 35.6% 95% CI [-24.4%, 143.2%] | 0.308 |
| | Fringe mov. | G-Trends | -38.6% 95% CI [-71.8%, 33.8%] | 0.220 |
| | | Wikipedia | -1.4% 95% CI [-34.0%, 47.2%] | 0.945 |

In this fashion, we consider four different models, each including different interactions. These models allow us to isolate whether platforming is more or less effective when considering 1) whether bans are temporary; 2) the reason for banning; 3) entity-level labels describing the influencers; and 4) the attention that the entities received in the pre-deplatforming period. In the latter case, we sort entities into two groups, according to the attention they received in the 12th month before deplatforming: high attention (top 1/3rd of influencers ranked per attention) *vs.* low attention (bottom 2/3rds).[X]

We show the coefficients associated with the interactions in Table 5.4. Our key findings are threefold. First, we find that temporary bans do not significantly differ from permanent ones at a 0.05 significance level (Model 1). This is true both on Google Trends ($p = 0.739$) and Wikipedia ($p = 0.791$). Second, we find that the effects of deplatforming are significantly stronger for entities associated with higher attention in the pre-deplatforming period, both on Google Trends (−58.8% 95% CI [−72.6%, −37.9%]) and Wikipedia (−33.7% 95% CI [−48.2%, −15.1%]) (Model 2). Third, we find that entities deplatformed because they spread misinformation suffer significantly larger decreases in online attention, again both on Google Trends (−59.8% 95% CI [−82.4%, −7.7%]) and Wikipedia (−48.7% 95% CI [−67.5%, −19.2%]) (Model 2). We do not find systematic differences associated with the entity labels (Model 4).

[X] High-attention influencers received, on average, 12,977 pageviews per day in the 12th month before deplatforming, whereas low-attention influencers received 100.

A noteworthy methodological concern related to Model 2, where we study differences in the effect for high *vs.* low-attention influencers, is that we might observe a regression to the mean. In other words, we may find the decrease particularly salient for high-attention influencers because we filtered them to be high-attention before the intervention, selecting influencers with an unusually high-attention month. However, we circumvent this issue by selecting high-attention users in a month far removed from the intervention (12 months before). Implicit in our DiD specification in Eq. (5.2) is that the reference month is set to $-3$, *i.e.*, we compare the effect relative to three months before the intervention.[XI] This is long after the month used for selecting high-attention influencers, and, therefore, any "regression to the mean" would likely have already occurred (and not influence the post-intervention results).

## 5.5 Discussion

In this section, we discuss the implications of our findings, methodological caveats, and promising venues for future work.

**Deplatforming decreases online attention.** Our quasi-experimental analysis of 165 deplatforming events associated with 101 influencers indicates that deplatforming decreases online attention toward influencers. This key finding largely aligns with previous work on the deplatforming of influencers on Twitter [115]. Yet, we argue that evidence from this previous study is, by itself, inconclusive due to the scope of the study (few events, one platform, active engagement only). In that context, we expect that our analysis gives stakeholders (including platforms and legislators) a more comprehensive and nuanced understanding of deplatforming.

**How and when to deplatform?** Deplatforming can be enacted in different ways (*e.g.*, temporary or permanent); upon different kinds of influencers (*e.g.*, politicians or media personalities); and for different reasons (*e.g.*, hate speech or misinformation). Unlike previous work, *e.g.*, [115, 99, 265], we consider many deplatforming events, which allow us to capture variations in the effectiveness of the intervention. We find that deplatforming is more effective when targeting popular influencers disseminating misinformation; yet, we argue that the key policy guidance our chapter can provide comes from a null result: that temporary deplatforming was similar to permanent deplatforming in how it reduced online attention toward influencers. We speculate that this may be partially due to deterrence, *i.e.*, temporarily banned influencers may avoid rule-breaking to prevent harsher sanctions. Here we did not discriminate between short (*e.g.*, one week) *vs.* long (*e.g.*, one month) temporary bans, which could be an interesting venue for future work. Broadly, permanent bans of prominent influencers have caused controversy as: the rationale behind the bans is often muddy [274]; there is no clear "reinstatement" procedure for banned accounts [62]; it is unclear if social media companies should be the ones to regulate speech [178]. Having bans of limited time is aligned with the broader call for a more transparent [250], consistent [134], and accountable [83] approach to social media moderation.

[XI] Other reference months, *e.g.*, $-2$, yield similar results.

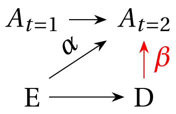


Figure 5.7: Directed Acyclic Graph depicting online attention before ($A_{t=1}$) and after ($A_{t=2}$) deplatforming ($D$). External events ($E$) may cause both the deplatforming of an entity and increase their online attention.

**External events causing changes to attention and the deplatforming.** In a prominent deplatforming event, Twitter, Facebook, YouTube, and Instagram banned former US President Donald Trump's account shortly after the 6 January 2021 U.S. Capitol Attack, alleging that he was inciting violence. Here, it is likely that the online attention received by Trump after his deplatforming would be influenced by both his suspension from major platforms *and* by the Capitol attack and his reaction to it. This illustrates a problem faced by our analysis, as we want to estimate the effect of deplatforming, not of the event that triggered the deplatforming. We argue this issue does not threaten the validity of our findings, given that we estimate that deplatforming harmed, rather than boosted, subsequent online attention. This is because a potential confounding event is only likely to *increase* online attention post-deplatforming; and since the effect that we estimated with our model is negative, this confounder can only weaken the effect observed. Therefore, the results obtained here can be interpreted as a lower bound of the true effect of deplatforming. In other words, if we could account for confounding due to an external event, the estimated effect of deplatforming would be even more strongly negative than the negative effects estimated by our model.

We more sharply illustrate this scenario with a directed acyclic graph as depicted in Fig. 5.7. This simple model considers the online attention $A$ that an entity receives in two time periods $t \in \{1,2\}$, and the attention in $t = 2$ is caused by 1) how much attention the entity received in $t = 1$, 2) the events $E$ associated with the entity in $t = 1$, and 3) whether the entity has been deplatformed ($D$). Our methodology is unable to distinguish between the causal effects associated with $E \rightarrow A_{t=2}$ (which we refer to as $\alpha$) and $D \rightarrow A_{t=2}$ (which we refer to as $\beta$), estimating both jointly ($\alpha + \beta$). Yet, as the total effect is negative ($\alpha + \beta < 0$) and as the effect from the event is likely to increase online attention ($\alpha > 0$), it follows that the real effect $\beta$ of deplatforming must be negative, and smaller or equal to the estimated effect $\alpha + \beta$.

**Beyond engagement on social media platforms.** An important way our research differs from previous work is that we go beyond social-media indicators of attention: we examine changes in overall online attention to these entities through search engine behavior and Wikipedia pageviews. This is, in part, what allowed us to circumvent the methodological limitations of previous work [221, 99, 115, 169, 265, 266]. Likes, posts, and comments are platform-specific and may underestimate the popularity of fringe influencers that migrate to alternative platforms (e.g., Gab, Rumble). In contrast, online attention traces like the ones we used are arguably better suited to capture the impact of moderation decisions on the broader information "ecosystem."

Researching social media has become increasingly difficult in what Freelon has named "the post-API age," the systematic degradation of infrastructure researchers use to study online platforms [69]. Lazer [136] wrote that "the Internet should be viewed as an [...] experiment, manipulating what people see and how they see it. The access that science has to this information is quite limited, however." Here, we consider holistic attention across the Web while studying phenomena inherently linked to social media. Other researchers, too, may use this approach to "dip their toes" in online attention data without partnering with social media companies.

Another approach that has been growing in popularity is the use of online panels captured by media companies [99, 2], where panelist, paid or voluntary, install browser extensions or mobile phone apps that allow companies and researchers to peek into their online habits, *e.g.*, analyzing their web history. In contrast to this strategy, using Google Trends and Wikipedia Pageviews is easier and remarkably cheaper. Nonetheless, it is worth noting the goal of initiatives such as the Observatory for Online Human and Platform Behavior [137] is to provide researchers with access to this data for free, so this can be a convenient way to conduct similar research in future work (the authors of this chapter are not involved in this initiative).

**Deliberate *vs.* coincidental attention.** Our findings are in contrast with, but do not contradict, previous work analyzing the impact of deplatforming Parler, a far-right social media platform, from Amazon Web Hosting Services [99, 2]. Previous work has found that banning Parler drove users to other alternative social media platforms (*e.g.*, Gab, Bitchute), such that the overall activity across platforms remained the same [99, 2]. At the same time, past research has found that engagement with radical content on YouTube is driven primarily by self-selection [102, 103, 37]. These findings suggest a fundamental difference between deplatforming influencers exposed to everyday users that do not necessarily wish to engage with them on mainstream platforms but *coincidentally* do so (*e.g.*, Alex Jones on Twitter) and influencers who are engaging mostly with a dedicated audience, that *deliberately* go out of their way to consume their content (*e.g.*, Alex Jones on Parler). Considering these previous findings, we conjecture that deplatforming is most effective in decreasing "spontaneous" attention, but further exploring these different "kinds" of online attention might be an interesting avenue for future work.

**Continuously tracking moderation interventions.** Social media platforms such as Facebook, YouTube, Twitter, and Reddit have changed the fabric of society. In part, the effects of these platforms on society are mediated by how they moderate content. Thus, we argue that comprehensive research examining the impact of moderation practices (*e.g.*, deplatforming) is key so that platform decisions are not made in a vacuum or solely to preserve the platforms' own financial interests. An essential element of this research agenda is the continued monitoring of the effectiveness of moderation interventions. Future work could continue to track and analyze moderation sanctions using the methodology provided here. It would be interesting to see if there are differences across countries and years as the digital ecosystem evolves.

**Ethical considerations.** In this work, we only used data publicly available on the Web and did not interact with online users in any way. We study and make several deplatforming events available. Yet, each event was covered in news pieces and widely discussed on Reddit, leading us to believe that we are not infringing on reasonable privacy expectations. We make available the names of deplatformed individuals both in the main text and in the data accompanying the research. We argue this exposure is reasonable and not in conflict with these individuals 'right to be forgotten,' as these are public figures, and these events are of public interest (as is the research at hand).

# Moderating content Part III

# 6 Removing content

**This chapter was the result of a collaboration with Meta (Facebook).**

## 6.1 Introduction

Many online platforms enforce community guidelines using automated content moderation systems that detect and intervene when rule-breaking occurs, *i.e.*, when user behavior violates community guidelines [163, 81, 264, 267]. These systems prevent harm by removing or reducing the visibility of rule-breaking content [85], *e.g.*, by reducing the number of people who see such content [72]. However, content removal or visibility reduction may also affect *on-platform* user behavior [130, 246]. Moderation interventions may increase compliance with community guidelines, *e.g.*, as deleted comments may prevent a conversation from derailing [300], or, conversely, backfire and increase rule-breaking, *e.g.*, because sanctioned users perceive the decision as unfair [35].

Understanding the causal effect of automated content moderation practices on user behavior is vital for evaluating these systems' effectiveness and can inform their design and use. However, measuring the causal effect of content moderation is difficult because of the ethical and technical challenges in using randomized experiments (e.g., A/B testing) to study content moderation practices [246]. Allowing some users not to be moderated implies not removing content that may harm others, and malicious actors could exploit the randomization of potential experiments to post harmful content.

Previous work has extensively documented the role of "manual" content moderation in online communities [130, 234], *i.e.*, where volunteer moderators find and remove content that breaches community guidelines. Some research has sought to estimate the effect of such manual content moderation on online communities, finding that it positively impacts user behavior [236, 246]. Nevertheless, these effects may not generalize to *automated*, platform-level content moderation. Further, research on content moderation has been typically descriptive [40, 64, 33, 74, 25, 116] rather than causal, and the quasi-experimental designs used in previous work are not readily adapted to an automated setting. For example, some approaches that rely on the randomness in time it takes for human moderators to intervene upon rule-breaking content to estimate the effect of content moderation [246] do not work for automated systems, in which moderation occurs immediately after content is created.

**Present work.** We study the effect of automatically enforcing community guidelines for violence and incitement for Facebook comments on user behavior with a quasi-experimental approach illustrated in Fig. 6.1. We examined subsequent rule-breaking behavior and commenting activity among users whose comments were moderated (user-level scenario) and among users in threads where these comments were posted (thread-level scenario). Analyzing over 412M comments, we measured the effect of two different interventions (hiding and deletion; see Fig. 6.2) on outcomes capturing commenting activity and rule-breaking behavior. Specifically, we estimated the causal effect of content moderation using a fuzzy regression discontinuity design [107], an approach that capitalizes on two design choices commonly used in automated content moderation systems [252, 195]. First, on many platforms, machine learning models that predict whether content breaches community guidelines assign a score of $S$ to each piece of content, reflecting its likelihood of breaking community guidelines. Second, when these assigned scores are sufficiently high, platforms may *automatically* enforce community guidelines. In other words, if a score exceeds a predetermined threshold $t$ and certain other conditions are met, content may be hidden or deleted immediately. Our approach allows us to mimic a randomized control trial around the threshold $t$ since content with a score right above the threshold (*i.e.*, $S = t + \epsilon$) is similar to content with a score right below (*i.e.*, $S = t - \epsilon$), but only the former is automatically intervened upon by the content moderation system.

**Results.** Overall, deleting comments reduced rule-breaking behavior in the thread where the comment was originally posted. Deleting comments also reduced rule-breaking among users whose comments were deleted, *i.e.*, other comments in the thread or that the user subsequently posted were hidden and deleted less often after the intervention. At the thread level, deleting rule-breaking comments significantly decreased rule-breaking behavior in threads with 20 or fewer comments before the intervention, even among other participants in the thread. This effect was statistically insignificant for threads with more than 20 comments. Deletion at the user level led to a decrease in subsequent rule breaking and posting activity. However, while the decline in rule breaking persisted with time, the decrease in posting activity waned. In other words, users gradually returned to making posts or comments at a rate similar to before their comments were deleted but were less likely to post comments that would subsequently be hidden or deleted. Hiding (rather than deleting) content had small and statistically insignificant effects on subsequent user activity and rule-breaking behavior at both the user and thread levels.

**Implications.** Deletions of rule-breaking content by automated content moderation, as currently applied on Facebook, decrease the subsequent creation of content that goes against community guidelines. Our results suggest two ways that this may happen. First, users whose comments are deleted are less likely to produce subsequent rule-breaking content. Second, other users are also less likely to create rule-breaking comments in the thread where the content was deleted. Building on previous work that found that “manual” content moderation [236, 246] can prevent rule-breaking behavior, here we show that these effects generalize to automated systems responsible for a substantial fraction of moderation interventions carried out by major social networking platforms [163, 81, 264, 267]. Though our results are limited in that we can only measure the effect of content moderation interventions triggered by classifiers at the thresholds at which they are applied, this study may clarify their present impact on online platforms such as Facebook. And while automated content moderation systems are typically assessed using precision and recall, this work shows how they may also be evaluated in terms of their effects on subsequent user behavior in an observational manner that does not require experimentation.

## 6.2 Background

**Violence and incitement policy.** In this chapter, we study a classifier and associated interventions used to help in enforcing Facebook’s community standards for violence and incitement. The policy[I] has the following rationale: *“We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence. (...)”*

**Interventions.** In this chapter, we study two interventions applied to rule-breaking content in the context of enforcing community guidelines. These interventions, illustrated in Fig. 6.1, are applied incrementally. Content whose score is greater than the first threshold $t_{\text{hide}}$ is hidden. Then, if the score crosses the second threshold $t_{\text{delete}}$, it is immediately deleted, and a warning is sent to the offending user. To other users, there is no indication that a post was made and later deleted. This approach aims to incrementally intervene upon content, acknowledging that some content that is borderline to community standards may remain in the social network with reduced visibility.[II]

**Scope.** The violence and incitement classifier studied here is only one of the ways that Facebook ensures that content follows community standards for violence and incitement. Other mechanisms also exist to ensure that content on Facebook adheres to these guidelines and that other community standards are enforced. These are beyond the scope of this chapter.

## 6.3 Materials and Methods

We studied the effect of automatically enforcing community guidelines with two quasi-experiments (Fig. 6.3). A *post* is a piece of content posted on Facebook, a *comment* is a response to that piece of content, and a *thread* comprises comments associated with a post.

[I] https://transparency.fb.com/policies/community-standards/violence-incitement/.

[II] https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote

**Data.** For both quasi-experiments, we used a dataset of public comments and posts posted by adult U.S. users in English between June 1st and August 31st, 2022. This comprised 412 million comments made in 1.5 million posts by 1.3 million distinct users. All data was de-identified and analyzed in aggregate, and no individual-level data was viewed by the researchers. (Data was provided by Facebook/Meta and is, unfortunately, unavailable for reproducibility purposes.)

**Thread level.** In this first scenario (Fig. 6.3a), we studied the impact of automatic moderation on the thread where comments were intervened upon. For each post in our data, we looked for the first comment $c_0$ whose score was in the 5 percentage point range of either of the two thresholds where the "hide" and the "delete" interventions are applied, *i.e.*, $S \in [t_{\text{hide}} - 0.05, t_{\text{hide}} + 0.05]$ or $S \in [t_{\text{delete}} - 0.05, t_{\text{delete}} + 0.05]$; recall that $S \in [0,1]$. If a thread had no comments that met the above criteria, it was excluded from this analysis. For each comment $c_0$ selected this way, we considered all comments made before (in the *pre-assignment period*) and after $c_0$ in the same thread (in the *follow-up period*). After computing the outcome measures using data from the follow-up period, we used fuzzy regression discontinuity (see Sec. 6.3) to determine the effect of hiding or deleting the comment. To study effect heterogeneity, we considered four different setups in this quasi-experiment, varying (1) whether we included other comments from the author of the selected comment $c_0$ when calculating the outcomes of interest in the follow-up period; and (2) whether we considered threads that had more than 20 comments. We choose 20 as a cutoff point as it induces an 80/20 split, *i.e.*, around 80% of the threads have less than 20 comments. A 75/25 or 85/15 split yielded similar results.

**User level.** In the second scenario (Fig. 6.3b), we studied the impact of automatic moderation on the users whose comments were intervened upon. For each user $u$ in our data, we looked for the first comment in the study period $c_0$ whose score was in the 5 percentage point range of the "hide" and "delete" thresholds. If a user had no comments meeting the above criteria, they were excluded from this analysis. For each user/comment tuple $(u_0, c_0)$ selected this way, we additionally considered all comments the user $u_0$ made in the $k$ days before (in the *pre-assignment period*) and after posting $c_0$ (in the *follow-up period*). Again, data from the follow-up period was used to calculate outcomes, and a fuzzy regression discontinuity design was used to determine the effect of the interventions. We studied the heterogeneity of the effect of these interventions in two ways. First, we varied the value of $k$, the number of days in the follow-up period (we considered $k \in \{7, 14, 21, 28\}$). Second, we separately considered (1) users who had not violated community guidelines recently and only received a warning after having their comment deleted; and (2) users who had violated community guidelines once in the recent past and were thus suspended from posting on Facebook for a day after their comment was deleted.[III]

**Outcomes.** The outcomes considered in this study are shown in Table 6.1. One outcome is associated with subsequent rule-breaking behavior (interventions), while one is associated with subsequent activity on the platform (comments).

[III] https://transparency.fb.com/en-gb/enforcement/taking-action/restricting-accounts/

Table 6.1: Outcomes considered in this study.

| Outcome | Description |
|---|---|
| Interventions in follow-up period | The number of interventions that, in the follow-up period, targeted either the comments made by the user (in the user-level scenario) or the subsequent comments in the thread (in the thread-level scenario). |
| Comments | The number of comments made during the follow-up period. In the user-level scenario, we also include posts. |

**Regression Discontinuity Designs.** Regression discontinuity (RD) is a quasi-experimental study design widely used in the social sciences since the 1990s [107]. Here, we provide an overview of the fuzzy regression discontinuity design (an extension of RD), explaining how we use it for the quasi-experiments described in Section 6.3 and illustrated in Fig. 6.3.

**Fuzzy Regression Discontinuity (FRD).** Let each comment $c$ be assigned a score $S_c \in [0,1]$, and $X_c$ be an indicator variable that equals 1 if the content has been intervened upon and 0 otherwise. If the $S$ score is beyond a threshold $t$, the probability of that comment getting intervened upon increases sharply:

$$P[X_c = 1 \mid S_c] = \begin{cases} f_1(S_c) \text{ if } S_c \geq t \\ f_0(S_c) \text{ if } S_c < t \end{cases} \quad \text{where } f_1(a) > f_0(a) \quad \forall a. \tag{6.1}$$

Note that this is a generalization of *sharp* regression discontinuity designs, where $f_0(S_c) = 0$ and $f_1(S_c) = 1$ , *i.e.* $P[X_c = 1 \mid S_c]$ jumps from 0 to 1 around the threshold. This is more suited to the scenario we are studying since mechanisms other than the classifier may come into play, *e.g.*, comments may be removed due to user reports when the score is below the threshold ($S_c < t$), and other automated systems may prevent comments above the threshold from being removed when the score is above the threshold ($S_c \geq t$). A directed acyclic graph (DAG) illustrating the causal relationship between the score, the treatment, and the outcomes we are interested in measuring is shown in Fig. 6.4a. The treatment $X$ is determined by the score $S$ given by the classifier and other factors unobserved in the present study (represented by $U$).

The key insight of fuzzy regression discontinuity designs is to estimate the effect of the intervention $X$ on the outcome $Y$, even with unknown confounders $U$, for comments with scores in the interval $S_c \in [t-\epsilon, t+\epsilon], \epsilon \to 0$. We assume that comments that lie right before or right after the threshold are indistinguishable, but those above the threshold are more likely to receive the treatment than those below. Thus, around the threshold, we can consider a new DAG where there is no arrow $U \to S$, as shown in Fig. 6.4b. Here, $S$ has a causal effect on $Y$ only through $X$, and thus we can use the same idea behind instrumental variable (IV) designs [10] to study the effect of $X$ on $Y$. In IV designs, we estimate the Local Average Treatment

Effect (LATE), the treatment effect for the subset of the comments that take the treatment (*i.e.*, $X_c = 1$) if and only if they were "assigned" to the treatment (*i.e.*, $S_c > t$):

$$LATE = \frac{ITT}{ITT_d}, \tag{6.2}$$

where ITT is the average effect of assigning comments to the treatment group (regardless of them being treated), and $ITT_d$ is the proportion of subjects treated when assigned to the treated group. As $S$ is only an instrument close to the threshold $t$, we estimate the LATE at the cutoff point (LATEC), rewriting Equation (6.2) as:

$$LATEC = \frac{E[Y_c \mid S_c = t+\epsilon] - E[Y_c \mid S_c = t-\epsilon]}{E[X_c \mid S_c = t+\epsilon] - E[X_c \mid S_c = t-\epsilon]}, \ \epsilon \to 0. \tag{6.3}$$

In practice, we can estimate the LATEC with 2-stage least squares regression, *i.e.*, regressing the treatment $X$ on the score $S$ (first-stage), and then the outcome $Y$ on the values $\hat{X}$ predicted on the first-stage (second-stage), see [9] for details. However, we do not have infinite data, and we cannot consider only comments with $S_c \in [t-\epsilon, t+\epsilon]$, $\epsilon \to 0$. This creates a bias–variance trade-off in the estimation of the LATEC. On the one hand, the wider the range we consider around the threshold $t$, the more the unmeasured confounders can bias our estimator. On the other hand, the narrower the range, the less data we have, and thus the larger the variance of our estimator.

A common solution to navigating this trade-off consists of using a local linear regression [89], where data points (here, comments) receive importance proportional to how far they are from the threshold, using a triangular weighting kernel defined as

$$K(S) = \mathbf{1}_{|S-t|<h}\left(1 - \frac{S-t}{h}\right), \tag{6.4}$$

where $h$ is the bandwidth of the kernel that controls the bias–variance trade-off, and $\mathbf{1}_{|S-t|<h}$ is an indicator variable that equals 1 if $|S-t| < h$ and 0 otherwise. We empirically determine the bandwidth $h$, choosing the bandwidth that yields the optimal mean squared error (MSE) of the LATEC estimator [106].

**Example.** Fig. 6.5 illustrates our fuzzy regression discontinuity design. It uses a random sample of users who did not previously violate community guidelines and examines interventions in a 7-day follow-up period following the first comment of interest ($c_0$). Figure 6.5 (top) shows the percentage of first comments ($c_0$) that received the "Delete" treatment (*i.e.* $E[X|S]$; in the $y$-axis) for different scores received by first comments $c_0$ (in the $x$-axis). Figure 6.5 (bottom) depicts the outcome "Interventions in the follow-up period" ($E[Y|S]$; in the $y$-axis) for different scores by first comments $c_0$ (in the $x$-axis). Intuitively, the regression discontinuity design estimates the treatment effect of $X$ on $Y$ around the threshold $t_{\text{delete}}$ by dividing the discontinuity in $E[Y|S]$ [corresponding to the numerator in Eq. (6.2) and Eq. (6.3)] by the discontinuity in $E[X|S]$ [corresponding to the denominator in Eq. (6.2) and Eq. (6.3)].

**Robustness checks.** To ensure the validity of our regression discontinuity design, we additionally conduct several robustness checks suggested by guides outlining best practices [139, 110]. These robustness checks can be found in Appendix A.1.

## 6.4 Results

### 6.4.1 Thread level

Fig. 6.6(a) shows the standardized effect of deleting comments on the number of comments and interventions in the follow-up period in the thread-level scenario). Comment deletion had a significant effect on both the number of subsequent interventions and the number of subsequent comments in threads that had fewer than 20 posts prior to the intervention. When comments from the original commenter were included (**≤20/All**), the intervention reduced the number of comments by −13.16 (95% CI: −21.23, −5.10) and the number of subsequent interventions by −0.946 (95% CI: −1.59, −0.299; see non-standardized effects shown in Tables 6.2 and 6.3.). To get a sense of the effect size, we calculated the average number of comments and interventions in the follow-up period received by threads right below the intervention threshold, where $S \in [t_{\text{delete}} - 0.01, t_{\text{delete}})$. Threads in the **≤20/All** scenario just before the threshold received 1.5 interventions (95% CI: 1.3, 1.9) and 27.7 comments (95% CI: 23.2, 33.,2) on average, suggesting that these effects were substantial.

Still considering the **All** scenario, for both outcomes, the effect of deletions was neither substantial nor significant for threads that already had more than 20 comments when the delete intervention happened [*e.g.*, see >**20/All** in Fig. 6.6(a)]. Deleting any single comment may have less of an effect in longer threads because participants are less likely to see such a comment (e.g., because such a comment may already have been hidden or because there are at least 19 other comments to see). Effects may also have been more difficult to observe because of the smaller sample size—there were fewer threads with more than 20 comments than threads with 20 or fewer comments. Hiding comments, as opposed to deleting, had small and statistically insignificant effects on the number of subsequent interventions and comments for all setups considered (see Table 6.3; we discuss this further in Sec. 6.5).

We also computed the same outcomes in the follow-up period, not considering comments by the original commenter to understand if the effect was due to changes in the behavior of the individual who had their comment intervened upon or other users in the thread (scenarios **≤20/Other** and >**20/Other** in Fig. 6.6(a) and Tables 6.2 and 6.3). We found that standardized effects remained qualitatively similar, *e.g.* comments were reduced −0.069 standard deviations (SDs) in the **≤20/All** setup *vs.* −0.037 SDs in the **≤20/Other** setup, suggesting that the intervention discouraged *other users* from posting rule-breaking comments.

### 6.4.2 User level

For the user-level scenario, we considered both cases where users did and did not have comments deleted previously. We make this distinction as the interventions for these users differ: “first-time offenders” only receive a warning, whereas “repeat offenders” additionally have their posting privileges suspended for 24 hours. The suspension period for repeat offenders is not considered in the follow-up period as it could explain behavior differences.

**User level: first-time offenders** Fig. 6.6(b) shows the effect of deletions for first-time offenders. Again, deleting comments had significant effects on both outcomes. Considering the 7 days following the intervention, deletion decreased the number of comments by 4.6 and decreased the number of subsequent interventions by 0.12. To get a sense of the effect size, we calculated outcomes for users right below the intervention threshold, $S \in [t_{\text{delete}} - 0.01, t_{\text{delete}})$. These users received on average 0.23 interventions in the follow-up period (95% CI: 0.21, 0.25) and made on average 13 comments (95% CI: 12.6, 13.4), suggesting that the effects are substantial. Setups that considered larger intervention periods (21 and 28 days) showed that, while the effect on the subsequent number of comments waned with time (*i.e.*, effects were smaller for longer follow-up periods), the effect on the number of subsequent interventions was largely stable. This indicates that automated content moderation has positive, long-lasting effects on subsequent rule-breaking behavior. Hiding comments had small and statistically insignificant effects on the number of subsequent interventions and comments.

**User level: repeat offenders** Fig. 6.6(c) shows the effect of deletions for repeat offenders. For these users, deleting comments yielded decreases in both the number of interventions and comments in the follow-up period. The wider confidence intervals here may be partially explained by the smaller sample, as fewer users had their comments deleted a second time. Nonetheless, for 3 out of the 4 time periods considered (7, 14, 28 days), we again observed significant effects that were similar in magnitude to the effects observed in the “first-time offender” setup. Considering a 28-day follow-up period, deletions decreased interventions received by repeat offenders by 0.28 (95% CI: −0.48, −0.078) *vs.* 0.192 for first-time offenders. This suggests that deletions are also effective for users who have previously broken community guidelines. Hiding comments had small and statistically insignificant effects on the considered outcomes.

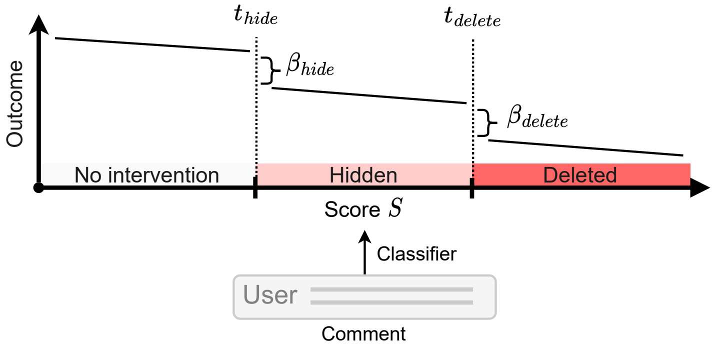
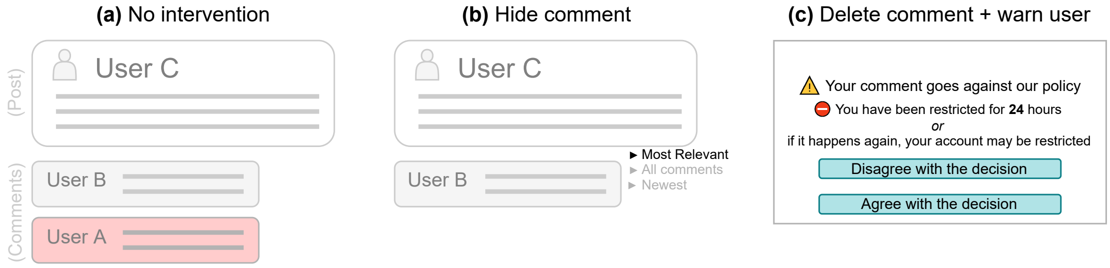


Figure 6.1: Comments posted on Facebook are scored by classifiers that measure adherence to community standards. When $S$ crosses specific thresholds ($t_{\text{hide}}$ and $t_{\text{delete}}$ in the figure), different interventions are applied. Though comments around each threshold are similar, they receive different interventions, *e.g.*, a comment with score $t_{\text{delete}} - \epsilon$ is hidden, while a comment with score $t_{\text{delete}} + \epsilon$ is deleted. Exploiting this fact, this work measures the impact of these interventions on user behavior outcomes by studying the discontinuities ($\beta_{\text{hide}}$ and $\beta_{\text{delete}}$) around the thresholds $t_{\text{hide}}$ and $t_{\text{delete}}$.



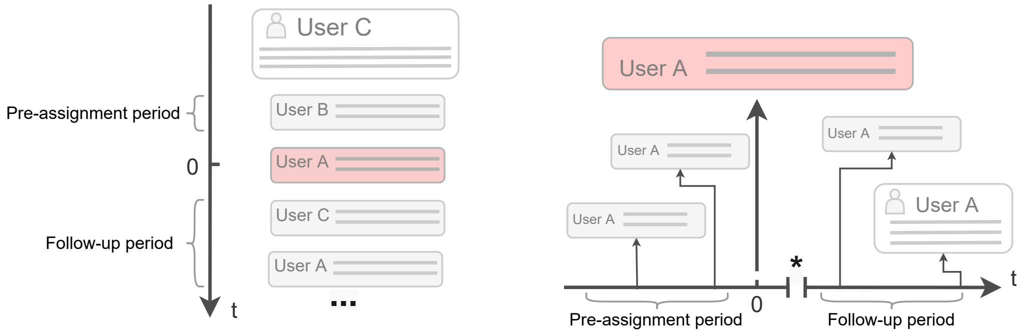
⚠ Your comment goes against our policy

⛔ You have been restricted for **24** hours

*or*

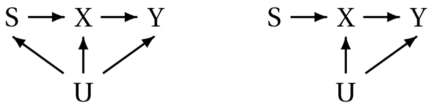
if it happens again, your account may be restricted

Figure 6.2: Moderation interventions – We depict a hypothetical scenario where *User A* writes a rule-breaking comment (in red) on a post by *User C* (white) that has received a comment by *User B*, (grey). Depending on the score a comment receives, (a) no intervention may be applied, in which case the comment is posted, (b) the comment may be *hidden*, and it will only be visible if a viewer changes the default comment ranking setting to show all comments, or (c) the comment may be *deleted* and the user who posted the comment warned (an additional sanction may be applied depending on their previous rule-breaking behavior).



(a) Thread-level scenario (b) User-level scenario

Figure 6.3: The study approximates a real experiment where comments were intervened upon at random using observational data using a fuzzy regression discontinuity design. We depict the thread-level and user-level scenarios in (a) and (b) and describe them in Sec. 6.3. In (b), the asterisk denotes the setup where users are suspended, and the suspension period is not considered when calculating the outcomes.



(a) (b)

Figure 6.4: Causal Directed Acyclic Graphs (DAGs) illustrating the fuzzy regression discontinuity design. $S$ is the score attributed to a comment, $X$ is an indicator variable representing whether the comment was intervened upon, $U$ are unmeasured confounders, and $Y$ is the outcome of interest. While estimating the effect of $X$ on $Y$ is not possible in (a), around a specific threshold $t$ where there is a discontinuity around the probability of treatment ($P[X=1|S]$), we can remove the arrow $U \rightarrow X$ [see (b)] and use the same idea behind instrumental variable designs to measure the effect of $X$ on $Y$ (see main text).

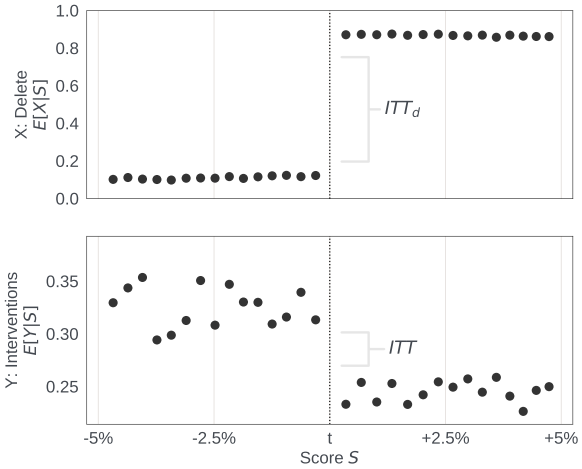


Figure 6.5: A real example of our fuzzy regression discontinuity approach, considering the output of the violence and incitement classifier as the running variable $S$, deletions as the treatment $X$, and the number of interventions in a 7 day follow-up period as the outcome $Y$. We estimate the causal effect as the ratio between two discontinuities ($ITT/ITT_d$). $ITT_d$ (top figure) is the discontinuity in the treatment around the threshold $t$ (i.e., the probability of deletion), while $ITT$ (bottom figure) is the discontinuity in the outcome of interest around the same threshold (i.e., the number of interventions in a 7-day follow-up period).

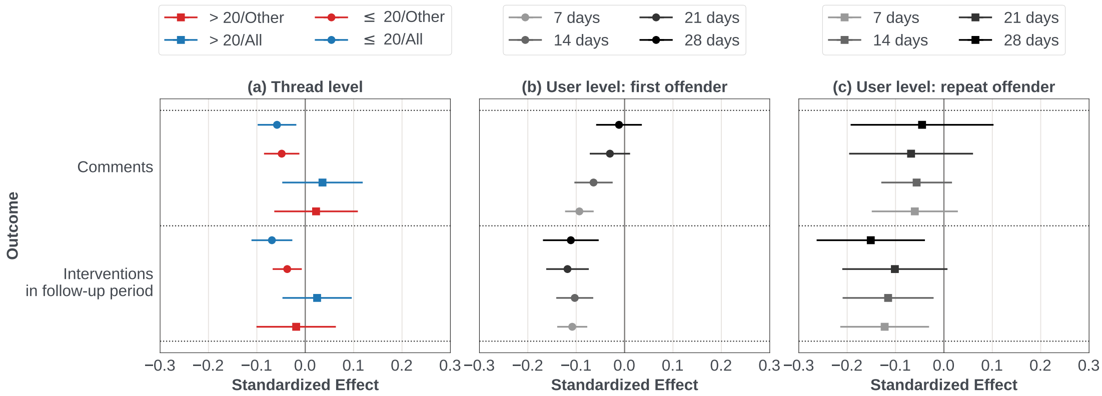


Figure 6.6: We depict the estimated standardized effect of *deleting* comments at the thread (a) and user level (b and c). Error bars represent 95% CIs. Comment deletions can reduce subsequent activity (as measured by comments) and rule-breaking behavior (as measured by interventions in the follow-up period) across both (a) thread- and (b/c) user-level scenarios.

Table 6.2: Summary of the effects across all settings for the **Delete** intervention. For the thread-level scenario, the setups where we consider only threads with less than 20 comments are marked with ≤ 20 in the "Setup" column (*vs.* >20 for the setup considering more than 20 comments), and the setups where the original commenter is not considered are marked as "Other" (*vs.* "All" for when they are). For the user-level scenario, the "Setup" column shows the number of days considered in the follow-up period. Stars ∗ indicate statistically significant effects, *i.e.* $p < 0.05$

**Delete**: Thread-level

| Outcome | Setup | Effect | Effect (Standardized) | n |
|---|---|---|---|---|
| Comments | ≤20 / All | -13.16 (-21.23, -5.10)$^*$ | -0.069 (-0.111, -0.027)$^*$ | 190885 |
| | ≤20 / Other | -7.41 (-13.39, -1.42)$^*$ | -0.037 (-0.067, -0.007)$^*$ | 200655 |
| | >20 / All | 43.22 (-82.98, 169.41) | 0.025 (-0.047, 0.096) | 49645 |
| | >20 / Other | -34.03 (-183.04, 114.97) | -0.019 (-0.101, 0.063) | 52241 |
| Interventions | ≤20 / All | -0.946 (-1.59, -0.299)$^*$ | -0.058 (-0.098, -0.018)$^*$ | 190885 |
| | ≤20 / Other | -0.876 (-1.53, -0.218)$^*$ | -0.049 (-0.085, -0.012)$^*$ | 200655 |
| | >20 / All | 2.27 (-3.02, 7.56) | 0.036 (-0.048, 0.119) | 49645 |
| | >20 / Other | 1.39 (-3.97, 6.74) | 0.022 (-0.064, 0.109) | 52241 |
| **Delete**: User-level (first offender) | | | | |
| Comments | 7 | -4.55 (-6.00, -3.11)$^*$ | -0.093 (-0.123, -0.064)$^*$ | 162149 |
| | 14 | -5.72 (-9.25, -2.19)$^*$ | -0.064 (-0.104, -0.025)$^*$ | 112793 |
| | 21 | -3.95 (-9.38, 1.48) | -0.030 (-0.072, 0.011) | 84592 |
| | 28 | -1.99 (-10.12, 6.14) | -0.012 (-0.059, 0.036) | 60175 |
| Interventions | 7 | -0.117 (-0.151, -0.084)$^*$ | -0.108 (-0.139, -0.077)$^*$ | 162149 |
| | 14 | -0.144 (-0.197, -0.090)$^*$ | -0.103 (-0.141, -0.065)$^*$ | 112793 |
| | 21 | -0.196 (-0.270, -0.123)$^*$ | -0.118 (-0.162, -0.074)$^*$ | 84592 |
| | 28 | -0.192 (-0.291, -0.092)$^*$ | -0.111 (-0.169, -0.053)$^*$ | 60175 |
| **Delete**: User-level (repeat offender) | | | | |
| Comments | 7 | -3.37 (-8.35, 1.61) | -0.060 (-0.149, 0.029) | 29825 |
| | 14 | -6.11 (-14.02, 1.81) | -0.056 (-0.129, 0.017) | 26596 |
| | 21 | -11.07 (-31.95, 9.81) | -0.068 (-0.196, 0.060) | 21693 |
| | 28 | -9.84 (-42.03, 22.34) | -0.045 (-0.193, 0.103) | 18468 |
| Interventions | 7 | -0.107 (-0.187, -0.027)$^*$ | -0.122 (-0.214, -0.031)$^*$ | 29825 |
| | 14 | -0.155 (-0.281, -0.029)$^*$ | -0.115 (-0.209, -0.021)$^*$ | 26596 |
| | 21 | -0.166 (-0.344, 0.012) | -0.101 (-0.210, 0.007) | 21693 |
| | 28 | -0.277 (-0.483, -0.072)$^*$ | -0.151 (-0.263, -0.039)$^*$ | 18468 |

Table 6.3: Summary of the effects across all setting for the **Hide** intervention, see Table 6.2 for details.

| **Hide**: Thread-level | | | | |
|---|---|---|---|---|
| Outcome | Setup | Effect | Effect (Standardized) | n |
| Comments | ≤20 / All | 0.718 (-1.93, 3.36) | 0.004 (-0.011, 0.019) | 868632 |
| | ≤20 / Other | 0.928 (-1.31, 3.17) | 0.005 (-0.007, 0.018) | 907871 |
| | >20 / All | -8.90 (-67.40, 49.60) | -0.003 (-0.026, 0.019) | 300716 |
| | >20 / Other | -10.94 (-68.34, 46.45) | -0.004 (-0.025, 0.017) | 314287 |
| Interventions | ≤20 / All | 0.023 (-0.096, 0.142) | 0.003 (-0.011, 0.017) | 868632 |
| | ≤20 / Other | 0.049 (-0.065, 0.163) | 0.006 (-0.008, 0.019) | 907871 |
| | >20 / All | 0.285 (-0.492, 1.06) | 0.011 (-0.018, 0.040) | 300716 |
| | >20 / Other | 0.241 (-0.492, 0.974) | 0.009 (-0.018, 0.036) | 314287 |
| **Hide**: User-level (first offender) | | | | |
| Comments | 7 | -0.568 (-1.41, 0.272) | -0.010 (-0.024, 0.005) | 723278 |
| | 14 | -1.19 (-2.90, 0.526) | -0.011 (-0.027, 0.005) | 542780 |
| | 21 | -1.98 (-4.97, 1.00) | -0.013 (-0.031, 0.006) | 422996 |
| | 28 | -0.350 (-6.11, 5.41) | -0.002 (-0.029, 0.026) | 291712 |
| Interventions | 7 | 0.007 (-0.011, 0.026) | 0.008 (-0.012, 0.027) | 723278 |
| | 14 | 0.005 (-0.015, 0.025) | 0.004 (-0.012, 0.020) | 542780 |
| | 21 | 0.018 (-0.015, 0.050) | 0.012 (-0.010, 0.033) | 422996 |
| | 28 | 0.032 (-0.007, 0.071) | 0.019 (-0.004, 0.041) | 291712 |
| **Hide**: User-level (repeat offender) | | | | |
| Comments | 7 | 0.948 (-1.11, 3.01) | 0.014 (-0.017, 0.045) | 126587 |
| | 14 | 2.68 (-1.44, 6.79) | 0.021 (-0.011, 0.053) | 122545 |
| | 21 | 3.30 (-3.64, 10.24) | 0.017 (-0.019, 0.053) | 103467 |
| | 28 | 3.92 (-6.41, 14.24) | 0.015 (-0.025, 0.056) | 82067 |
| Interventions | 7 | -0.002 (-0.034, 0.029) | -0.003 (-0.039, 0.033) | 126587 |
| | 14 | 0.017 (-0.025, 0.059) | 0.013 (-0.018, 0.044) | 122545 |
| | 21 | 0.013 (-0.066, 0.092) | 0.008 (-0.039, 0.054) | 103467 |
| | 28 | 0.000 (-0.113, 0.113) | 0.000 (-0.059, 0.060) | 82067 |

## 6.5 Discussion

Content moderation systems are essential to the functioning of mainstream social networks [75] and can prevent harm by removing rule-breaking content before anyone sees or interacts with it [85]. In this work, we studied how these systems may also positively impact on-platform user behavior. Using a fuzzy regression discontinuity design [107], we found that comment deletion had substantial and statistically significant effects on subsequent rule-breaking behavior and user activity. At the user level, for "first-time offenders," deletions had long-lasting effects on reducing rule breaking, but only temporary effects on posting activity, suggesting that comment moderation does not necessarily require making a trade-off between safety and engagement. This result is qualitatively aligned with the findings of Srinivasan et al. [246] on the r/ChangeMyView community on Reddit and suggests that automated platform-level moderation may yield the same effects as manual community-level moderation. At the thread level, we found that content moderation reduced rule-breaking activity even for other users who were not intervened upon. This result is qualitatively aligned with previous work suggesting that uncivil behavior is contagious [39, 128], further highlighting the importance of proactive content moderation.

We also found that hiding comments did not have substantial or significant effects. This may be linked to an important limitation of our work: we were able to measure the effect of content moderation only at the thresholds at which they were applied. The hiding intervention may have a stronger effect at a different threshold. Importantly, this study does not necessarily imply that comment hiding is not useful, as hiding comments can still prevent harm by reducing exposure to borderline content and may have other beneficial effects that we did not measure. In that context, future work could also find ways to estimate the effect of moderation across various thresholds. At the same time, the effects of deletion reported here may also be an underestimate. As interventions on Facebook are "cumulative," when we study the effect of deletion, we do not compare "deletion" with "no deletion," but instead can only compare "deletion" with "hiding." Therefore, it could be that the effect of deleting content is even stronger, but that part of the effect is masked by the "hiding" intervention (which, as previously stated, might itself be impactful if enacted at higher thresholds). Last, our study is also limited in that we consider specific interventions enacted only upon U.S.-based Facebook users, with effects that could be heterogeneous across other platforms and countries. Despite the aforementioned limitations, we argue that, even in this specific setting, understanding the impact of in-production content moderation systems is of great importance as a first step toward a more holistic understanding of how automated moderation systems impact online platforms such as Facebook.

Last, we argue that the methodology discussed and applied in this chapter can be used to assess moderation interventions across different scenarios and platforms. While much of the literature on harmful content has focused on developing methods to accurately detect such content, here we provide a way to measure the effects of deploying these systems (and their associated interventions) on our information ecosystem.

# 7 Pre-approving content

**This chapter was the result of a collaboration with Meta (Facebook).**

## 7.1 Introduction

Online communities are partially shaped by the design affordances of the platforms they inhabit [23]. In Facebook Groups, administrators can turn on “post approvals,” a setting that requires members’ posts to be accepted by community leaders (i.e., administrators and moderators) before others in the group can see and interact with them [166]. This setting changes *when* norms are enforced in a community or group, as illustrated in Figure 7.1. If the setting is turned off, community leaders must reactively moderate the posts in the community, e.g., by browsing posts in the group as they appear or by responding to reports from other members or the platform. If the setting is turned on, leaders can proactively moderate the community, prescreening posts that are low quality or that break the rules.

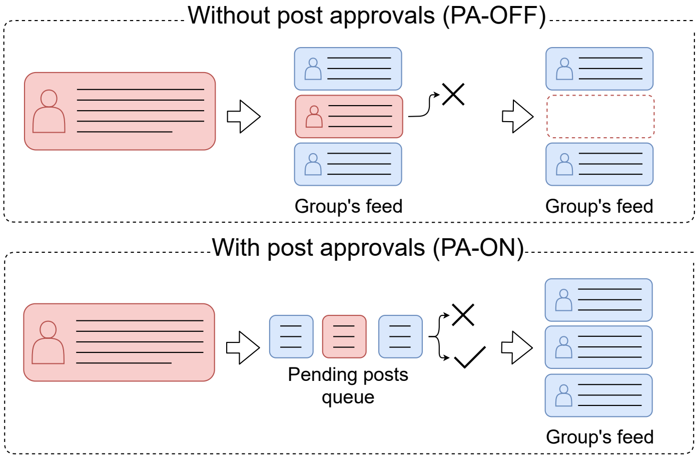


Figure 7.1: Post approvals allow posts that violate a community’s guidelines (in red) to be filtered *before* other members in the community see them. Without post approvals, these posts can only be moderated *after* they are posted in the group.

Well-moderated spaces are more attractive to users [289] and can improve the quality of users' contributions [50]. However, over-enforcement of rules can discourage participation [112, 125], and moderation creates more work for leaders [146, 56]. Thus, post approvals, a proactive moderation strategy, involves several trade-offs. On the one hand, it may prevent harm caused by violations of a community's guidelines and improve members' overall experience. On the other hand, it introduces participation friction and may increase leaders' workloads.

**Present work.** This chapter presents an observational study of the adoption and the impact of post approvals in online communities. We ask:

- **RQ1** What leads communities to adopt post approvals?
- **RQ2** How do post approvals shape user activity and moderation in online communities?
- **RQ3** Does the impact of post approvals depend on community properties and how the setting is used?

Using a longitudinal dataset of user activity- and moderation-related traces from 233,402 Facebook Groups from March to July 2021, we compared communities that enabled post approvals (PA-ON; $n$ = 8,767) to communities that did not change any moderation-related settings (PA-OFF; $n$ = 224,635).

To examine the factors that led to the adoption of post approvals **(RQ1)**, we studied activity in PA-ON and PA-OFF communities in the 4 weeks before the former enabled the setting. During this period, PA-ON communities experienced greater growth in user activity (*e.g., Comments*) and moderation (*e.g., Posts reported*) compared to PA-OFF communities. Further, right before PA-ON communities enabled post approvals, they experienced a sudden increase in moderation, which may have been the final straw that led administrators to turn on the setting. These findings continue to hold when using propensity score matching to control for initial baseline user and moderation activity, and are further confirmed when examining how user activity or moderation predicts if post approvals will be turned on in future weeks.

To study how post approvals shape online communities **(RQ2)**, we matched PA-ON and PA-OFF communities on user activity and moderation traces in the 4 weeks prior to post approvals being turned on and compared differences in their subsequent activity. We found that, while fewer posts were shared in groups that enabled post approvals, the posts that were shared received more comments, more reactions, more time spent, and fewer reports, suggesting improvements in the quality of content being posted. Further, post approvals did not significantly increase the average time leaders spent in their groups, though groups that enabled the setting tended to increase their moderation team.

Last, post approvals may differently impact a community depending on its properties and on how post approvals are used in practice **(RQ3)**. To understand these differences, we studied how the effects of post approvals varied with group size, leaders' response time for submitted posts (*i.e.*, how much time did it take for a post to be approved) and the post approval rate (*i.e.*, what fraction of posts submitted in a given group were approved).

For all three factors, we found significant interactions with time spent by leaders in the group and with changes in activity in the group following the adoption of the setting. Leaders spent significantly more time after adopting post approvals in larger groups, groups with higher post approval rates, and groups with faster response times. There were sharper decreases in the number of posts and increases in the number of comments, reactions, and time spent per post in larger groups, groups with lower post approval rates, and groups with slower response times. Still, other changes persisted across different communities. Independent of group size, response time, or approval rate, the fraction of posts reported decreased significantly after post approvals was adopted. This suggests that regardless of how post approvals was enforced, the setting nonetheless reduced content perceived by members as problematic or rule-breaking.

Overall, our findings suggest that post approvals substantially change how online communities work and that the setting creates communities centered around fewer, higher-quality posts. These insights may guide improvements to community-level moderation processes and the quasi-experimental approach we adopted can be easily extended to analyze other opt-in features provided by social media platforms.

## 7.2 Background

Moderation in online communities increases their attractiveness to newcomers [289], improves the quality of contributions [50], and decreases anti-social behavior [236].

Nonetheless, effective moderation is difficult – platforms experience several challenges related to the scale, the legitimacy, and the contextual nature of content moderation [75, 65]. Beyond work to better predict when content may violate community guidelines [231, 184], community-oriented social media (and moderation) has been suggested as part of the solution to these challenges because community leaders may better incorporate local and cultural context into moderation decisions [234] and because the decisions taken would be considered more legitimate [65]. To this end, some research has examined how moderators engage and regulate their communities [237] by developing specific design guidance from fundamental theories in the social sciences [130].

Most relevant to the present work are existing studies that explored how technological affordances provided by platforms shape content moderation. For example, participation controls [130] limit what specific users are allowed to see or do within a social media platform or a specific online community. In the development of open-source software, collaborators receive “commit rights” as they offer evidence of their technical expertise [58]; on PalTalk (an early video group chat service), moderators could impose “activity quotas” to chat room users, limiting their participation [130]; on Twitch, moderation includes chat “modes” that change how users can participate, for instance allowing only emotes to be sent [236]; on Reddit, Jhaver et al. (2019) [116] studied the usage of *AutoModerator*, a system that allows moderators to define “rules” to be automatically applied to posts in their communities.

This work examines post approvals, a participation control that is central to community-level moderation in Facebook Groups but whose specific effects have not yet been systematically studied. Post approvals change the dynamics of content moderation by allowing community leaders to *proactively* moderate posts before they ever land in the communities' feeds. Moreover, when someone attempts to contribute to a community with post approvals turned on, posts may take hours or even days to get published (if they do). Communities could thrive in the better-moderated spaces enabled by post approvals [289] and the participation friction could disrupt mindless interactions [160]. However, the setting could also discourage participation [125] and create unnecessary work for leaders [146, 56].

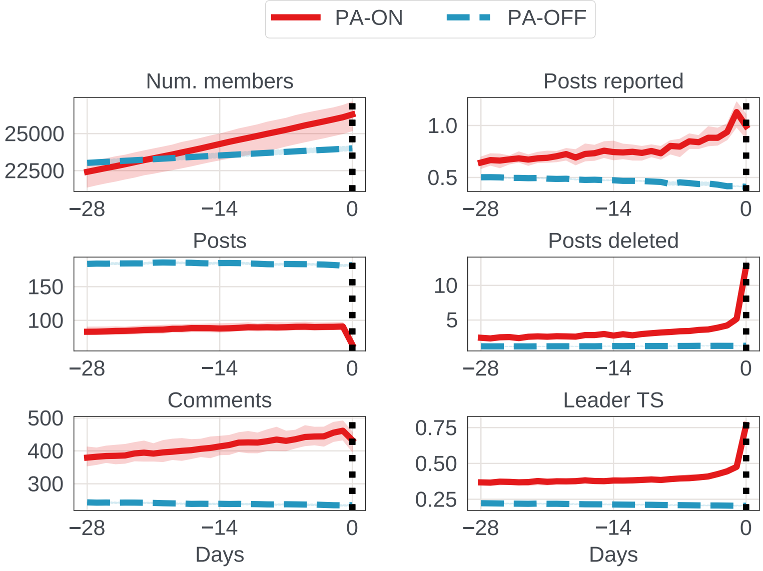
Studying the impact of post approvals (and other participation controls) in online communities can help create better governance practices and further our understanding of how participation friction and proactive moderation can improve online spaces.

## 7.3 Materials and Methods

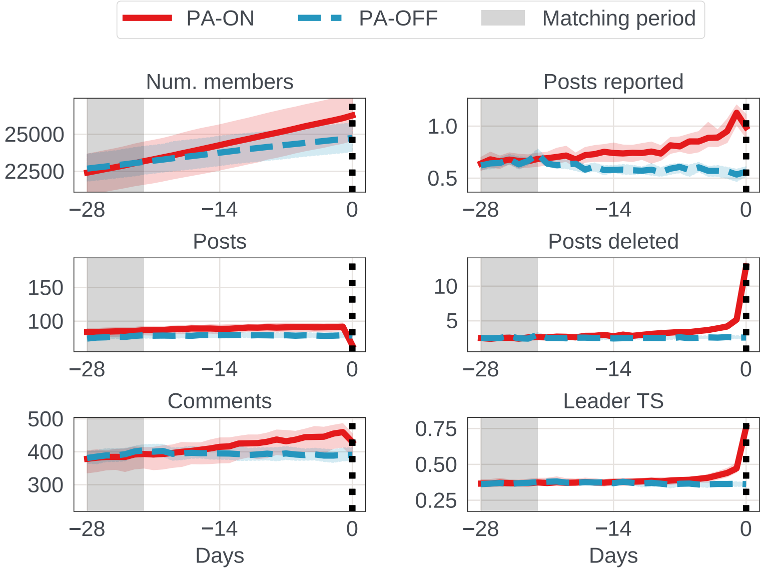
Between March 28, 2021, and July 11, 2021, we collected data on 1) communities that turned on post approvals and did not change other moderation-related settings (PA-ON; $n$ = 8,767); and 2) a random sample (50%) of communities that did not change *any* moderation-related setting (PA-OFF; $n$ = 224,635). For PA-ON groups, we considered only communities that enabled post approvals at least 28 days after the start of the study period and at least 28 days before its end. For both PA-ON and PA-OFF groups, we considered only communities with 128 or more members and at least one comment and one post over any 7-day window.

We analyzed PA-ON communities relative to when they turned on post approvals, referring to the day when they enabled the setting as day 0. For PA-OFF communities, we randomly assigned a pseudo-intervention date drawn from the distribution of dates (day and hour) when PA-ON groups enabled post approvals (*cf.* Appendix A.3.3; Fig. A.5). We considered the set of variables described in Table 7.1 in the 28 days before and after each intervention, for a total of 57 days (from −28 to 28). Some variables are 1) time-dependent, capturing group activity and group moderation (*e.g.*, number of posts, number of posts deleted), while others are 2) time-invariant, capturing group topic, demographics, and moderation settings (*e.g.*, group visibility, group category, if a group was a buy-and-sell group, *etc.*).

All data was de-identified and analyzed in aggregate, and no individual-level data was viewed by the researchers. In the analyses that follow, variables were 95%-winsorized (*i.e.*, the 2.5% smallest and largest values were replaced with the most extreme remaining values [287]) prior to aggregation unless otherwise stated. This ensured that trends/effects were not dominated by a few large groups. Nonetheless, results were qualitatively similar without winsorization. (Data was provided by Facebook/Meta, and is, unfortunately, unavailable for reproducibility.)



(a) All



(b) Matched

Figure 7.2: Average values for user activity- and moderation-related variables in the four weeks before communities enabled post approvals. Values for communities that enabled post approvals (PA-ON) are in red and those for communities that did not (PA-OFF) are in blue. For PA-OFF communities, day 0 corresponds to a pseudo-intervention date selected at random. We show trends for all communities in our dataset in *(a)* (PA-ON $n = 8{,}767$; PA-OFF $n = 224{,}635$) and for matched pairs of communities in *(b)* (PA-ON/PA-OFF $n = 8{,}643$). The period when matching was done is marked in gray. Error bars represent 95% CIs.

| | |
|---|---|
| **Group characteristics** (not time-dependent) | |
| **Visibility**$^\star$ | Whether group is private or public. |
| **Join approvals**$^\star$ | Whether leaders have to manually approve new members. |
| **Average age**$^\star$ | Average age of the members in the group. |
| **% women**$^\star$ | The percentage of women in the group. |
| **Buy-&-Sell**$^\star$ | Whether the group is a buy and sell group or not (specified by admin). |
| **Group categories**$^\star$ | Lexical categories obtained from the groups' description and title using Empath [63]. See Appendix A.3.2 for details. |
| **Moderation-related** | |
| **Moderating TS** | Average time leaders spent in moderation-related interfaces (e.g., approving posts). |
| **Leader TS** | Average time leaders spent in the group. |
| **Members Removed** | Number of members removed. |
| **Posts deleted** | Number of posts by regular members deleted by leaders. |
| **Posts reported** | Number of posts reported in the community by users. |
| **Num leaders** | Number of leaders in the community. |
| **Activity-related** | |
| **Posts** | Number of posts. |
| **Comments** | Number of comments. |
| **Time spent** | Total time users spent browsing posts in the group (in hours). |
| **Reactions** | Number of Likes and of other reactions (Sad, Happy, Wow, Laugh, Angry). |
| **Num members** | Number of members in the community. |

Table 7.1: Description of the group-level variables considered in this chapter. Variables marked with a star ($\star$) were measured on the day prior to the intervention (for PA-ON groups) or the pseudo-intervention (for PA-OFF groups). They were not analyzed in the result sections of this chapter, but were used in the matching to ensure the two sets of communities were comparable (*cf.*, Appendix A.3.1).

## 7.4 Results

### 7.4.1 What Leads to the Adoption of Post Approvals?

This subsection examines *why* communities adopt post approvals to begin with (**RQ1**). We focus on what happens *before* the setting was enabled, contrasting PA-ON and PA-OFF groups. All analyses in this section were done at the group level, and each group was given equal weight.

**Case–control analysis.** Our first analysis follows a case–control design [230]: we compared user activity and moderation traces of groups that enabled post approvals (PA-ON; the "case") with those that did not change any moderation settings (PA-OFF; the "control"). We considered three variables related to user activity (*Num Members*, *Posts*, and *Comments*) and three variables related to moderation activity (*Posts reported*, *Posts deleted*, *Leader TS*) in the 28 days before the intervention (*cf.* Table 7.1 for descriptions). We refer to this scenario as "all" since, in what follows, we examine a subset of this data corresponding to matched pairs of PA-ON and PA-OFF groups.

Fig. 7.2a shows the average value of each of the variables mentioned above. We found significant differences between PA-ON and PA-OFF groups that are consistent across the 28-day period considered ($p < 10^{-4}$ for independent t-tests conducted each day). PA-ON groups have significantly more reported and deleted posts, more comments, and fewer posts than PA-OFF groups. Leaders also spent more time in PA-ON groups than in PA-OFF groups.

Temporal trends also differed significantly between the two sets of groups: PA-ON groups experienced larger increases in all considered variables in the weeks before enabling post approvals. For moderation-related metrics, we observed a sharp spike on the day (or in the case of posts reported, on the day before) post approvals was turned on. These changes were not observed in PA-OFF groups.

The shifts in user activity (*e.g.*, *Comments*) and in moderation (*e.g.*, *Posts deleted*) before post approvals was turned on suggests that leaders enable the setting in response to new (and perhaps more chaotic) group dynamics. Specifically, the setting was commonly enabled in groups that were quickly growing and that experienced a surge in moderation-related events, which may have been the final straw that led administrators to enable post approvals. This finding is consistent with prior work suggesting that major changes in moderation (*e.g.*, changing settings, creating new rules) happen in reaction to problems that emerge [237].

**Matched analysis.** While indicative, the previous analysis conflates two factors. Not only do PA-ON and PA-OFF communities differ in their baseline user and moderation activity, but they also differ in the way the studied variables change over time. Thus, observed differences in temporal trends may come from the fact that groups that adopt post approvals are different from those that do not. To more fairly compare these communities, we matched PA-ON

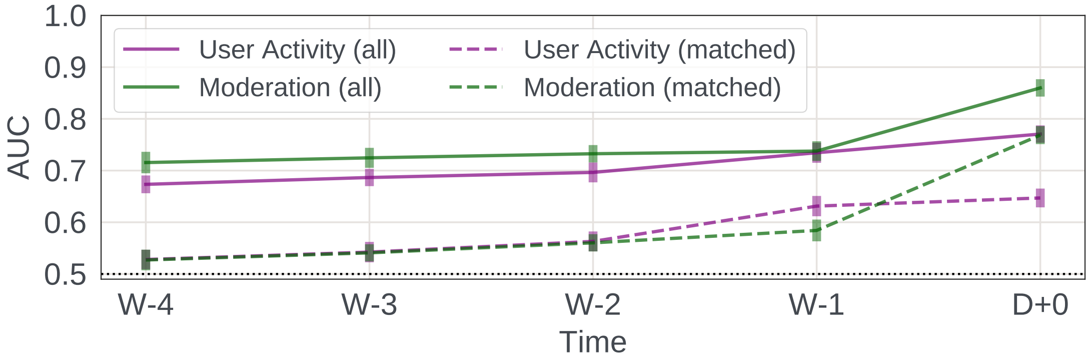


Figure 7.3: AUCs for classifiers trained to distinguish PA-ON and PA-OFF groups using data from different time spans and different sets of features. Error bars represent 95% CIs obtained through a 20-fold cross validation.

and PA-OFF groups on user activity and moderation-related metrics between days −28 and −22. This matching ensures that communities were similar in the first week of the study period. Specifically, we performed one-to-one propensity score matching of PA-ON and PA-OFF communities using moderation and activity-related variables, as well as general group characteristics (*e.g.*, *Group categories*). Details of the matching procedure can be found in Appendix A.3.1.

After performing this matching, we repeated the same analysis as in the previous subsection (Fig. 7.2b). Again, we found that PA-ON groups experienced a gradual increase in moderation-related traces which was accentuated right before post approvals was turned on — changes that were not observed in PA-OFF groups. Differences in user activity were subtler. Both PA-ON and PA-OFF groups experienced growth in the number of members, but this growth was higher for PA-ON groups. Moreover, PA-ON groups experienced a significant increase in the number of daily comments received, while comments received remained largely unchanged in PA-OFF groups. Last, in both PA-ON and PA-OFF groups the number of posts increased slightly during the 28 days considered.

Overall, the matched case–control analysis confirms that there are differences in the temporal trends of moderation and user activity of PA-ON and PA-OFF communities. Even when considering communities that were initially similar, PA-ON communities experienced larger increases in user and moderation activity prior to the day when they turned on post approvals.

**Predicting if post approvals will be turned on in future weeks.** While findings thus far indicate that both changes in user activity and in moderation precede the use of post approvals, is one a stronger indicator than the other? And how far in advance might they predict the adoption of the setting? To answer these questions, we examined if user activity or moderation can be used to distinguish PA-ON groups from PA-OFF groups.

We created two balanced samples of groups. The first sample (*all*) comprised 10,000 groups — half PA-ON and half PA-OFF. The second (*matched*) comprised the same PA-ON groups as in the *all* sample, while corresponding PA-OFF groups were obtained using one-to-one propensity score matching previously described. We considered two sets of features (group activity-related and moderation-related; *cf.* Table 7.1) and five time spans,[I] calculating the value of each feature in each time span by taking its average.

We trained Gradient Boosting classifiers to distinguish PA-ON and PA-OFF groups, varying the feature set and the time span used. Fig. 7.3 shows the AUC of the classifiers trained in each of the different settings (all *vs.* matched; moderation *vs.* activity) considering the features up to the time specified on the $x$-axis. For instance, the points shown above $x$ = W-3 correspond to the AUC of classifiers trained with features associated with W-4 and W-3.

For the *all* sample, we found that moderation features were more predictive in earlier weeks (W-4 to W-2). However, for $x$ = W-1, there was an increase in the AUC of the classifier trained with activity features. We observed a similar pattern in the *matched* sample: classifiers started with similar AUC values at $x$ = W-4 (user activity: 0.53 AUC *vs.* moderation: 0.53 AUC), but the classifiers trained with activity features saw a larger increase in performance at $x$ = W-1 (0.58 for activity *vs.* 0.63 for moderation). Including data up until the time of intervention, $x$ = D+0, the performance of classifiers trained with moderation features increased sharply (*e.g.*, in the *matched* sample: 0.65 activity *vs.* 0.77 moderation). Overall, these results suggest that the adoption of post approvals was associated with gradual changes in user activity in the weeks before the adoption of the setting and sudden changes in moderation activity on the day the setting was enabled. Repeating this analysis but instead training classifiers using features belonging to each individual time span (e.g., using only W-1 vs. using W-4 to W-1 for prediction) resulted in qualitatively similar findings.

### 7.4.2 How do Post Approvals Shape Online Communities?

Having explored changes in user activity- and moderation-related signals that *precede* the adoption of post approvals, we now turn our attention to what happens *after* communities choose to adopt the setting. Here, we examine how user and moderation activity in online communities change following the adoption of the setting (**RQ2**). To do so, we matched communities that turned on post approvals (PA-ON) with similar communities that did not (PA-OFF) and validated the observed differences using regression. To obtain this matching, we performed one-to-one propensity score matching on moderation and activity-related variables as well as general group characteristics. Matching was done across the entire pre-intervention period (day −28 up to day 0 right before the intervention). See Appendix A for details. All analyses in this section were done at the group level, with each group weighted equally.

[I] Days -28 to -22 (week -4; W-4), -21 to -15 (week -3; W-3), -14 to -8 (week -2; W-2), -7 to -1 (week -1; W-1), and day 0 (D+0).

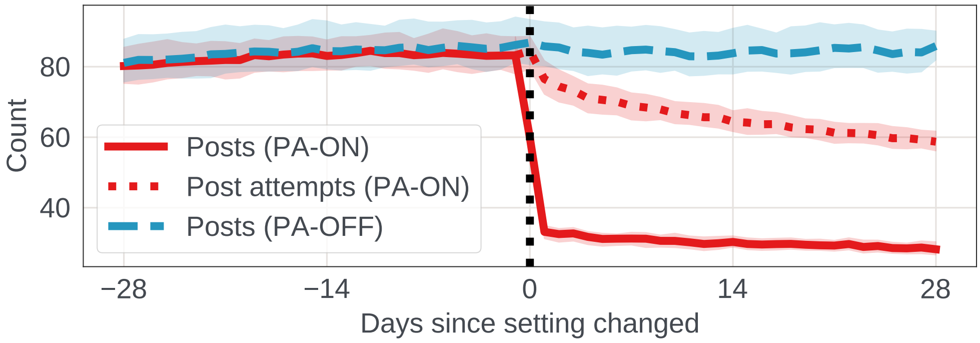


Figure 7.4: The average number of posts in PA-ON (solid red) and PA-OFF (dashed blue) communities. For PA-ON communities, the average number of post attempts after post approvals was enabled is shown in dotted red. Error bars represent bootstrapped 95% CIs.

**Posts.** First, we examined how posting behavior changed after the adoption of post approvals. In Fig. 7.4, we show both the number of posts that were actually published (*Posts*), but also, for PA-ON communities, the number of post attempts, i.e., post requests initiated by regular group members following the adoption of post approvals. For PA-ON communities, we found a significant decrease in the number of posts following the adoption of post approvals. The average number of posts went from roughly 80 posts a day pre-intervention to around 30 posts a day post-intervention, a decrease not observed in the matched set of PA-OFF communities.

What explains this decrease in posting? Was it because posts were being filtered? Or were people more hesitant to even post? To understand the relative contribution of these factors, we examined two corresponding quantities that make up the decrease in posting: 1) the difference between the number of posts in the control (PA-OFF) and treatment setting (PA-ON) (blue line vs. dotted red line); and 2) the difference between the number of post attempts and actual posts in PA-ON communities (dotted red line vs. solid red line). We observed a gradual decrease in the average number of posts submitted (the first component mentioned above), from 76 posts a day on day 1 to 58 posts a day on day 28. However, the fraction of posts approved per community (the second component) remained largely stable at around 63% of posts (note that this was calculated without winsorization and per group, instead of dividing the overall averages; *cf.* Fig. A.4, Appendix A.3.3). These findings suggest that post approvals reduce the number of posts by directly filtering out undesired posts, but also by reducing the likelihood of people to attempt to post in the first place.

**Other user activity-related signals.** Second, we looked at other user activity-related metrics (*Comments*, *Reactions*, and *Time spent*). These are shown in Fig. 7.5 both in absolute terms (first column) and normalized per number of posts (second column). Comparing PA-ON with PA-OFF communities, there was an absolute *decrease* but relative *increase* in all of these user activity metrics for PA-ON communities following the enabling of post approvals. In other words, there were fewer posts, but each post received, on average, more comments, reactions, and time spent.

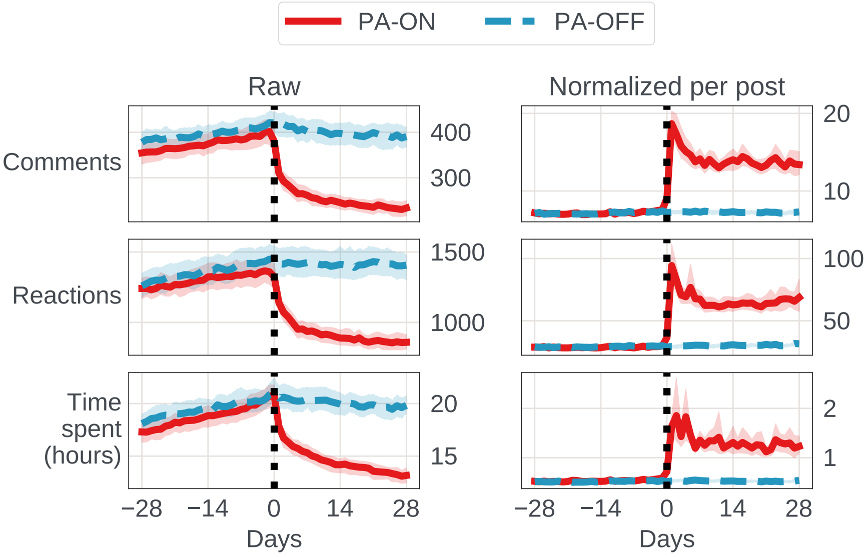


Figure 7.5: User activity-related signals before and after the adoption of the post approval setting. Signals are shown both in absolute terms (left) and normalized per number of posts (right). Error bars represent bootstrapped 95% CIs.

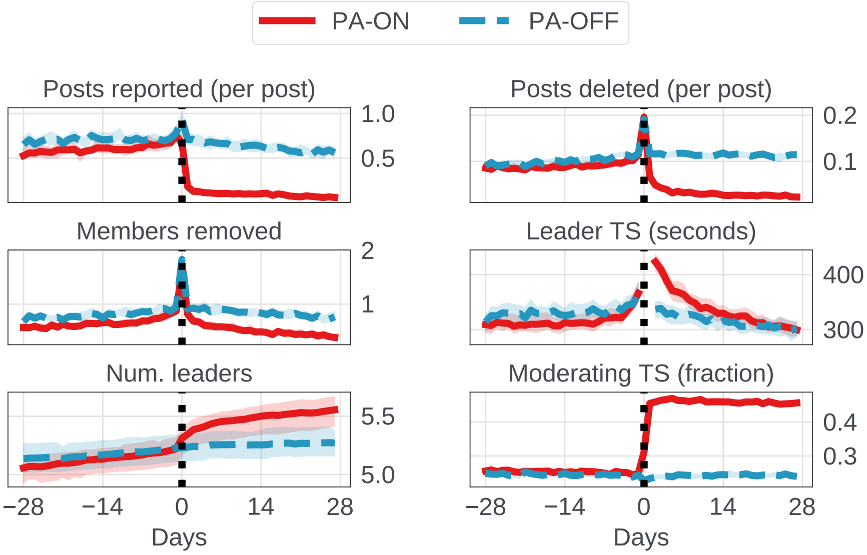


Figure 7.6: Moderation-related signals before and after the adoption of post approvals. Error bars represent 95% CI. We omit days 0 and 1 from the plot showing the *Leader TS*, as it contains a sharp peak.

For instance, before the intervention (day $-1$), PA-ON and PA-OFF communities received an average of around 402 and 421 daily comments and around 7.6 and 7.4 comments per post. (Note that although the two sets of communities are matched, their averages are not perfectly identical.) After day 0, when PA-ON communities enabled post approvals, the number of daily comments declined substantially, reaching an average of 233 daily comments on day 28. Meanwhile, the number of comments per post nearly doubled to around 13.4. This change was not observed in the matched PA-OFF groups.

**Moderation-related signals.** Third, in Fig. 7.6, we examined moderation-related metrics – *Posts reported*, *Posts deleted*, *Leader TS* and *Moderating TS* (*cf.* Table 7.1 for descriptions). We normalized the number of posts reported and deleted by the total number of posts, and moderating time spent by the total leader time spent. Recall that the moderating time spent encompasses activities such as responding to reported content and approving posts.

Following day 0, PA-ON groups had fewer members removed and fewer posts reported/deleted (per post) than PA-OFF communities. For example, the percentage of reported posts decreased from around 0.75% Around the time post approvals was enabled for PA-ON groups, time spent per admin increased substantially, likely because leaders were getting used to the new moderation style. However, by week 4, time spent per leader in PA-ON groups had returned to pre-intervention levels, though PA-ON groups had also tended to appoint new leaders. Examining the fraction of time spent in admin surfaces, leaders went from spending around 25% of their time using group moderation tools to around 45%, suggesting that leaders spent a substantial fraction of their time approving posts. While this increase appears large, group leaders in groups without post approvals may nonetheless informally vet posts by browsing posts in a group as they appear.

**Regression analysis.** Previously, we compared the user and moderation signals between PA-ON and PA-OFF communities before and after the post approvals are turned on after matching. Here, we performed a more rigorous analysis of the same signals under a regression framework.

We considered the average value of each variable of interest in week $-4$ (days $-28$ to $-22$) and week 4 (days 22 to 28). Then, for the variables in the post-intervention period, we estimate the impact of adopting post approvals using a linear model:

$$y = \alpha \mathbf{X} + \beta \mathbf{1}_{[\text{PA-ON = True}]}, \tag{7.1}$$

where $y$ represents the average value of one of the variables we studied in week 4 (*i.e.*, after the intervention), *e.g.* post approvals, $\mathbf{X}$ represents an array with all the variables we did the matching with in week $-4$ (*i.e.*, before the intervention), and $\beta$ represents the coefficient associated with turning on post approvals, as it multiplies an indicator variable that equals 1 for PA-ON, and 0 for PA-OFF, communities.

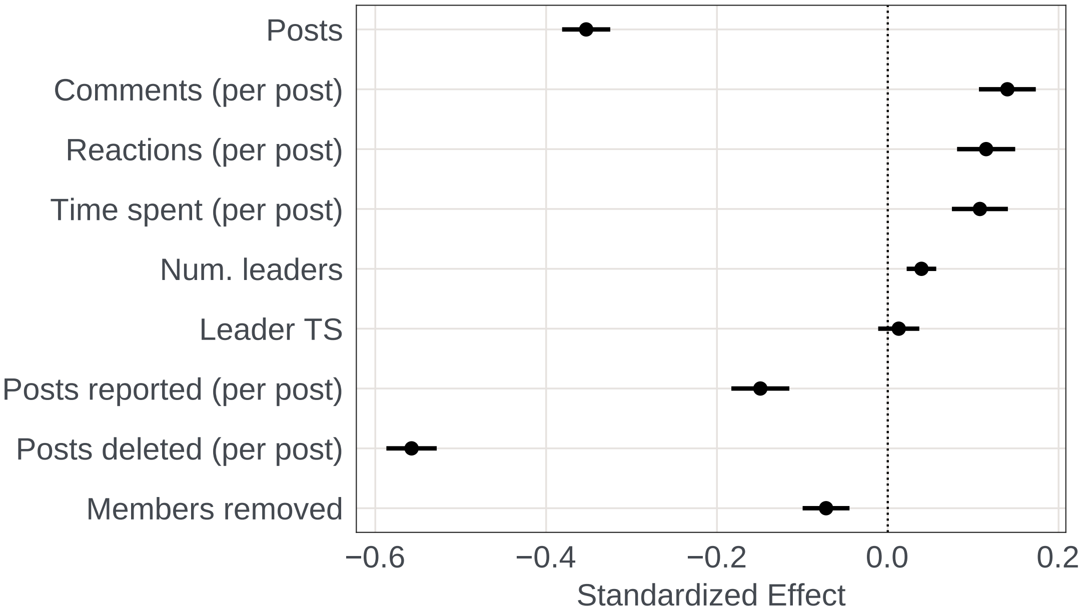


Figure 7.7: Standardized effect of enabling post approvals on user activity- and moderation-related variables. Error bars represent 95% confidence intervals. Data was not winsorized prior to this analysis.

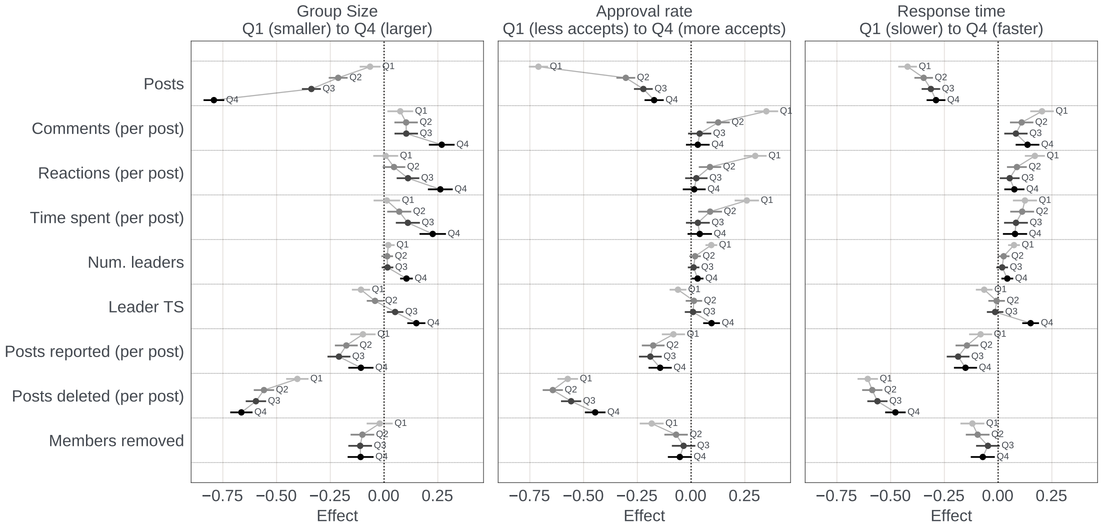


Figure 7.8: Effects of enabling post approvals for different stratifications of the data. Here, we report quartile-specific effects for group size (*i.e.*, number of members in a group; first column), the post approval rate (*i.e.*, percentage of posts that get approved in a group; second column) and the response time (*i.e.*, average time taken to accept posts in a group; third column). Data was not winsorized prior to this analysis. Error bars represent 95% CIs.

To facilitate interpretability, we standardized the dependent variable $y$, so that the coefficients represent (pooled) standard deviations. The coefficient $\beta$ captures the difference between our treatment and control groups in the matched setting. Since the coefficient is associated with an indicator variable, the effects reported represent the differences between PA-ON and PA-OFF groups in standard deviations. We report $\beta$ for all outcomes of interest in Fig. 7.7.

This analysis largely confirms results shown in Fig. 7.4, Fig. 7.5, and Fig. 7.6. The use of post approvals was significantly associated with a reduction in the number of posts ($-0.35$ standard deviations) and an increase in the number of comments, reactions, and time spent per post. Use of the setting was also associated with a decrease in the number of posts reported per post ($-0.15$ SDs), posts deleted per post ($-0.55$ SDs) and number of members removed ($-0.07$ SDs). Taken along with the previous analyses, these results suggest that the setting improves the quality of posts. Further, post approvals do not significantly increase the average time leaders spend in the group, although groups that enable the setting tend to increase the size of their leadership team (around 0.04 SDs).

### 7.4.3 Heterogeneity of Post Approvals

While post approvals change how online communities function, the effect of the setting may vary by group size as well as how it is used. For example, we found that, on average, community leaders do not spend more time in their communities following the adoption of the setting. Yet, this may not be the case for all groups: very large groups (with possibly hundreds of daily post attempts) may actually require more time from leaders after the setting is turned on, while smaller groups may require less time. Thus, we analyzed the impact of post approvals in communities with 1) different member counts, as well as in communities that 2) approved different fractions of the posts submitted (approval rate); and 3) took a different amount of time to approve posts (response time).

For each of the aforementioned variables (number of members; response time and approval rate), we divided PA-ON communities into four quartiles.[II] Then, we used the same regression setup depicted in Equation (7.1), but estimated the effect of post approvals separately for communities in each of the quartiles. This amounts to a linear model of the form:

$$y = \alpha \mathbf{X} + \sum_{i=1}^{4} \beta_i \, \mathbf{1}_{[\text{PA-ON = True and Quartile = } i]}, \tag{7.2}$$

where $\beta_i$ is the effect for groups in a given quartile. We ran a different regression for each of the three setups described above (group size, approval rate, and response time). All three factors had noteworthy interactions with user activity- and moderation-related signals (*cf.* Fig. 7.8).

[II] Defined by three points: for approval rate: 0.45, 0.69, 0.84; for response time: 1.2, 2.7, 5.75 (hours); for group size: 2850, 8500, 24000 (rounded) Approval rate and response time were measured across the entire post-intervention period. Group size was measured on day 0.

**Time spent by leaders.** Leaders in larger groups (group size Q4), groups with lower approval rates (approval rate Q4), and groups with faster response time (response time Q4) spent more time in their communities following the adoption of post approvals (*Leader TS*: 0.15; 0.10; and 0.15 SDs). These trends were gradual across quartiles, and contrast with the overall null effect reported in Fig. 7.7. In other words, the moderation burden after enabling post approvals depends on the kind of group and on how community leaders proactively moderate the community.

**User activity.** Larger groups experienced larger decreases in the number of posts (*e.g.*, Q4: −0.79 SDs) and larger increases in relative activity (*e.g.*, 0.22 SDs for *Time spent* per post). Groups with a higher approval rate (Q3/Q4) experienced smaller decreases in the number of posts and smaller increases in relative activity. The higher the approval rate, the smaller the deviations were from PA-OFF matched communities. To a lesser extent, this was also observed for response time: the faster the response time, the smaller the deviations were. These results suggest that activity-related changes were greater in larger communities and that the strictness and speed of community leaders in approving or rejecting posts mediated changes in user activity.

**Moderation.** Regardless of group size, response time, or approval rate, the number of posts reported and deleted decreased significantly across quartiles in all three analyses. Other moderation-related metrics such as members removed also decreased in most cases. Overall, this suggests that the decrease in potentially problematic content following the enabling of post approvals is robust and that it holds even when groups are very large (*e.g.* −0.11 SDs for *Posts reported* per post for group size Q4) or when a vast majority of posts are approved (*e.g.* −0.15 SDs for approval rate Q4). In Fig. 7.4, we saw that post approvals changed the number of posts through both behavior change and through the filtering done by community leaders. Here, we see evidence that, regardless of the strictness of this filtering, the number of posts reported and deleted (normalized per post) decreases.

## 7.5 Discussion

In this work, we presented a large study of post approvals in Facebook Groups, examining both their adoption and their subsequent impact. Post approvals was adopted after changes in the groups' dynamics in the weeks prior: user activity and moderation increased in the weeks before the setting is enabled, and, on the day when the setting was turned on, there was often a surge in moderation activity. After the setting is adopted, communities become, on average, centered around fewer posts that receive more comments and reactions, and which users interact with for longer. These posts were less likely to be reported, and members were less likely to be removed, which suggests an increase in the quality of the discussions happening in the group. However, the strength of these effects varied with group size and with how proactive moderation was carried out – *e.g.* in larger groups, leaders spent more time in the group after the setting was enabled, while in smaller groups, they spent less time.

Overall, the findings provide preliminary insight on how proactive moderation may improve online information ecosystems: by adding participation friction in online communities, post approvals elicit behavior change (*cf.* Fig. 7.4).

A limitation of our work is that we focus on a limited set of community-level analyses over a short period without considering spillover effects. This limitation suggests several potential extensions. First, future work could analyze how post approvals impact the participation of different kinds of users (including leaders); for instance, the setting may disproportionately affect highly-active users or may affect newcomers more than veteran members of a group, discouraging the former from participating. Related, future work could examine the reasons for the decrease in post attempts – to what extent do post approvals discourage lower-quality posts vs. all posts? Second, future work could investigate the impact of post approvals in the long run. In other words, do the patterns we observe here continue for months or even years? How do communities evolve with and without the setting? Third, future work might examine spillover effects across different communities. If two communities have many overlapping members and one adopts post approvals (as well as stricter moderation practices), does this influence the behavior of users in the other community which did not adopt the setting? Fourth, future work could explore other community-level variables, such as within-group friendship network properties. This work may include examining if these variables can help explain the adoption of post approvals (as in Sec. 7.4.1), if they change suddenly after post approvals are enabled (Sec. 7.4.2), or if the effect of post approvals is heterogeneous across these variables (Sec. 7.4.3).

Last, we note that our analysis was limited to Facebook Groups. Adapting the methodology here to explore participation controls studied qualitatively in other platforms such as “chat modes” in live streaming platforms [236] or software such as Reddit’s *AutoModerator* [116] remains future work. As argued by Kraut (2012) [130], participation controls are “design levers” that shape how people connect with others in online communities. Thus, understanding how they work may inform the design of better-governed online spaces.

# 8 Guiding content creation

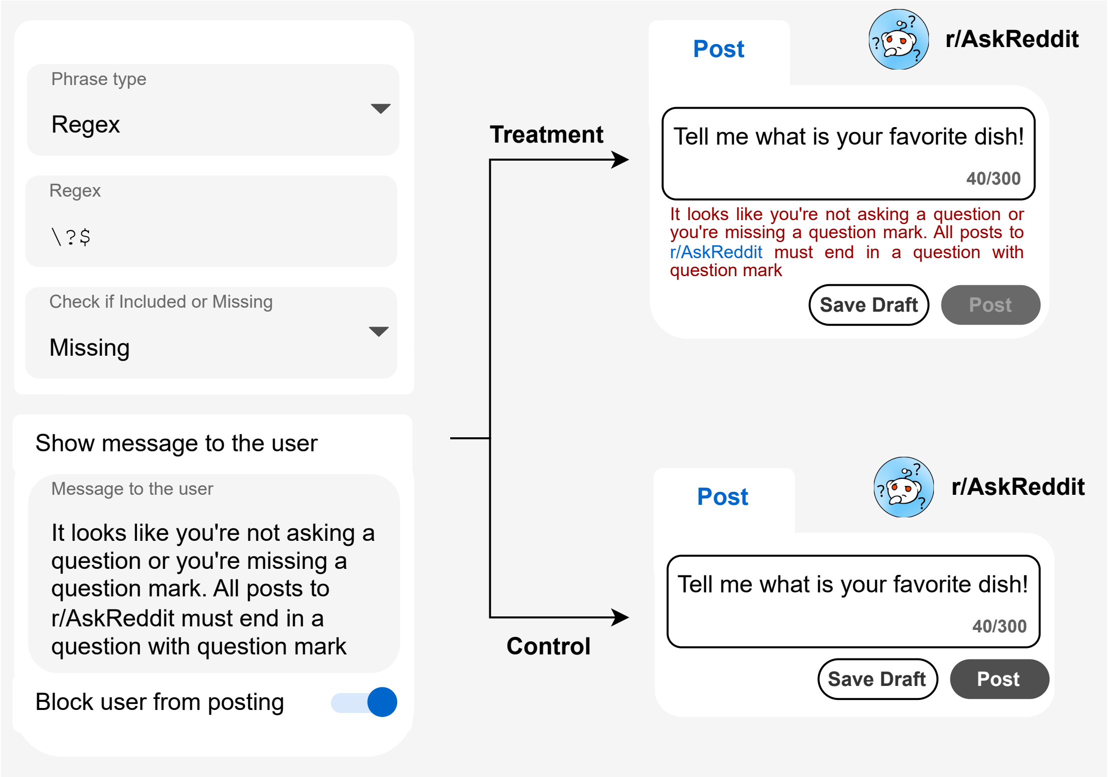
**This chapter was the result of a collaboration with Reddit.**

## 8.1 Introduction

Rules and norms are crucial for online communities to thrive [127, 64, 33] and achieve their varied goals [281, 235, 224, 156]. In *r/relationshipadvice*, for instance, a community on Reddit where "users can request others' opinions on a specific situation between two people," all posts must include the age and gender of the two referenced people using a stylized format, (*e.g.*, "I [*24M*] have an issue with my aunt [*45F*]." ) Posts breaching this guideline will likely be removed either manually after review by the moderators of *r/relationshipadvice* or automatically, through programmable moderation tooling, such as Reddit's 'AutoModerator.'

Moderators in communities such as *r/relationshipadvice*, which have many contributions, must often decide whether to rely more heavily on automated tooling, which can scalably capture submissions that might deviate from the goals of the community, or to invest more time in manual moderation, making precise judgments about which content should or should not be accepted [237]. For instance, moderators might configure automated tools to remove all posts with keywords that correlate with some rule-breaking behavior; however, they then risk removing posts unfairly, leading disgruntled users to leave the community [116]. In contrast, moderators may consider each contribution manually and become overwhelmed by the large and constant task of evaluating content [232, 155]. Thus, there is a need for solutions that combine the scalability of automation with the nuanced understanding that humans can bring to evaluate content.

A third promising direction focuses on shifting the knowledge and work associated with evaluating content from moderators to the contributors themselves to help them better understand what content will or will not be accepted within the community. Prior research, for example, has explored simple 'nudge'-like approaches, such as making rules salient [155] or asking users to reconsider posts that may breach community guidelines [122]. These approaches are *proactive* and *user-centric*, guiding *users* in how they should contribute to the



Phrase type
Regex

Regex
\?$

Check if Included or Missing
Missing

Show message to the user

Message to the user
It looks like you're not asking a question or you're missing a question mark. All posts to r/AskReddit must end in a question with question mark

Block user from posting

Tell me what is your favorite dish!
40/300

It looks like you're not asking a question or you're missing a question mark. All posts to r/AskReddit must end in a question with question mark

Tell me what is your favorite dish!
40/300

Figure 8.1: Depiction of how Post Guidance works. When users try to write posts in a community, their contribution attempts are matched against a set of rules configured by the moderators of that community (left). Posts can only be submitted if they fulfill these rules, e.g., in the image, the post cannot be submitted as Post Guidance requires posts to end with a question mark (top right). Here, we consider a large, user-level experiment ($n_{\text{users}}$=97,616) where users were either exposed to this feature ('treatment;' top right) or not ('control;' bottom right) across 33 communities on Reddit.

community *before* they break any rules, but also universal, in that this guidance does not adapt to the community or specific content. Other research has hypothesized that personalized guidance following moderation decisions may help encourage future contributions [112, 113]. These approaches are *community-specific*, offering users the opportunity to improve their skills at evaluating what content will be accepted within a given community, but reactive and still requiring time-consuming manual intervention from moderators.

### 8.1.1 Post Guidance

In this paper, we evaluate *Post Guidance*, a new approach to moderating online communities that allows moderators to create rules that trigger interventions as users draft posts in an online community. Post Guidance is a paradigm for moderating user-generated content where:

- *Users* draft and edit contributions within online platforms (*e.g.*, posts, videos, or images). We refer to these contributions in the making as “post drafts;”
- *Platforms* implement “interventions” to prevent rule-breaking behavior. These can be customizable, *e.g.*, sending custom messages.
- *Moderators* create “rules” that, upon changes to the “post draft” (or attempts to submit it), trigger “interventions” if certain conditions are met.

While the overall approach is applicable across a broad range of platforms, rules, and interventions, Fig. 8.1 illustrates the implementation of Post Guidance on Reddit evaluated in this paper. In *r/AskReddit*, a prominent Q&A community, all post titles must end with a question mark. Post Guidance can enforce this with a rule requiring the regular expression `"\?$"` to match every post title. As users edit their post draft, Post Guidance prevents the submission of the post and warns the user until the rule is met.

Despite its simplicity, Post Guidance differs substantially from other content moderation interventions extant in large social media platforms; it is simultaneously *community-specific*, *proactive*, and *user-centric*. This approach offers the potential to improve adherence to rules and encourage contributions by educating individual users through a mechanism that actually reduces the effort required from moderators.

### 8.1.2 Hypotheses

In many communities on Reddit, moderators remove most or all rule-breaking posts, either through automated filtering or manual removal. The contextually relevant guidance provided by Post Guidance may lead some users to adapt their contributions to match the rules of the community, such that a subset of initially rule-breaking posts might avoid removal; thus, we hypothesize that:

*Post Guidance increases the number of non-removed contributions to communities adopting it.* (**H1**)

Post Guidance will also prevent the submission of some subset of rule-breaking posts that would otherwise have required manual evaluation by moderators (either directly in the feed or within the AutoModerator queue); thus, we hypothesize that:

*Post Guidance decreases the workload for moderators in communities adopting it.* (**H2**)

The first two hypotheses capture the "direct effects" of Post Guidance, but the feature may also indirectly shape communities. First, by ensuring that more posts align with the norms and goals of the community, it is possible that Post Guidance could contribute to an overall increase in a community-specified notion of post "quality", as measured by the degree to which other members engage with posts through votes and comments; thus, we hypothesize that:

*Post Guidance increases the quality of contributions to communities adopting it.* (**H3**)

Finally, it is possible that by helping more users successfully post and receive positive feedback from others within their communities, Post Guidance might help improve users' (especially newcomers') experience in and connection to those communities, as reflected in their frequency of visits and participation in the community; thus, we hypothesize that:

*Post Guidance increases users' engagement (other than posting) to communities adopting it.* (**H4**)

### 8.1.3 Experimental Setup

To understand how Post Guidance shapes online communities, we conducted a large-scale field experiment with 33 subreddits. Subreddits that opted-in to participate were onboarded onto the feature to ensure both familiarity with the tool and the creation of rule-based post constraints. Over 63 days, 97,616 users who started drafting at least one post in any of these subreddits were assigned randomly into treatment and control groups, enabling us to make within-subreddit comparisons. Using behavioral logs of the 28 days after enrollment, we calculated aggregated measures of (1) the users' posting activity, (2) moderation received on those posts, (3) community engagement with those posts, and (4) the users' overall engagement in the subreddit. This randomized setup allows us to analyze the causal effects of Post Guidance using simple regression analysis.

### 8.1.4 Summary of Results

Our experiment offered strong evidence in favor of **H1**, **H2**, and **H3** and against **H4**. Post Guidance:

**H1** Increased the number of "successful" contributions, which we operationalize as posts not removed 72 hours after being submitted (+5.8% relative increase; $p$=0.03). Curiously, the feature reduced the number of posts started (−5.7%; $p$<0.001) and submitted (−13%; $p$<0.001), but posts created with Post Guidance were removed less often, which explains the increase.

**H2** Decreased moderation workload: users exposed to the feature had their posts reported less often (−9.4%; $p$=0.01) and removed by the AutoModerator less often (−34.9%; $p$=0.01). Note that moderators spend time reviewing both reported and automatically removed posts. Post Guidance did not significantly change the number of mod-removed posts (+2.7%; $p$=0.236).

**H3** Increased the quality of contributions. As previously discussed, posts created with the feature were reported and removed less. But they also received more comments (+28.6%; $p$=0.004), screen views (+26.6%; $p$=0.027), and more upvotes (+36.1%; $p$=0.004), indicating that they abide by the subjective quality criteria in the community they were created.

**H4** Did not increase user participation. For users that engaged with the feature, we observed a small, not statistically significant *decrease* in the number of days active (−5%; $p$=0.04) as well as votes (−5%; $p$=0.04) and comments to the community (-5%; $p$=0.04).

Subsequent secondary analyses revealed that the effect of Post Guidance was similar across Reddit veterans and newcomers. Additionally, subreddits that relied heavily on AutoModerator (*i.e.*, automated, reactive content moderation) before the experiment and those that set up many Post Guidance rules saw the biggest increases in the number of 'successful contributions.' This suggests that extensive use of the feature makes a difference and that going from reactive to proactive content moderation may yield considerable advantages to online communities.

### 8.1.5 Implications and Future Directions

This study finds Post Guidance to be effective as a scaleable and flexible content moderation paradigm, with the potential to improve user-generated content across the Web. Though this approach is currently implemented for posts on Reddit, it would easily adapt to various contribution types and platforms, *e.g.*, comments on social media, direct messages in instant messaging platforms, and wiki contributions in collaborative encyclopedias. Moreover, the Post Guidance paradigm can easily be extended using machine learning models that allow the feature to (1) handle a broader set of rules (e.g., "Be kind" and others that are hard to map to specific textual patterns); (2) handle multi-media content; and (3) provide even more personalized feedback (e.g., suggest changes that would make contributions adhere better to community standards). In that context, future work could extend this new paradigm "breadth-wise," *i.e.*, adapting it to other parts of the Web, or "depth-wise," *i.e.*, increasing the possibilities for rules and nudges and quality of the feedback.

## 8.2 Background

Reddit comprises over 100,000 active communities (sometimes called *subreddits*) [204], where users can contribute submissions, comment on others' submissions, and upvote or downvote others' comments and submissions. These communities, independently created and managed by users called *moderators*, cover a broad range of interests and topics; as Reddit puts it, "whether you're into breaking news, sports, TV fan theories, or a never-ending stream of the internet's cutest animals, there's a community on Reddit for you" [207].

Activity within communities must abide by Reddit's platform-level Content Policy [203] and Moderator Code of Conduct [205], as well as the "Rediquette" values and practices that broadly apply to the platform [206], which are enforced by administrators (or *admins*), who are Reddit employees. Communities themselves create their own community-specific rules and norms, which are enforced by the community moderators [33, 64]. Community members can engage in distributed moderation by upvoting or downvoting content, controlling its visibility within the feed on a specific community. Reddit supports automated content moderation via its AutoModerator system, used extensively across the service, particularly in large subreddits [167]. Reddit also has a reputation system called 'karma,' calculated using upvote data [207].

## 8.3 Materials and Methods

We describe a large ($n$=97,616) randomized experiment to test the effectiveness of Post Guidance in improving community-level content moderation on Reddit. The experiment was conducted over a period of 63 days days on 33 subreddits.

**Pre-experimental setup.** Since Post Guidance is a community-level moderation strategy, the critical pre-experimental setup step in the study was to onboard subreddits. Post Guidance only works if there are rules in place, and therefore, we needed to give moderators access to the tool and encourage them to adopt it as part of their day-to-day moderation operations. Subreddits either self-enrolled after an announcement in a private subreddit gathering moderators of large communities or were invited to participate by three Reddit employees, who helped moderators in these communities familiarize themselves with the tool and, in some cases, optimize their rules. While enrollment to use the feature was continuous, we focused on the first 33 subreddits that adopted the feature and created at least two post-guidance rules.

**Assignment procedures.** All individuals on desktop (excluding users of 'old Reddit,' a pre-2018 version of the interface) who opened the interface to draft a post in one of the 33 enrolled subreddits were enrolled in the experiment. They were randomly assigned to either treatment or control with equal probability (meaning we have roughly 49,000 users in each condition; see Fig. 8.4b for the exact figures). Importantly, once a user is enrolled in a specific subreddit, we only consider the subsequent activity of that user in that subreddit, even though they will be sorted in the treatment or control group for all subreddits in the study; *e.g.*, for a user who first opened the interface to draft a post on *r/AskReddit* and was assigned to the treatment group, if they then went on to open the interface to draft a post on *r/healthyfoods*, they would also have Post Guidance enabled. However, we do not consider cross-community effects in the experiment at hand, as we find that the initial subreddits users visited when they were enrolled are responsible for most subsequent activity (*e.g.*, 97% of subsequent posts; 98% of subsequent daily activity).

**Start date and follow-up.** Enrollment started on 24 July 2023 and ended on 28 August 2023, and for each enrolled user, we considered a follow-up period of 28 days. In other words, we track users for 28 days after they first attempt to create a post in one of the 43 considered subreddits; e.g., if someone were enrolled on 28 August, we would track them until 25 September. In total, 97,616 users enrolled in the experiment. Fig. 8.2 depicts the number of enrolled users per day and the fraction of users enrolled per subreddit. Fig. 8.3 illustrates the experimental timeline; the enrollment period is shown with a red line along the main axis of the figure and grey lines above the main axis depict the tracking period of hypothetical users enrolling on different days.

**Outcomes.** We consider thirteen different outcomes, which we describe in detail in Table 8.1. In short, we consider variables related to the post creation flow (*e.g.*, Post starts), content moderation (*e.g.* AutoModerator removals), post engagement (*e.g.*, Received comments), and user activity (*e.g.*, Days active). We tie each outcome to one of our research hypotheses, with the exception of 'Number of reports,' which is associated with both **H2** and **H3**. We plot the distribution of the outcomes in Fig. A.9 (at the end of the paper).

**Statistical analyses.** Given that outcomes are heavy-tailed counts, we estimate the average treatment effect with a Poisson Regression model

$$\log \mathrm{E}(Y|Z) = \alpha + \beta \cdot Z, \tag{8.1}$$

where $Y$ is the outcome we care about, and $Z \in \{0,1\}$ is a binary indicator marking whether the user is in the treatment or the control group. Note that $\beta$ captures the log ratio between the treated and control groups, *i.e.*, $\beta = \log \frac{\mathrm{E}(Y|Z=1)}{\mathrm{E}(Y|Z=0)}$. Coefficients estimated with the Poisson Regression are *consistent* and *unbiased*, even if the dependent variable $Y$ is not Poisson-distributed [27]. Yet, Poisson Regression assumes that $\mathrm{E}(Y) = \mathrm{Var}(Y)$, which creates problems in estimating standard errors. We address this issue using a robust covariance matrix estimator (commonly known as *HC0* or Huber estimator), see [183].

**Heterogeneity of the effect.** We study how the effect varies depending on the user's and the subreddit's characteristics. To do so, we stratify the effect according to user- and subreddit-level variables. For example, assuming a dummy variable $X \in \{0,1\}$, we extend Eq. 8.1 to

$$\log \mathrm{E}(Y|Z,X) = \alpha + \beta \cdot Z + \eta \cdot X + \gamma \cdot Z \cdot X, \tag{8.2}$$

where $\gamma$, captures whether the effect is 'stronger' when $X$=1; it captures the ratio between the relative effect in units where $X$=1 and units where $X$=0, *i.e.*, $\gamma$=log$[\frac{\mathrm{E}(Y|Z=1,X=1)}{\mathrm{E}(Y|Z=0,X=1)} / \frac{\mathrm{E}(Y|Z=1,X=0)}{\mathrm{E}(Y|Z=0,X=0)}]$. We consider stratifying the effect along four distinct covariates, depicted in Table 8.1, capturing whether the user was a newcomer (A) or highly active (B), and whether the subreddit implemented many Post Guidance rules (C) or used AutoModerator substantially before the start of the experiment (D).

## 8.4 Results

Our experiment provide strong evidence in favor of **H1**, **H2**, and **H3** and against **H4**. We present our results in Table 8.2 and discuss them in further detail below.

### 8.4.1 H1: Effect of Post Guidance on Contribution Success

Post Guidance has significantly increased the number of non-removed contributions (+5.8% relative increase; $p$=0.03). This happens even though there was a significant *decrease* in post starts (−5.7%; $p$<0.001) and an even larger and significant decrease in the number of submitted posts (−13.0%; $p$<0.001). We further illustrate this finding in Fig. 8.4 (In Appendix A.4), plotting the effect size (left); and showing the percentage of (potential) contributions "lost" in each step of the posting pipeline (right). In the control group, more users start (85,421) and submit (53,500) posts compared to the treatment group (start: 80,593); submit: 46,522).[I]

### 8.4.2 H2: Effect of Post Guidance on Content Moderation

Contributions created using Post Guidance are more likely to remain in the platform because they are less likely to be removed. In particular, users in the treatment condition had their posts removed less often by the AutoModerator (−34.9%; $p$<0.001) and by Reddit administrators (−9.2%; $p$=0.03), besides being reported less often (−9.4%; $p$=0.001). Reviewing reports and AutoModerator removals is a substantial content moderation task, which implies that Post Guidance decreased the moderation workload. Interestingly, we did not find a significant decrease in removals by moderators. We conjecture that Post Guidance may, at the same time, (1) decrease rule-breaking posts from being created — *e.g.*, users might see that what they want to Post is not allowed and give up posting altogether; and (2) allow users to skirt rules, creating rule-breaking posts that the AutoModerator would otherwise catch — *e.g.*, upon seeing that their post break the rules, users might slightly alter their contribution in a way that leads to avoiding the AutoModerator, since the rules for Post Guidance and the AutoModerator are often similar. To explore this further, we re-ran the analysis, considering only users who submitted a post in the follow-up period, a scenario that isolates the second mechanism outlined above. We find that users in the treated group who submitted a post were significantly more likely to have their posts removed by moderators (17.7%; $p$<0.001), providing evidence that the two mechanisms conjectured above are at play.

[I] But more submitted posts are removed in the control group (56.5%; vs 47.3% in the treatment group).

Table 8.1: Description of variables considered in the study. We describe both main outcomes (#1—#13) and variables used to study the heterogeneity of the effect (A—D). For the former, we indicate the hypothesis they are tied to; for the latter, we indicate whether they were calculated at the user (User) or subreddit-level (SR).

| # | Name | Description | Hyp./Kind |
|---|---:|---|---:|
| 1 | Post starts | Number of times the user has entered the post creation interface in their assigned community. | **H1** |
| 2 | Post submitted | Number of times the user has submitted posts in their assigned community. | **H1** |
| 3 | Post non-removed | Number of posts made by the user in their assigned community that were not removed in the 24 hours after they were submitted. | **H1** |
| 4 | Automod removals | Number of times the AutoModerator has removed posts or comments from the user in their assigned community. | **H2** |
| 5 | Mod removals | Number of times a moderator has removed posts or comments from the user in their assigned community. | **H2** |
| 6 | Admin removals | Number of times an admin (a Reddit employee with Reddit-wide moderation capacities) has removed posts or comments from the user in their assigned community. | **H2** |
| 7 | Num. reports | Number of times other users reported posts by the user in their assigned community. | **H2**, **H3** |
| 8 | Rec. comments | Number of comments (from other users) in posts made by the user in their assigned community. | **H3** |
| 9 | Rec. screen views | Number of screen views (from other users) in posts made by the user in their assigned community. | **H3** |
| 10 | Rec. upvotes | Number of upvotes (from other users) received in posts made by the user in their assigned community. | **H3** |
| 11 | Days contributing | Numbers of days in the follow-up period (maximum 28) that the user has created a post or a comment in their assigned community. | **H4** |
| 12 | Days voting | Numbers of days in the follow-up period (maximum 28) that the user has up or downvoted a post or comment in their assigned community. | **H4** |
| 13 | Days active | Numbers of days in the follow-up period (maximum 28) that the user visited their assigned community. | **H4** |
| A | Newcomers | Whether the user visited the subreddit in the 90 days before enrolling in the experiment. | User |
| B | Low activity | Whether the user is in the bottom quartile of activity in the 90 days before enrolling in the experiment. | User |
| C | High rule count | Whether the subreddit created more than X rules in the 90 days before the start of the experiment. | SR |
| D | High automod-use | Whether the AutoModerator was triggered more than X times in the 90 days before the start of the experiment. | SR |

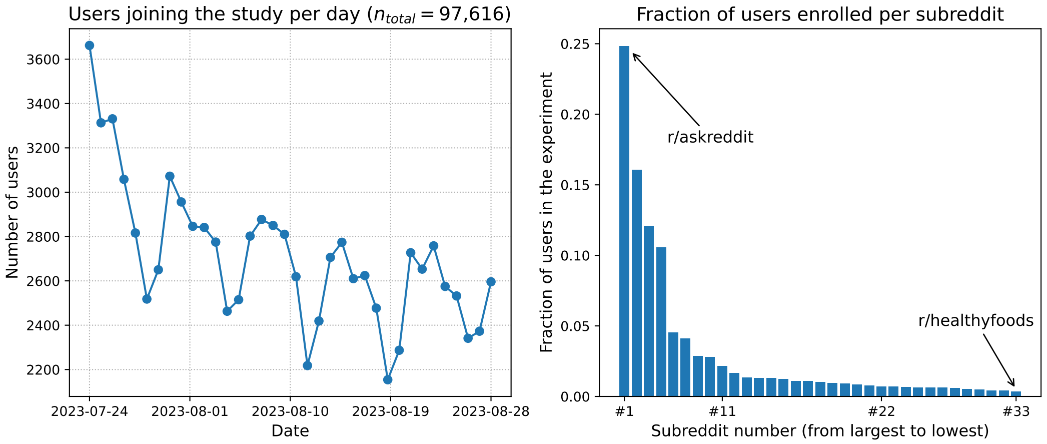


Figure 8.2: Details about experiment enrollment. On the left, we show the number of users enrolled in the experiment per day; there is seasonality (with lower enrollment during weekends) and a continuous drop in the number of enrolled users in the first few weeks (as users can only enroll once). On the right, we show the fraction of users in the experiment per subreddit considered; there is a long tail of subreddits with less than 1% of users in the experiment. Note that subreddits are sorted by the number of distinct users that entered the posting interface during the study period.

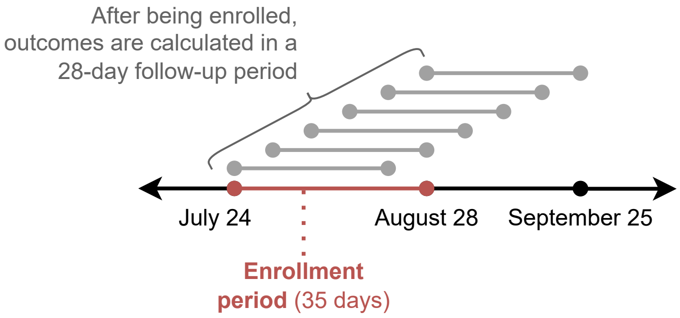
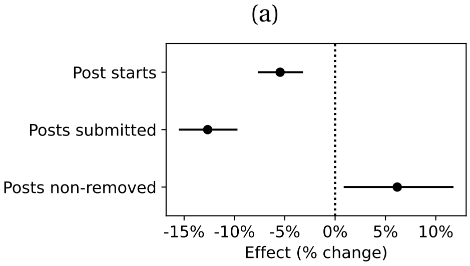


Figure 8.3: Timeline of the experiment. Users were enrolled in a 35-day period between 24 July 2023 and 28 August 2023. After enrolling, outcomes are calculated in a 28-day follow-up period. In the figure, we use gray lines to symbolize the tracking period of hypothetical users enrolling in different days.



(b)

| | Control | | Treatment | |
|---|---|---|---|---|
| Posts (...) | # | %Δ | # | %Δ |
| starts | 85421 | — | 80593 | — |
| submitted | 53500 | 37.4% | 46522 | 42.3% |
| non-removed | 23268 | 56.5% | 24527 | 47.3% |
| Total users | 48793 | | 48823 | |

Figure 8.4: Effect of Post Guidance on contribution success. (a) Post guidance significantly reduces the number of non-removed posts, even though fewer posts are started and submitted; (b) We further detail the fraction of (potential) posts lost at each step of the creation pipeline, indicating the number (#) and the percentage of post "lost" at each step (%Δ) for treatment and control groups.

### 8.4.3 H3: Effect of Post Guidance on contribution quality

Posts from users in the treatment group received more comments (28.6%, $p$=0.004), screen views (26.6%, $p$=0.027), and upvotes (36.1%, $p$=0.004), suggesting they are overall better contributions. In addition, as we previously mentioned, they were reported less often (−9.4%; $p$=0.001), which can also act as a proxy for very low-quality posts. Note that posts created with *vs.* without Post Guidance co-existed in the subreddits during the experiment; these posts "compete" for users' upvotes, screenviews, and comments. Assuming that Post Guidance increases the quality of posts, it could be that engagement levels do not significantly increase as much as we see here once the feature is rolled out for all posts within a subreddit, as twice as many posts would increase in quality and thus relative engagement across all posts might decrease.

### 8.4.4 H4: Effect of Post Guidance on user engagement

Last, we find small and statistically insignificant effects when considering outcomes related to user engagement in their assigned subreddit. This contradicts our hypothesis that Post Guidance would increase user involvement. Similar to in **H2**, we conjecture that two simultaneous effects may be at play here. On the one hand, Post Guidance increases the number of "successful" contributions, which could increase subsequent engagement. However, at the same time, Post Guidance raises the bar for users to participate in the community. To explore this further, we re-ran the analysis, again considering only users who submitted a post in the follow-up period. We find that, indeed, when considering this population, users exposed to Post Guidance experienced significant increases in subsequent engagement (Days contributing: 29.5%; Days voting: 25.9%; Days active 14.0%; $p$<0.001).

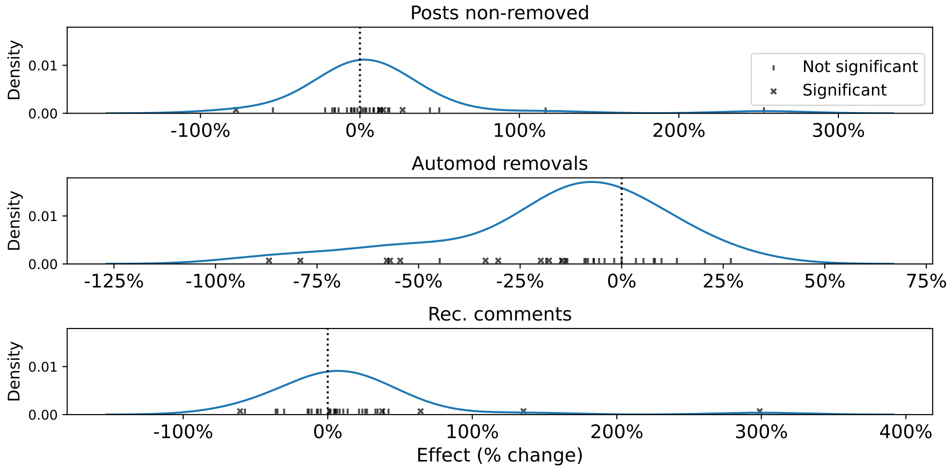


Figure 8.5: Depiction of the heterogeneity of the effect. For three of the outcomes considered (see Table 8.1), we show the distribution of subreddit-level effects with a kernel density estimate (KDE) plot. Close to the $x$-axis, we plot the actual effect sizes observed for each community using different markers for significant (×) and not significant (|) subreddit-level effects. Note that effect sizes vary widely, going from negative to positive across all four outcomes (although, in many cases, the estimated effects are not statistically significant considering only one subreddit).

### 8.4.5 Heterogeneity of the effect

Not all communities are impacted by the Post Guidance in the same way. This is illustrated in Fig. 8.5, where we show that effect sizes vary widely across communities for four out of the thirteen outcomes considered. To further explore when Post Guidance is effective *vs.* when it is not, we decompose the effect size into five components: one "baseline" effect and four "interaction" effects associated with subreddit and user characteristics (see Table 8.1); see Equation 8.2. Recall that we can interpret the coefficients associated with the interaction ($\gamma$) as whether the effect is stronger or weaker when users or subreddits have specific characteristics.

**Newcomers.** We operationalize newcomers as users who, in the 90 days before enrolling in the experiment, did not visit their assigned community (54% of users). Surprisingly, we find that the effect for Post Guidance was not significantly different for newcomers when compared to other users for any of the 13 outcome variables considered (see Fig. 8.6a). This indicates that people who are somewhat familiar with a community benefited equally from the feature compared to those who were not.

**Low activity.** "Newcomer," as defined above, is a community-specific concept. A user might be a newcomer to r/AskReddit and, at the same time, be active in other Reddit communities. We additionally consider users with 'low activity' in the entirety of Reddit, operationalized as users with three or fewer votes on Reddit in the 90 days before enrolling in the experiment (50% of users). Results here are very similar to what we found for newcomers (see Fig. 8.6b),

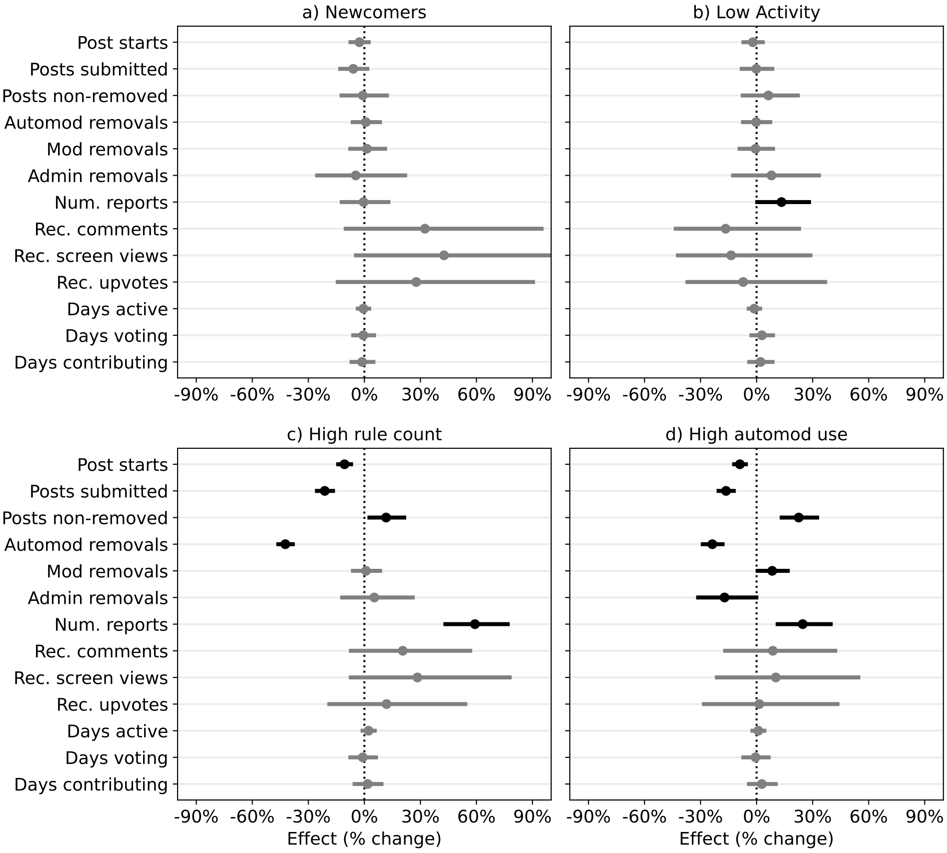


Figure 8.6: Poisson regression results for effect heterogeneity. Recall that estimates should be interpreted as measuring whether the effect is amplified or attenuated given specific conditions. For each binary variable $X$ (one per plot), we show the ratio between the effect between for users where this variable equals one and users where this variable equals zero.

with the sole exception of reports, which increased significantly more for these users relative to the others (relative increase of 12.5%; $p$=0.04). We did not manage to hypothesize a credible reason for this observed effect.

**High rule count and AutoModerator use.** Subreddits were free to choose to which extent they adopted Post Guidance, and the number of rules in the 33 considered subreddits ranged from 2 to 32. To contrast extensive and intensive Post Guidance use, we split our communities into two groups: those with a high rule count (more than seven rules; 46% of users) and those with a low rule count (seven or fewer rules). Further, we note that the functionalities of the Automoderator, another automated content moderation tool at Reddit, overlap with Post Guidance. For example, a community could either configure the AutoModerator to remove posts containing a specific keyword or configure Post Guidance to prevent posts from this keyword from being submitted. In that context, we split out communities into two groups, those with high Post Guidance use (8% of posts or less use Post Guidance; 52% of the communities) and those with low Post Guidance use.

We find that the effect of Post Guidance differs in high rule count and high AutoModerator usage communities. Notably, we found a significant increase in the number of non-removed posts (11.0% for high rule-count, $p$=0.03; 20.3% for high AutoModerator removals, $p$=0.03); a significant decrease in AutoModerator removals (−55.0%, $p$ <0.001; 27.0%, $p$ <0.001); and a significant increase in the number of reported posts (46.5%, $p$ <0.001; 22.0%, $p$ <0.001). Here, we attribute the significant increase in reported posts to the associated increase in non-removed posts. It may be that Post Guidance creates more 'borderline' posts that are not considered rule-breaking by moderators, but that are perceived as so by users of the community. Interestingly, we also find a significant increase in the number of posts removed by moderators in communities with high AutoModerator use (8.0%, $p$=0.03). This may indicate that rules ported in this community from AutoModerator to Post Guidance allowed "bad-faith" rule-breaking users to try to skirt the rules (we discuss this further in Sec. 8.5).

## 8.5 Discussion

Here, we present Post Guidance, a new feature allowing proactive, automatic content moderation. We show with a large-scale experiment that Post Guidance: increased the number of non-removed contributions to communities adopting it **(H1)**; decreased moderator workload **(H2)**; and increased the quality of contributions **(H3)**. Interestingly, we find that Post Guidance did not increase user engagement in the communities adopting it **(H4)**. Our analyses also allow us to hypothesize the mechanisms by which Post Guidance shapes online communities. For example, we find that it *decreases* the number of posts submitted, but those that are submitted are less likely to be removed.

The effectiveness of Post Guidance was similar across more experienced users and newcomers (both to Reddit and to specific communities). Yet, its effect varied across communities; those that saw the largest increases in the number of non-removed posts were the ones that (1) set up many Post Guidance rules before the experiment started; and (2) frequently used the AutoModerator feature before the implementation of Post Guidance. This suggests a possible "dose–response" relationship between the changes in the community and the extent to which they use Post Guidance, as communities that set up more Post Guidance rules are bound to prevent more rule-breaking content, and as communities that heavily used the AutoModerator feature were likely to have rules that were easy to enforce with simple regex expressions.

**Design friction.** While online platforms typically aim to simplify participation, Post Guidance is an example of 'participation friction', adding extra hurdles with the goal of ensuring that more submitted posts are successful. The friction added by Post Guidance may explain our findings regarding **H4**. On the one hand, the feature increases subsequent user engagement for those users who are successful, similar to findings by Srinivasan (2023) [245]. On the other hand, the feature may 'backfire' and discourage users whose posts do not follow the community's rules. Often, when a post or comment is removed *after* submission, users are not even aware [112]; these users may continue to engage as if they had a successful post. By creating friction at the time of submission, Post Guidance may discourage this subsequent engagement. We find evidence supporting this interpretation in Sec. 8.4.4. While the overall result is null, if we limit our analysis to those who ended up submitting their posts, we find a strong positive effect for users exposed to the feature.

**Good *vs.* bad faith rule breaking.** Rule-breaking behavior can be broadly split between “good-faith” rule breaking, when users do not adhere to community norms because they do not know them, and “bad-faith” rule breaking, when users are well aware of the rules but break them regardless. The iteration of Post Guidance studied here is mostly effective against good-faith rule-breaking, as it relies largely on regex patterns that can often be circumvented. In contrast, this current iteration may facilitate bad-faith rule-breaking by helping users skirt the rules (see Sec. 8.4.2). For example, a user may discover via Post Guidance that a specific word is forbidden in a community and then proceed to ‘fuzz’ the word using punctuation to circumvent the rule. While the overall effect of Post Guidance, as studied here, is still positive, this balance can be further shifted in the future through the addition of machine-learning-powered evaluations of content, which go beyond keyword or regex matching and reduce the ability of bad-faith actors to circumvent community rules.

**Proactive *vs.* reactive moderation.** The introduction of Post Guidance enables moderators to moderate content both proactively (Post Guidance) and reactively (using AutoModerator). Interestingly, in Sec. 8.4.5, we find that subreddits that relied heavily on automated *reactive* moderation (high AutoModerator use) had significantly more non-removed posts and more manually removed posts than those that did not. These differences could be explained by the overlapping functionalities between AutoModerator and Post Guidance: if subreddits already had meaningful AutoModerator rules (that triggered in many posts), it may have been particularly easy to ‘port’ these meaningful rules to the new Post Guidance feature. However, it may also be that these subreddits experienced more dramatic changes because they were the communities where a substantial chunk of automated *reactive* moderation turned *proactive*. Last, it is important to stress that proactive and reactive moderation are complementary paradigms. On the one hand, AutoModerator is better suited to implement regex rules to “bad-faith” rule-breaking (*e.g.*, spam), as adversarial agents will likely exploit the Post Guidance interface to create rule-breaking posts. On the other hand, Post Guidance is better suited to educate users who want to conform to community guidelines.

**User-centricity.** We argue that an interesting way to examine content moderation features is to ask: who is burdened by it? In the case of a simple ‘delete button’ that reactively removes rule-breaking content, the answer is simple: it burdens moderators who will use it. But even proactive content moderation features may burden content moderators; for instance, Horta Ribeiro et al. [96] studied “Post Approvals,” a feature used in Facebook groups where every post has to be manually approved by moderators before landing in the groups’ feed. Post Approvals effectively reduced low-quality posts but led moderators to create around-the-clock shifts to handle the demand of highly active communities [1]. In contrast to these moderation practices that are “moderator-centric,” in the sense that they add work to moderators, Post

Guidance is “user-centric,” *i.e.*, it burdens users. The work at hand, as well as previous work using nudges, suggest that user-centric approaches are effective.

### 8.5.1 Limitations

We conduct a large-scale randomized experiment "in the wild" on Reddit, with a diverse set of communities, meaning that our key findings (*i.e.*, answers to **H1**–**H4**) have high internal and external validity. However, our secondary analyses have limitations worth discussing. First, in Sec. 8.4.2 and Sec. 8.4.4, we conduct an analysis considering only users who submitted posts. This can yield biased results because there may be confounders that cause both users to submit posts and the other outcomes. Nonetheless, we argue that this analysis, albeit imperfect, provides us with further insight into the tradeoffs involved with Post Guidance, and we were not able to think of any particular confounder that would hinder the conclusions drawn in Sec. 8.4.2 and Sec. 8.4.4. Second, when we conduct the analyses on the heterogeneity of the effect (Sec. 8.4.5), we stress that we cannot attribute “cause-and-effect” interpretations to how the considered variables modify the effect. For example, there could be other feature that causes both effect modification and lead subreddits to have a high rule count. Therefore, results on the heterogeneity of the effect should be interpreted as descriptive, *i.e.*, how the effect differs for users with different characteristics.

### 8.5.2 Implications and Future Work

Post Guidance is a promising content moderation strategy that can improve online communities while reducing the moderation workload. More broadly, we believe that research evaluating the effect and nuances around content moderation (like this paper) can help improve the public debate around online platforms, making the available toolkit of interventions more transparent to stakeholders in academia, industry, and government.

**Engaging users.** The *user-centric* aspect of Post Guidance refers to the ways in which the overall approach requires users to actively engage with the rules of the community in order to contribute. While this study focused on the immediate effects, future work could explore how engaging users in this way could shape their relationships with their community. Prior work has shown how resolving ambiguity around specific applications of community rules can help moderators to iteratively build an understanding of a community’s goals [51]; shifting this effort towards users could confer similar benefits. Future work might also explore whether stronger engagement with the rules and goals of a community leads to the formation of stronger feelings of attachment to the community.

**Adapting to Post Guidance.** While we evaluate one specific implementation of Post Guidance in this work, the paradigm itself represents a tool that can flexibly be used by moderators alongside other tools. In our exploration of the heterogeneity of the effect, we find that Post Guidance was particularly efficient in increasing the number of non-removed posts in certain communities. Additionally, we find that the effectiveness of Post Guidance varied along with the use of existing moderation tools, such as Automoderator. Future work could explore how Post Guidance is, and could most effectively be, integrated by community moderators into a broader set of strategies and tools over time and adapted as the goals and needs of a community evolve.

**Going beyond Reddit.** Post Guidance could be used on other platforms hosting online communities, such as Discord and Facebook Groups. Adapting the Post Guidance approach to the particularities of each platform could be an interesting venue for future work and would help improve online communities across the Web. We highlight that a proactive moderation feature like Post Guidance could be particularly interesting for Wikipedia, a large online community centered around building an encyclopedia. Wikipedia is one of the internet's greatest "public goods," and struggles with retaining newcomers, especially since edits must follow a strict set of rules, and potential Wikipedians often give up after having their edits reverted [170, 90]. We conjecture that an intervention like Post Guidance could diminish these negative experiences and increase contribution to the world's largest encyclopedia.

**Going beyond regex.** Post Guidance uses only simple rules programmable with regex (besides other simple features, like account age). But with more complicated models, *e.g.*, Large Language (multimodal) Models, one could imagine creating a more flexible version of Post Guidance. Some rules can be very easily enforced using regex (*e.g.*, all posts must end with a question mark), whereas other can't (*e.g.*, "be kind," "only post pictures of cats"). More complex models could, therefore, increase the range of rules enforced by the community, allowing subreddits to deliver personalized prompts to a larger variety of scenarios. That being said, a big challenge in that direction is the loss of transparency — although limited, a great virtue of the pattern-matching approach is that it gives community leaders very fine-grained control over automated moderation, which would perhaps be lost with more complex models.

### 8.5.3 Ethical Considerations

All communities involved in this experiment sought out and consented to participate. All data used in the final analyses for this paper was de-identified and analyzed in aggregate, with care not to single out any individual nor violate users' privacy. Consent was received from the moderators of any community referenced by name. We argue that the benefits of this study greatly outweigh its potential harms, given that designing tools to understand online communities better can greatly help improve the Web as a whole.

Table 8.2: Poisson regression results. Standard errors were calculated using the Huber robust estimator. We report effect sizes and confidence intervals as relative changes, e.g., a $-5.7\%$ effect means that treated units experienced a relative decrease of 5.7% relative to control units. Note that, to obtain effects as relative change, we simply transform the estimated effects using the formula $(e^{\beta}-1)$. Given that $\beta = \log \frac{\mathrm{E}(Y|Z=1)}{\mathrm{E}(Y|Z=0)}$, we have $(e^{\beta}-1) = \frac{\mathrm{E}(Y|Z=1)-\mathrm{E}(Y|Z=0)}{\mathrm{E}(Y|Z=0)}$.

| # | Outcome | Effect | *p*-value | Hyp. |
|---|---|---|---|---|
| 1 | Post starts | -5.7%; 95% CI [-7.8%, -3.5%] | <0.001 | H1 |
| 2 | Posts submitted | -13.0%; 95% CI [-15.8%, -10.2%] | <0.001 | H1 |
| 3 | Posts non-removed | 5.8%; 95% CI [0.6%, 11.2%] | 0.03 | H1 |
| 4 | Automod removals | -34.9%; 95% CI [-37.0%, -32.8%] | <0.001 | H2 |
| 5 | Mod removals | 2.7%; 95% CI [-1.7%, 7.2%] | 0.236 | H2 |
| 6 | Admin removals | -9.2%; 95% CI [-17.3%, -0.4%] | 0.042 | H2 |
| 7 | Num. reports | -9.4%; 95% CI [-14.4%, -4.1%] | 0.001 | H2, H3 |
| 8 | Rec. comments | 28.6%; 95% CI [8.2%, 52.9%] | 0.004 | H3 |
| 9 | Rec. screen views | 26.6%; 95% CI [2.8%, 56.0%] | 0.027 | H3 |
| 10 | Rec. upvotes | 36.1%; 95% CI [10.1%, 68.1%] | 0.004 | H3 |
| 11 | Days contributing | -2.0%; 95% CI [-5.2%, 1.3%] | 0.233 | H4 |
| 12 | Days voting | -1.9%; 95% CI [-5.0%, 1.2%] | 0.229 | H4 |
| 13 | Days active | -1.4%; 95% CI [-2.9%, 0.1%] | 0.059 | H4 |

# Discussion and Conclusion Part IV

# 9 Discussion and Conclusion

This thesis aims to advance the understanding of content curation practices by measuring the causal effect of content moderation interventions on user behavior in a series of studies. In the first part of the thesis, we motivated this enterprise by highlighting (1) the critical role that online platforms play in our lives and in our society; (2) that content curation is central to online platforms; and (3) that content curation practices are disproportionally guided by anecdotal or observational evidence. We also discussed work and background information on topics relevant to content moderation on online platforms: antisocial behavior online, content moderation practices and paradigms, and previous work studying the effect of content moderation interventions.

We conducted six studies roughly divided into two themes: moderating people and moderating content. In the second part of this thesis, we studied what happens when platforms ban individuals or collectives due to rule-breaking behavior. In the third part of this thesis, we studied different ways platforms can moderate rule-breaking content. Individually, these studies contribute to our understanding of different content moderation practices, but collectively, they form a proposal for a new way of doing content moderation—anchored in evidence rather than common sense. They advocate for a more empirical approach to content moderation, one that leans on data-driven insights to create, refine, and tailor interventions.

A central theme across the studies presented is the existence of trade-offs. Content moderation practices are no silver bullets. Instead, they embody a complex interplay between protecting free speech and safeguarding community standards; improving user behavior in a specific online platform and creating isolated but toxic communities elsewhere; decreasing rule-breaking behavior and reducing participation in online communities. These trade-offs underscore the importance of a tailored approach to content moderation that adapts to the evolving dynamics of online platforms and their user bases. Studies presented here do not estimate causal effects that generalize across time and platforms—and thus, to guide content moderation practices, we must continuously evaluate interventions and content moderation practices as our information ecosystem evolves.

While this thesis was written, for example, large language models (LLMs) grew tremendously in popularity and capacity, capturing the collective imagination. In recent work unrelated to this thesis, we showed that LLMs are widely used on crowd work platforms and that targeted mitigation strategies can reduce, but not eliminate, their usage [272]. This is concerning as LLM use can threaten the validity of research conducted on crowd work platforms like Prolific and Amazon Mechanical Turk, as researchers usually care about human (rather than model) behavior or preferences. It is even more urgent to understand how large language models will impact other online platforms, such as social media. These models can browse the Web, engage in conversation in a human-like fashion, and interact with multimodal content; thus, they are bound to create new challenges. Small radical communities already capable of shaping media narratives and harassing users will use these models to astroturf and artificially generate content capable of misleading and harming others. Whether this thesis' findings hold in a post-LLM internet begs the question and highlights the need to evaluate content moderation practices continuously.

The impact of the work presented here was partly due to industry collaborations—the entire third part of this thesis was done in partnership with companies. Chapters 6 and 7 were done in collaboration with researchers from Meta, and Chapter 8 with researchers and engineers from Reddit. One of the advantages of collaborating with industry was that results obtained directly impacted products used by millions of users. Our work showcasing the effectiveness of post approvals [96], a feature where moderators require posts in Facebook groups to be pre-approved, led to expanding the feature to comments and prioritizing additional features that decreased the effort to approve posts. Our research on content removal [95] has broadened the success metrics associated with content moderation within Facebook—leading them to consider second-order effects like subsequent rule-breaking behavior when moderating content. Last, Post Guidance was developed alongside Reddit, and our research there informed the creation and deployment of a new content moderation intervention from day zero.

At the same time, collaborating with industry is not sufficient to advance research on content moderation. For example, the studies presented in the second part of this thesis concern banning controversial influencers, communities, and platforms, which likely would not be approved by companies' internal research pipelines, as these controversial topics may bring political and regulatory trouble to online platforms. Nonetheless, these studies were conducted in broad collaborations with academics worldwide—which was often vital to obtaining the data needed, *e.g.*, access to data from The Nielsen Company (used in Chapter 4) was granted through a collaboration with researchers from The University of Pennsylvania. Rather than directly impacting user-facing products, these papers impacted public discourse around content moderation and deplatforming, appearing on outlets like NBC News,[I] WIRED,[II] and The Conversation.[III]

[I] https://go.epfl.ch/deplatforming-nbcn

[II] https://go.epfl.ch/deplatforming-wired

[III] https://go.epfl.ch/deplatforming-conv

Three ingredients—causal science, digital traces, and domain expertise in the technical infrastructure of online platforms—are critical to the evidence-informed content moderation paradigm championed in the thesis. The research done here would not be possible without quasi-experimental designs like difference-in-differences; Or without digital traces from Reddit, Wikipedia, and Facebook; Or without understanding how comments are removed from Facebook or how communities are banned from Reddit. Without digital traces, studies of online behavior lack ecological validity; Without causality, results cannot meaningfully inform the questions that matter (*e.g.*, *does deplatforming reduce online attention towards fringe social media?*) as these are "what if" questions [188]. Without understanding the technical infrastructure of online platforms, it is hard to find ways to identify causal effects and ensure the validity of our measurements.

## Future directions

The results presented in this thesis point to several future directions, some of which we sketch here, concluding the thesis.

### New ways to obtain data

Obtaining reliable and extensive digital traces has become increasingly complex due to technological, legal, and ethical challenges. Application Programming Interfaces (APIs) have traditionally been a primary channel for researchers to access data from various online platforms. However, in recent years, restrictions on data access imposed by these platforms limit the scope of research [69]. At the same time, web crawling has become increasingly difficult due to the sophisticated measures websites employ to detect and block automated access.

The lack of data may severely harm the quality and throughput of research informing content curation practices. Therefore, future work must find ways around these challenges. One promising avenue is facilitating and promoting data sharing among individuals, academic institutions, research organizations, and private entities. Initiatives to facilitate individuals donating their data already exist,[IV] and often piggyback on infrastructure that platforms provide users to extract their data (*e.g.*, Google Takeout). Another related direction is the establishment of online panels (like the one we used to study the deplatforming of Parler in Chapter 4), where consenting participants share their digital engagement data. Importantly, these panels can be tweaked and weighted so that they are representative of the general population.

[IV] E.g., see https://datadonation.uzh.ch/ddm/login/.

### Large language models and online platforms

The advent of large language models (LLMs) has substantial implications for online platforms. These widely accessible[V] models can generate human-sounding text and realistic images and, therefore, may redefine the landscape of digital content creation and consumption. In the pre-LLM era, concerns around mis/disinformation and influence operations on online platforms were already widespread [86]. However, LLMs may substantially improve existing practices, allowing arguments and persuasive strategies to be tailored to individuals and happen in conversations. In that context, future work may help assess the threats associated with LLM use on online platforms and explore how content moderation practices may prevent these issues. Taking persuasion as an example, future work may benchmark if the new conversational paradigm used by LLMs is worth worrying about and develop mitigation strategies, *e.g.*, flagging content online that is likely generated by LLMs.

Besides bringing forth new challenges, LLMs can also help improve content curation practices. Throughout this thesis, we stressed the importance of tailoring interventions, *e.g.*, in Chapter 8, we introduced an intervention where community leaders created trigger conditions that could prevent posting. LLMs could enable personalized moderation approaches like Post Guidance. For example, they may be used to (1) to help community leaders configure systems that require technical expertise (like Reddit's Automod [116]); (2) detect content that breaks the rules from specific communities (given their remarkable few-show capabilities). (3) recognize signs of escalating conflict or emotional distress in online interactions and intervene with programmed responses designed to calm tensions or direct users to supportive resources [36].

### Putting it all together

Observational and experimental studies like the ones presented in this thesis typically obtain point estimates in a specific platform at a particular moment. What was the effect of deleting comments on Facebook on subsequent rule-breaking behavior over 28 days during the summer of 2022? To what extent did deplatforming Parler in January 2023 decrease Parler users' consumption of Fringe social media? Future work could go beyond these point estimates and draw guidelines and best practices for content moderation, making research more broadly actionable.

Duncan Watts once said *"For 20 years, I thought my job was as a basic scientist. Publish papers and throw them over the wall for someone else to apply. I now realize that there's no one on the other side of the wall. Just a huge pile of papers that we've all thrown over."*[VI] In that context, work interpolating the point estimates obtained through rigorous, empirical studies like the ones presented in this thesis is key to preventing research on content curation practices from "piling up" on the other side of the wall.

[V] LinkedIn, for example, already allows users to "write with AI" as a platform feature, https://www.linkedin.com/help/linkedin/answer/a1517763

[VI] See https://www.nature.com/articles/d41586-024-00479-w

# A Appendix

## A.1 Removing content

**Visual analysis.** As a first sanity check, we visually inspect discontinuities in the outcome variables around the thresholds (*e.g.*, as shown in Fig. 6.5). We find that we can visualize the discontinuities right around the threshold, as expected in a fuzzy regression discontinuity design. In Fig. A.1, we further show the discontinuities for the thread-level scenario (setup: **≤20/All**).

**Manipulation at the cutpoint.** One well-established threat to the validity of RD designs is that individuals may have knowledge about the cutpoint and adjust the running variable (for us, the score *S*) to fall right before or right above it. For instance, in our scenario, if users knew exactly what score their comment would receive before posting it, they might re-word until they find it below the threshold. While in our case, we do not consider this threat to be credible, we entertain the hypothesis and conduct the standard robustness checks, inspecting the density of scores around the threshold and conducting the McCrary test [157] to assess whether the discontinuity in the density of the rating variable at the cutpoint equals zero. We plot the density in Fig. A.2, which shows no indication of manipulation around the threshold (as does the McCrary test, where $p > 0.05$).

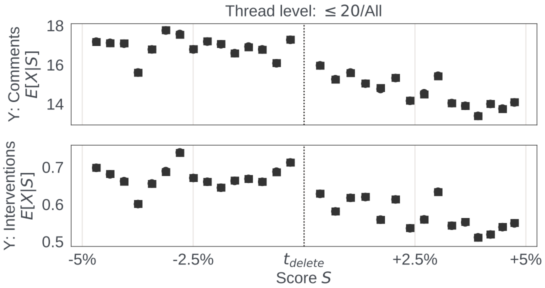


Figure A.1: Example of the discontinuities in the outcome (top: comments; bottom: interventions) we visually inspected to ensure the validity of our approach.

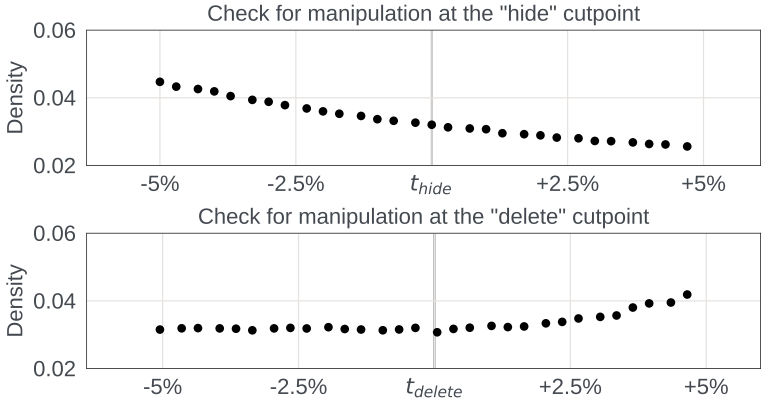


Figure A.2: Density of the running variable (*i.e.*, the score $S$) around the thresholds where content gets deleted ($t_{\text{delete}}$) and hidden ($t_{\text{hide}}$).

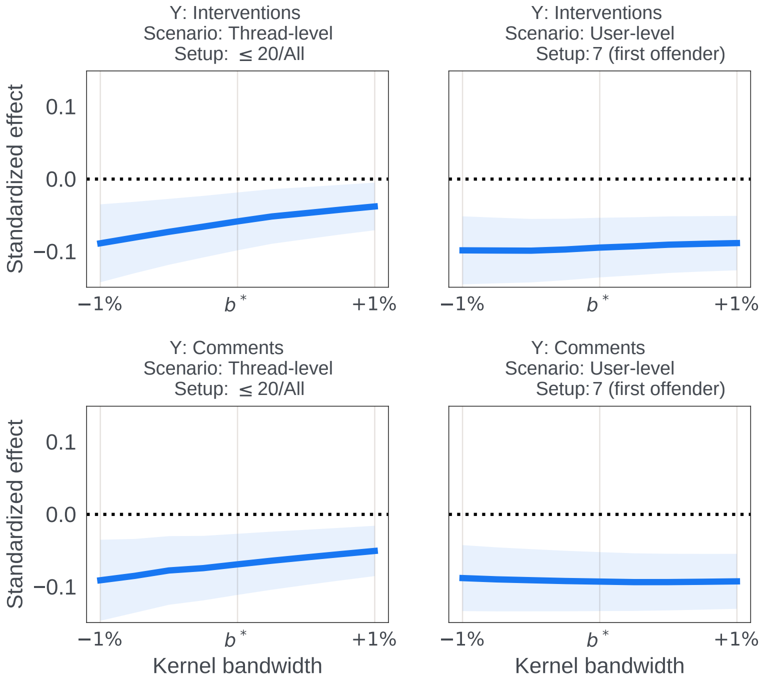


Figure A.3: We show how the standardized effect ($y$-axis) of four of the regression discontinuity designs vary with slight changes to the kernel bandwidth. $b^*$ corresponds to the MSE-optimal bandwidth for each FRD (as described in [106]).

**Placebo FRDs.** A key assumption of FRD is that around the threshold, units are exactly the same except for the fact that those above the threshold have an increased chance of receiving the treatment. As such, it is commonplace (*e.g.*, see [110]) to repeat the entire FRD analysis considering a variable that the treatment should not impact. If comments below and above the threshold are comparable, we should not see significant differences for these placebo FRDs. In our case, we run placebo FRDs considering the same outcome variables calculated in the pre-assignment period (see Fig. 6.3), where no intervention occurred. Thus, we should expect no discontinuity in the outcomes. Results are reported in Table A.1 and Table A.1 . We find that effects in the placebo FRDs are small and not statistically significant, *i.e.*, $p > 0.05$, suggesting that, indeed, comments below and above the threshold are comparable.

**Varying the bandwidth.** Finally, to ensure our findings were robust to slight changes in the kernel bandwidth, we repeated the FRDs with varying bandwidth sizes. In Fig. A.3 we show the changes in the standardized effect for four of the FRDs carried (at both the user level and the thread level) when varying the kernel bandwidth around the MSE-optimal bandwidth $b^*$. Overall, we find that our results are robust to slight changes in the kernel bandwidth.

Table A.1: Results for placebo FRDs. This table is exactly the same as Table 6.2 except that outcomes are calculated in the pre-assignment period and thus should not differ between treatment and control groups (*i.e.*, those whose comment fell above or below the intervention thresholds).

| **Delete**: Thread-level | | | | |
|---|---|---|---|---|
| Outcome | Setup | Effect | Effect (Standardized) | n |
| Comments | ≤20 / All | -0.125 (-0.319, 0.068) | -0.024 (-0.062, 0.013) | 190885 |
| | ≤20 / Other | -0.153 (-0.336, 0.031) | -0.030 (-0.065, 0.006) | 200655 |
| | >20 / All | -11.87 (-41.36, 17.62) | -0.022 (-0.076, 0.033) | 49645 |
| | >20 / Other | -20.34 (-49.02, 8.33) | -0.046 (-0.111, 0.019) | 52241 |
| Interventions | ≤20 / All | 0.003 (-0.011, 0.016) | 0.006 (-0.022, 0.034) | 190885 |
| | ≤20 / Other | 0.001 (-0.013, 0.015) | 0.002 (-0.028, 0.032) | 200655 |
| | >20 / All | 0.038 (-0.429, 0.505) | 0.005 (-0.061, 0.072) | 49645 |
| | >20 / Other | -0.005 (-0.450, 0.441) | -0.001 (-0.067, 0.065) | 52241 |
| **Delete**: User-level (first offender) | | | | |
| Comments | 7 | -0.186 (-1.77, 1.39) | -0.004 (-0.038, 0.030) | 162149 |
| | 14 | -0.545 (-3.38, 2.29) | -0.007 (-0.045, 0.031) | 112793 |
| | 21 | 0.569 (-4.11, 5.25) | 0.006 (-0.042, 0.054) | 84592 |
| | 28 | 3.68 (-2.35, 9.71) | 0.029 (-0.019, 0.078) | 60175 |
| Interventions | 7 | 0.013 (-0.008, 0.035) | 0.019 (-0.012, 0.050) | 162149 |
| | 14 | 0.001 (-0.029, 0.031) | 0.001 (-0.036, 0.039) | 112793 |
| | 21 | -0.012 (-0.052, 0.029) | -0.013 (-0.057, 0.031) | 84592 |
| | 28 | -0.008 (-0.065, 0.049) | -0.008 (-0.063, 0.048) | 60175 |
| **Delete**: User-level (repeat offender) | | | | |
| Comments | 7 | -0.369 (-5.73, 4.99) | -0.007 (-0.108, 0.094) | 29825 |
| | 14 | -0.906 (-7.57, 5.76) | -0.010 (-0.081, 0.061) | 26596 |
| | 21 | 1.94 (-12.45, 16.33) | 0.015 (-0.095, 0.124) | 21693 |
| | 28 | 2.01 (-15.87, 19.90) | 0.012 (-0.097, 0.121) | 18468 |
| Interventions | 7 | 0.048 (-0.022, 0.118) | 0.057 (-0.026, 0.139) | 29825 |
| | 14 | 0.036 (-0.069, 0.141) | 0.031 (-0.061, 0.124) | 26596 |
| | 21 | 0.086 (-0.047, 0.219) | 0.062 (-0.034, 0.159) | 21693 |
| | 28 | 0.021 (-0.128, 0.170) | 0.013 (-0.081, 0.108) | 18468 |

Table A.2: Results for placebo FRDs. This table is exactly the same as Table 6.3 except that outcomes are calculated in the pre-assignment period and thus should not differ between treatment and control groups (*i.e.*, those whose comment fell above or below the intervention thresholds).

| **Hide**: Thread-level | | | | |
|---|---|---|---|---|
| Outcome | Setup | Effect | Effect (Standardized) | n |
| Comments | ≤20 / All | -0.054 (-0.150, 0.042) | -0.010 (-0.029, 0.008) | 868632 |
| | ≤20 / Other | -0.039 (-0.140, 0.062) | -0.007 (-0.027, 0.012) | 907871 |
| | >20 / All | -18.16 (-40.09, 3.77) | -0.018 (-0.041, 0.004) | 300716 |
| | >20 / Other | -17.14 (-38.37, 4.08) | -0.017 (-0.037, 0.004) | 314287 |
| Interventions | ≤20 / All | 0.001 (-0.005, 0.006) | 0.002 (-0.014, 0.019) | 868632 |
| | ≤20 / Other | 0.001 (-0.005, 0.006) | 0.002 (-0.014, 0.018) | 907871 |
| | >20 / All | -0.021 (-0.146, 0.104) | -0.005 (-0.033, 0.023) | 300716 |
| | >20 / Other | -0.063 (-0.163, 0.038) | -0.014 (-0.036, 0.008) | 314287 |
| **Hide**: User-level (first offender) | | | | |
| Comments | 7 | -0.501 (-1.28, 0.279) | -0.009 (-0.023, 0.005) | 723278 |
| | 14 | -0.972 (-2.37, 0.422) | -0.011 (-0.027, 0.005) | 542780 |
| | 21 | -0.877 (-3.07, 1.32) | -0.008 (-0.026, 0.011) | 422996 |
| | 28 | 1.30 (-2.53, 5.14) | 0.008 (-0.016, 0.033) | 291712 |
| Interventions | 7 | 0.003 (-0.008, 0.015) | 0.006 (-0.014, 0.025) | 723278 |
| | 14 | -0.009 (-0.025, 0.008) | -0.013 (-0.037, 0.011) | 542780 |
| | 21 | -0.011 (-0.033, 0.010) | -0.014 (-0.042, 0.013) | 422996 |
| | 28 | -0.003 (-0.024, 0.018) | -0.003 (-0.026, 0.020) | 291712 |
| **Hide**: User-level (repeat offender) | | | | |
| Comments | 7 | 0.231 (-2.11, 2.58) | 0.004 (-0.033, 0.040) | 126587 |
| | 14 | 0.766 (-3.10, 4.63) | 0.007 (-0.028, 0.042) | 122545 |
| | 21 | 0.214 (-5.26, 5.69) | 0.001 (-0.034, 0.037) | 103467 |
| | 28 | 3.83 (-3.82, 11.49) | 0.020 (-0.020, 0.060) | 82067 |
| Interventions | 7 | 0.003 (-0.021, 0.026) | 0.004 (-0.028, 0.036) | 126587 |
| | 14 | 0.002 (-0.035, 0.038) | 0.002 (-0.035, 0.038) | 122545 |
| | 21 | 0.029 (-0.038, 0.096) | 0.024 (-0.032, 0.079) | 103467 |
| | 28 | 0.009 (-0.047, 0.065) | 0.007 (-0.034, 0.047) | 82067 |

## A.2 Banning influencers

Table A.3: Final patterns used to extract deplatforming events.

| **Patterns** |
|---|
| <ent> × on <plat> |
| <plat> employee deactivated <ent> |
| <plat> employee ‘deactivated‘ <ent> |
| <plat> employee deactivates <ent> |
| <plat> bans <ent> |
| <plat> suspends <ent> |
| <plat> suspended <ent> |
| <plat> banned <ent> |
| <plat> removes <ent> |
| <plat> ban <ent> |
| <plat> banning <ent> |
| <ent> locked out of <plat> |
| <ent> blocked on <plat> |
| <plat> deleted <ent> |
| <plat> takes down <ent> |
| <plat> blocks <ent> |

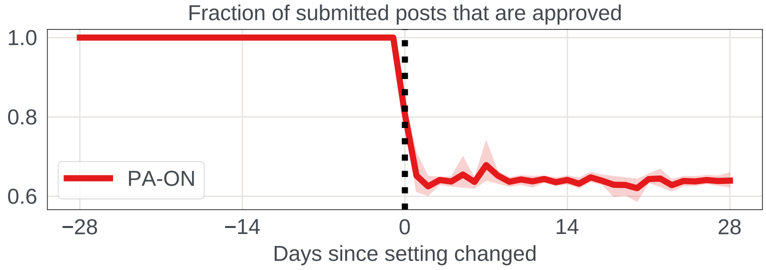


Figure A.4: For PA-ON groups, we show the fraction of posts submitted that were approved.

## A.3 Pre-approving content

### A.3.1 Propensity Score Matching

Throughout the paper, we performed one-to-one propensity score matching (PSM) of PA-ON and PA-OFF communities on the group-level variables. We considered all time-varying variables in Table 7.1 (under the headers "activity-related" and "moderation-related") as well as all time-invariant variables (under the header "group characteristics"), with the exception of *Group categories*. We did not match on the latter, but found good covariate balance nonetheless, *cf.* Appendix A.3.2. Matching was done using nearest-neighbor matching without replacement, as implemented in Ho et al. (2011) [93]. Propensity scores were obtained using a Gradient Boosting Classifier, as implemented in Pedregosa et al. (2011) [189]. For both matching procedures and for all continuous variables, we obtained absolute standardized mean differences (SMDs) smaller than the commonly-used 0.1 threshold that indicates imbalance [11].

**What leads to the adoption of post approvals?** For the analyses done in Sec. 7.4.1, we performed PSM using an absolute caliper of 0.5, discarding 125 PA-ON groups (1.4%) for which we were not able to find good matches. For time-varying variables (under the headers "activity-related" and "moderation-related" in Table 7.1), we considered three days of interest (days −22, −25, and −28). In other words, each covariate-day pair corresponds to a distinct feature used by the classifier to obtain the propensity scores (*e.g.* posts on day −22, posts on day −25, and posts on day −28). Day 0 was when the intervention took place for PA-ON group (*i.e.*, when they enabled the post approvals setting), and it was chosen at random for PA-OFF groups. Covariate balance for this matching is shown in Fig. A.7.

**How do post approvals shape online communities?** For the analyses done in Sec. 7.4.2 and Sec. 7.4.3, we performed PSM with a caliper of 0.0075, discarding 1426 PA-ON groups (16%) for which we were not able to find good matches. For time-varying variables (under the headers "activity-related" and "moderation-related" in Table 7.1), we considered five days of interest (days −28, −21, −14, −7 and −1). Additionally, we considered the number of posts, comments, reactions, deleted posts, reported posts and members removed on the day of the intervention (day 0) measured up to the hour when the intervention (or pseudo-intervention) was introduced. These were all variables that we were able to measure hourly. Matching on the day of the intervention was done on an hourly basis as PA-ON communities have periods of activity with and without the post approval setting. Covariate balance for this matching is shown in Fig. A.8.

**Additional observations.** We clarify a couple of decisions regarding the propensity score matching procedure:

- As mentioned in Sec. 7.3, to obtain candidate PA-OFF groups for matching, we used a random sample of 50% of all communities that had over 128 members and at least one post and one comment over any 7-day period. This reduces the matching space, as we could have used 100% of all communities. Yet, empirically, using more than 50% of the sample harmed the propensity matching score matching capacity to balance the variables of interest as the class imbalance was too extreme. Even with a 50% sample, before matching, the number of PA-OFF groups ($n$ = 224,635) outnumbered PA-ON groups ($n$ = 8,767) by approximately 25 to 1.
- Differences between the PSM done for Sec. 7.4.1 *vs.* for Sec. 7.4.2 and Sec. 7.4.3 are as follows:
  - Different calipers were used, as achieving covariate balance was harder for the matching in Sec. 7.4.2/7.4.3.
  - Different dates were considered for time-varying variables (−22, −25, −28 *vs.* −1, −7, −14, −21, −28).
  - The matching used in Sections 7.4.2 and 7.4.3 additionally included variables from the day when post approvals was turned on, measured up to the very hour when the setting was changed (*cf.* Fig. A.8b)

### A.3.2 Group Topics

To ensure that PA-ON and PA-OFF groups were topically comparable after propensity score matching, we used Empath [63]'s lexical categories. Lexicons have been used in causal inference by Koustuv et al. (2019) [223] and by Sridhar and Getoor (2019)[244]. We translated group titles and descriptions into English and measured the occurrence of words matching each of the 194 default Empath categories (e.g., *work*, *celebration*, *writing*, *etc.*). Even without explicitly matching groups by Empath category word frequency, propensity score matching yielded good covariate balance – the standardized mean difference for all 194 categories was below 0.1,[I] suggesting the two sets of groups had similar titles and descriptions. Fig. A.7c and Fig. A.8d show the covariate balance of the top 20 most common Empath categories before and after PSM.

[I] Except for the categories *communication* and *internet* in the PSM done for Sec. 7.4.2, where they had SMDs equals to 0.103/0.107.

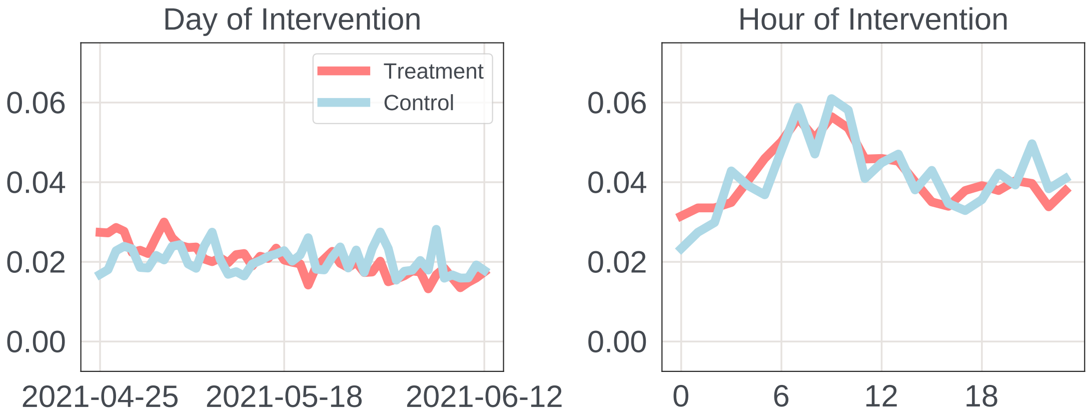


Figure A.5: We show the day (left) and hour (right) of the intervention (for PA-ON groups) and pseudo-intervention (for PA-OFF groups) after the matching done in Sections 7.4.2 and 7.4.3. Note that the hours are shown in UTC+0.

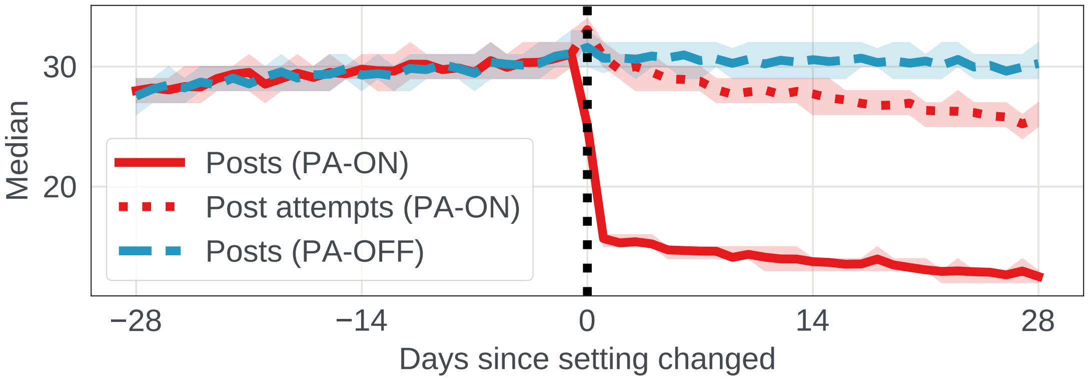


Figure A.6: This figure repeats the analysis done in Fig. 7.4 using the median instead of the winsorized mean.

### A.3.3 Additional Plots

We provide a couple of additional plots as sanity checks:

- Fig. A.5 shows the distribution of hour and day of the intervention for the matching done for Sections 7.4.2 and 7.4.3.
- Fig. A.6 reproduces Fig. 7.4 using the median instead of the (winsorized) mean.
- Complementing Fig. 7.4, Fig. A.4 shows the fraction of posts submitted that were approved.

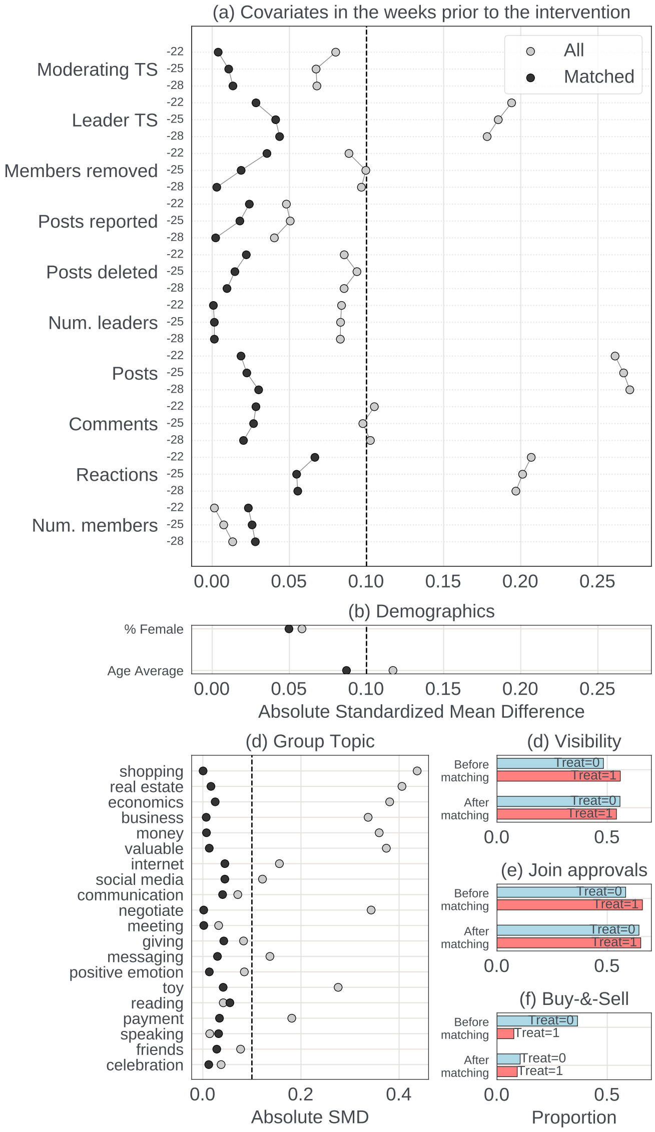


Figure A.7: Covariate balance for matching done in Sec. 7.4.1. *(a)* Absolute standardized mean difference (SMD) for all time-varying variables considered in days −28, −25 and −22. *(b)* SMD for demographic-related variables. *(c)* SMD for the top 20 most popular group categories (from Empath). *(d-f)* covariate balance pre- and post-matching for the three binary variables considered in the matching.

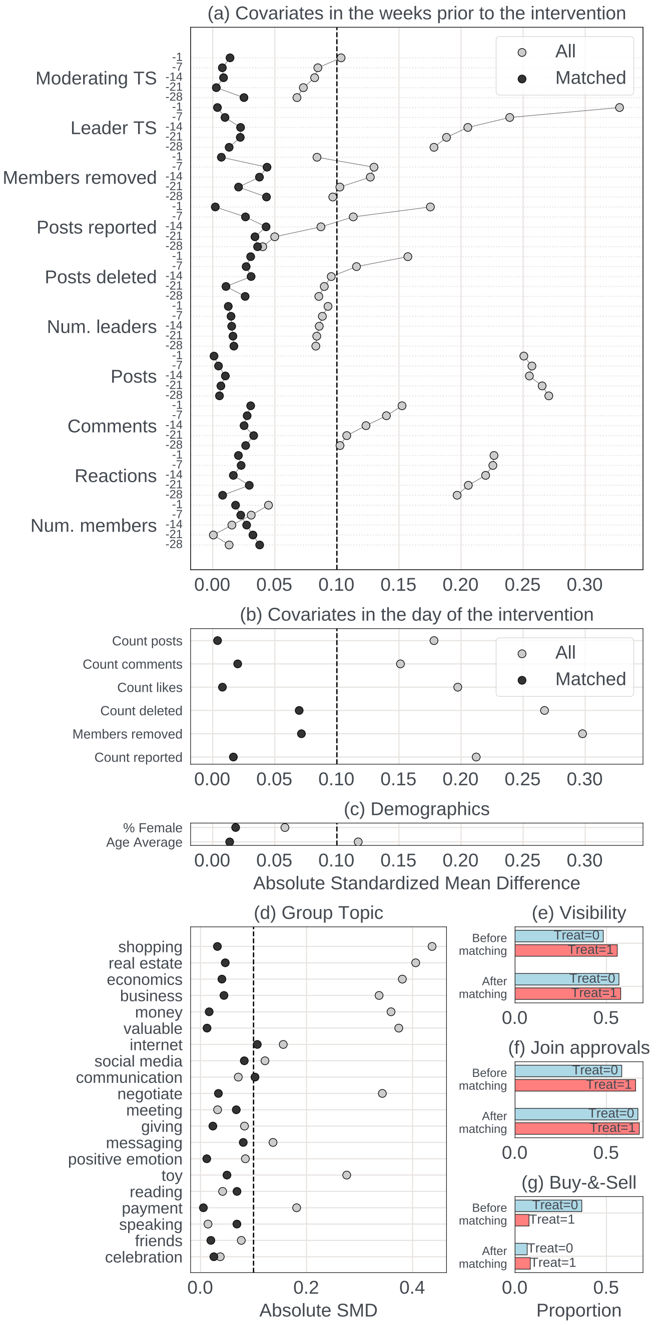


Figure A.8: Covariate balance for matching done in Sec. 7.4.2. *(a)* Absolute standardized mean difference (SMD) for all time-varying variables considered in days −1, −7, −14, −21, and −28. *(b)* SMD for variables measured in the day of the intervention. *(c)* SMD for demographic-related variables. *(d)* SMD for the top 20 most popular group categories (from Empath). *(e-g)* covariate balance pre- and post-matching for the three binary variables considered in the matching.

## A.4 Guiding content creation

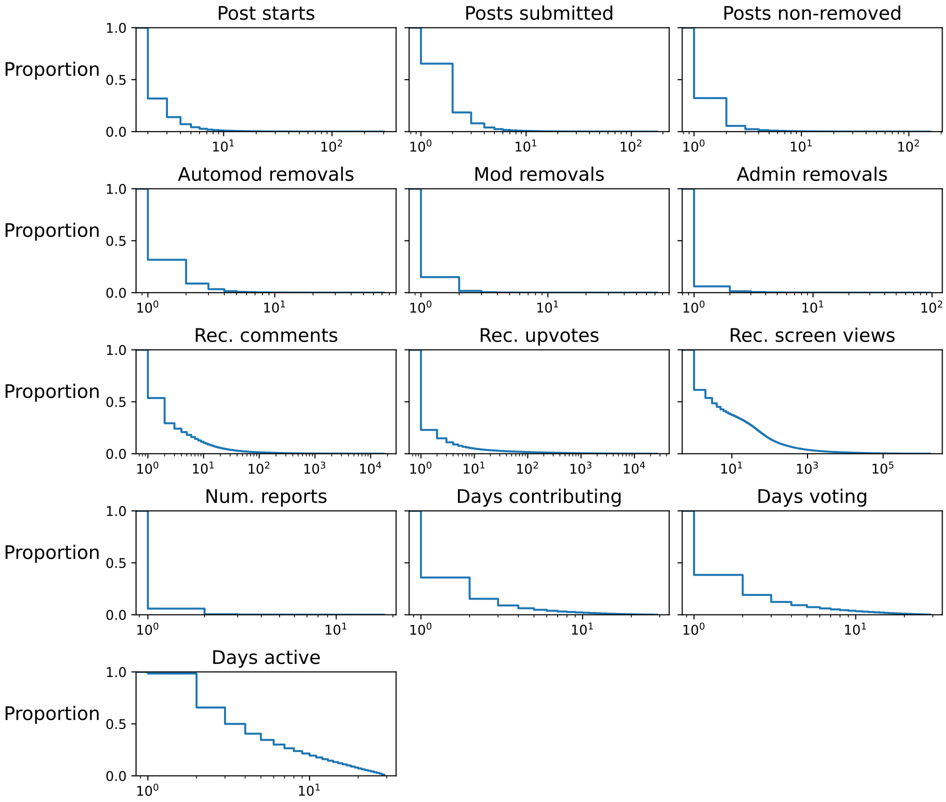


Figure A.9: Distribution of the outcomes considered in the experiment. Note that distributions are heavy-tailed, which motivated our choice to model the effect using a Poisson regression.

## A.5 Banning platforms

### A.5.1 Extended methods: Data

**Desktop panel description**

In the desktop panel, users are asked to install tracking software on their web browser(s). When the users open their browser and access a website, a "session" is initiated. A session is finished on four occasions: When the URL is changed, the tab is deactivated, the browser is closed, or the computer ceases to be in "awake" mode. A single row of the desktop panel data corresponds to a session. A session contains the timestamp of when the session started, the URL associated with the session, and the duration (the time spent between the beginning and the end of the session). Our desktop panel ($N$ = 76,677) had the following sociodemographic characteristics:

- Gender – 62.8% of participants were women.
- Race/Ethnicity – 17.7% of participants were Black, 3.8% of participants were Asian or Pacific Islanders, 1.3% of participants were American Indian or Alaska natives, 12.9% of participants were Hispanic.
- Education – 28.7% of participants had higher school-level education or lower, 41.1% of participants attended some college, 20.3% of participants were college graduates, 9.9% of participants completed post-graduate degrees.
- Income – 23.9% of participants' households earned less than 25 thousand US dollars per year, 28.4% earned between 25 and 50 thousand US dollars per year, 19.8% earned between 50 and 75 thousand US dollars per year, 12.5% earned between 75 and 100 thousand US dollars per year, 15.3% earned more than 100 thousand US dollars per year.
- Age — 6.2% of participants were between 18 and 20 years old, 43.4% of participants were between 21 and 44 years old, 31.5% of participants were between 45 and 64 years old, 13.9% of participants were 65 years old or older.

**Mobile panel description**

In the mobile panel, users are asked to install a tracking app on their phones. The app runs in the background and records users' Web history and app usage. Like the desktop panel, Nielsen then divides users' phone usage into sessions, each corresponding to a single row of the mobile panel data. However, sessions related to app usage (e.g., opening the Mail app) and Web browsing (i.e., opening a Web browser and surfing on the Web) are different: Web browsing sessions contain the domain of the URL visited, the timestamp of when the session started, and the duration of the session. App usage sessions contain the app's name (e.g., the

Mail app), the starting timestamp, and the session duration. Our desktop panel ($N = 36{,}028$) had the following sociodemographic characteristics:

- Gender – 56.6% of participants were women.
- Race/Ethnicity – 18.3% of participants were Black, 3.5% of participants were Asian or Pacific Islanders, 1.7% of participants were American Indian or Alaska natives, 21.3% of participants were Hispanic.
- Education – 29.7% of participants had higher school-level education or lower, 42.5% of participants attended some college, 18.7% of participants were college graduates, 9.0% of participants completed post-graduate degrees.
- Income – 0.8% of participants' households earned less than 25 thousand US dollars per year, 29.4% earned between 25 and 50 thousand US dollars per year, 17.5% earned between 50 and 75 thousand US dollars per year, 10.4% earned between 75 and 100 thousand US dollars per year, 12.0% earned more than 100 thousand US dollars per year.
- Age — 5.9% of participants were between 18 and 20 years old, 50.9% of participants were between 21 and 44 years old, 35.2% of participants were between 45 and 64 years old, 8.1% of participants were 65 years old or older.

### Data used to study platform-level trends

When analyzing the overall user activity across Parler and other fringe social media, we consider the entire panel between August 2020 and June 2021. In total, there were $N = 76{,}677$ unique participants in the desktop panel and $N = 36{,}028$ unique participants in the mobile panel. Across the considered time span, the average time in the panel was 5.5 months for desktop and 4.5 months for mobile. The Nielsen Company also provides weights accompanying each panel; each individual $i$ is assigned a weight $w_i$ such that the weights map to the US population ($\sum w_i \approx$ number of individuals in the US using desktop/mobile). We use these weights for the analysis done in Fig. 1 of the paper when analyzing the percentage of daily active users (%DAU). We define the %DAU as the percentage of people in the panel that accessed a social media platform (*e.g.*, Parler) or a category of social media platforms (*e.g.*, mainstream) on a given day. If on day $t$ the %DAU for Parler is 1%, this means that per 100 panelists enrolled on day $t$, 1 had a session where Parler was accessed. Note that while we calculate this percentage, we employ the demographic weights provided by the Nielsen Company to adjust our sample.

### Data used to study the user-level impact of deplatforming

When analyzing the effect of deplatforming on active users on Parler, we considered two sets of matched users, namely (1) those who spent over 3 minutes browsing Parler in December 2020

($N^{\text{Treated}}_{\text{Desktop}} = 135$; $N^{\text{Treated}}_{\text{Mobile}} = 209$), termed “treated”; and (2) those who spent over 3 minutes browsing other fringe social networking platforms and less than 3 minutes on Parler over the same period ($N^{\text{Control}}_{\text{Desktop}} = 265$; $N^{\text{Control}}_{\text{Mobile}} = 387$), termed “control”. The outcome we consider for our user-level analysis is the daily activity (1 if a user visited the domain/app associated with a social media platform at least once on the respective day, and 0 otherwise).

#### Further information about panels

Panelists receive up to \$60 in rewards points per year and participate in monthly \$10,000 sweepstakes that are spread over 400 winners (top prize earners win \$1,000). The panels rely on convenience recruitment with demographic targets. According to Nielsen, panels under-represent males, persons 18–24 and older than 65 years old, as well as incomes under \$75,000 per year. The skew is corrected by weighting to the universe of smartphone and tablet owners, separated by the operating system (for the mobile panel) and to the universe of individuals with access to a desktop computer at home or work (for the desktop panel). For the mobile panel, age, gender, race, Hispanic ethnicity, and income are used as controls. For the desktop panel, controls are gender, age, education, household size, income, Hispanic ethnicity, working status, designated metropolitan area, and Spanish-language dominance. There may be users in both panels, although we do not have this information at hand. Yet, we do not foresee this will impact the results obtained. In contrast, we argue that having analyses on two different panels (even if they share panelists) strengthens our findings.

#### List of mainstream social media considered

YouTube, Facebook, Instagram, Pinterest, LinkedIn, Snapchat, Twitter, WhatsApp, TikTok, Reddit, Nextdoor.

### A.5.2 Extended methods: Difference-in-differences

#### Difference-in-differences approach

To estimate the effect $\delta$ of the deplatforming of Parler on users’ social media usage, we use a difference-in-differences (DiD) model:

$$Y_{it} = \gamma P_t + \lambda T_i + \delta P_t T_i + \epsilon_{it}, \tag{A.1}$$

where the daily usage $Y_{it}$ of user $i$ on day $t$ is determined by whether the day $t$ came after the deplatforming of Parler ($P_t$) and whether the user was an active consumer of Parler before the intervention ($T_i$). With this specification, we estimate the coefficient $\delta$ associated with the interaction between the dummy variables $P_t$ and $T_i$ using OLS to obtain the average treatment effect on the treated (ATT). Given the parallel trends assumption, we have that $\hat{\delta}$ is an estimate of the effect of being active on Parler ($T_i = 1$) on the usage metric after the intervention ($P_t = 1$):

$$\hat{\delta} = E[Y_{ij} \mid P_t = 1, T_i = 1] - E[Y_{ij} \mid P_t = 1, T_i = 0]. \tag{A.2}$$

Following Yang et al. [291], we make our results more robust by estimating our DiD model using weights generated by coarsened exact matching [105] (CEM) and cluster standard errors at the level of the user $i$ [since the daily activity time-series of each user may be autocorrelated, see [17]]. Our ability to causally identify ATTs with the described DiD strategy is predicated on a key identifying assumption: in the absence of the deplatforming event (the "treatment"), the difference between users that were active on Parler and those that were not (treated and control groups) remains constant over time for the considered outcomes. Time series of outcomes of different matched samples shown in Fig. 2 of the paper suggest that this assumption is plausible, as the time series for treated and control groups appear to move in parallel before the deplatforming of Parler. This assumption is relaxed due to our usage of coarsened exact matching, as the CEM-based results are valid as long as any differences in how the two groups would have evolved in the absence/presence of the intervention are entirely explained by the panelist characteristics on which we match.

### Coarsened exact matching

To perform coarsened exact matching, we assign each panelist to a stratum based on their age, race, ethnicity, gender, education level, income, and pre-intervention level of consumption of fringe social media [binned using Scott's normal reference rule [233]]. For each panelist $i$ in a stratum $s$ containing a mixture of panelists that were and were not active Parler users before deplatforming ("treated" *vs.* "untreated"), we construct a CEM weight

$$w_i = \begin{cases} 1 & \text{if } T_i = 1, \\ \frac{N_{T=0}}{N_{T=1}} \frac{N^s_{T=1}}{N^s_{T=0}} & \text{if } T_i = 0, \end{cases} \tag{A.3}$$

where $N_{T=1}$ ($N_{T=0}$) is the total number of treated (untreated) panelists and $N^s_{T=1}$ ($N^s_{T=0}$) is the number of treated (untreated) panelists in stratum $s$. We perform this matching with the R package *MatchIt* [93].

### Placebo testing

To test the robustness of the parallel trends assumption, we carry out a placebo test. We use the same control and treatment groups as before and the same difference-in-differences models specified in Eq. A.1. However, we consider the first 15 days of December as the pre-treatment period (*i.e.*, $P_t = 0$ for $t$ between December 1 and 15, 2020) and the last 16 days of December as the post-treatment period (*i.e.*, $P_t = 1$ for $t$ between December 16 and 31, 2020). Since there was no intervention on December 15, the parallel trends assumption here implies that there

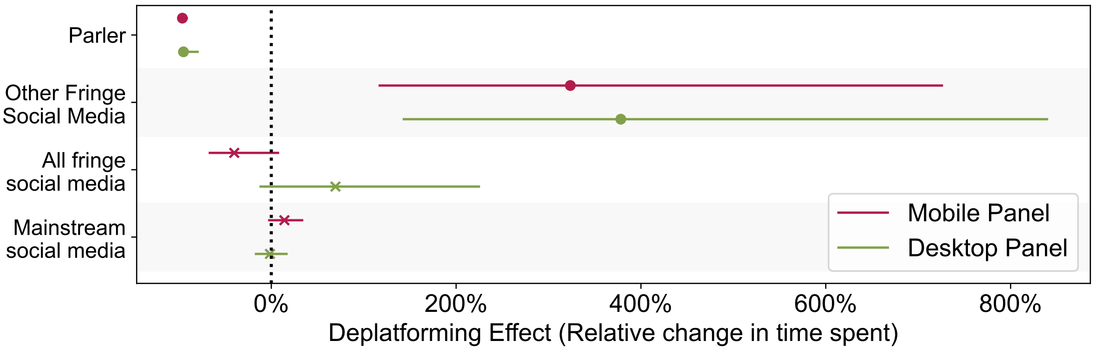


Figure A.10: Difference-in-differences analysis considering time spent on each set of platforms as the outcome. We use a Poisson regression instead of a linear regression (see main text).

should be no significant difference between treatment and control groups. This is indeed what we find: all coefficients obtained are small (smaller than 2.5 percentage points) and not statistically significant ($p > 0.05$).

### A.5.3 Extended results

Fig. A.10 shows the results of the same DiD analysis in the chapter considering another outcome—the total time in seconds users spent on Fringe platforms. Here, however, since the outcome variable is heavily skewed, we use a Poisson regression. We report effect sizes and confidence intervals as relative changes, e.g., a −5.7% effect means that treated units experienced a relative decrease of 5.7% relative to control units.[II]

[II] Note that, to obtain effects as relative change, we simply transform the estimated effects using the formula $(e^{\beta} - 1)$. Given that $\beta = \log \frac{E(Y|Z=1)}{E(Y|Z=0)}$, we have $(e^{\beta} - 1) = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(Y|Z=0)}$.

# Bibliography

[1] Crystal Abidin and Jing Zeng. “Subtle Asian Traits and COVID-19: Congregating and commiserating as East Asians in a Facebook group”. In: *First Monday* (2021).

[2] Saharsh Agarwal, Uttara M Ananthakrishnan, and Catherine E Tucker. “Deplatforming and the control of misinformation: Evidence from Parler”. In: *Available at SSRN* (2022).

[3] Kabir Ahuja et al. “Ordering in: The rapid evolution of food delivery”. In: *McKinsey & Company* (2021).

[4] Milam Aiken and Bennie Waller. “Flaming among first-time group support system users”. In: *Information & Management* (2000).

[5] Yavuz Akbulut, Yusuf Levent Şahin, and Bahadır Erişti. “Cyberbullying victimization among Turkish online social utility members”. In: *Educational Technology & Society* (2010).

[6] Shiza Ali et al. “Understanding the effect of deplatforming on social networks”. In: *Proceedings of the ACM Web Science Conference*. 2021.

[7] Max Aliapoulios et al. “A large open dataset from the Parler social network”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2021.

[8] Virgílio Almeida, Fernando Filgueiras, and Francisco Gaetani. “Digital governance and the tragedy of the commons”. In: *IEEE Internet Computing* (2020).

[9] Joshua D Angrist and Guido W Imbens. “Two-stage least squares estimation of average causal effects in models with variable treatment intensity”. In: *Journal of the American statistical Association* (1995).

[10] Joshua D Angrist and Jörn-Steffen Pischke. “Getting a little jumpy: Regression discontinuity designs”. In: *Mostly harmless econometrics*. 2008.

[11] Peter C Austin. “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* (2011).

[12] Axios. *Poll: Most conservatives think social media is censoring them*. https://www.axios.com/2018/08/29/conservatives-social-media-censorship-poll. 2018.

[13] Christopher P Barlett et al. “Cross-cultural differences in cyberbullying behavior: A short-term longitudinal study”. In: *Journal of cross-cultural psychology* (2014).

[14] Mehmet F Bastug, Aziz Douai, and Davut Akca. “Exploring the "demand side" of online radicalization: Evidence from the Canadian context”. In: *Studies in Conflict & Terrorism* (2020).

[15] Jason Baumgartner et al. “The Pushshift Reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2020.

[16] Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. 2018.

[17] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. “How much should we trust differences-in-differences estimates?” In: *The Quarterly journal of economics* (2004).

[18] Md Momen Bhuiyan et al. “NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2021.

[19] Sophie Bishop. “Managing visibility on YouTube through algorithmic gossip”. In: *New media & Society* (2019).

[20] Official google blog. *Introducing the knowledge graph: things, not strings*. https://blog.google/products/search/introducing-knowledge-graph-things-not/. 2012.

[21] David Bromell. “Deplatforming and Democratic Legitimacy”. In: *Regulating Free Speech in a Digital Age: Hate, Harm and the Limits of Censorship*. 2022.

[22] Andrea Broughton et al. “Precarious employment in Europe”. In: *Strasbourg: European Parliament* (2016).

[23] T Bucher and A Helmond. “The affordances of social media platforms”. In: *The SAGE handbook of social media*. 2018.

[24] Heleen Buldeo Rai, Sabrina Touami, and Laetitia Dablanc. “Not all e-commerce emits equally: Systematic quantitative review of online and store purchases’ carbon footprint”. In: *Environmental Science & Technology* (2022).

[25] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. “Don’t look now, but we’ve created a bureaucracy: the nature and roles of policies and rules in wikipedia”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2008.

[26] A Colin Cameron and Douglas L Miller. “A practitioner’s guide to cluster-robust inference”. In: *Journal of human resources* (2015).

[27] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*. 2013.

[28] Robyn Caplan. *Content or context moderation?* 2018.

[29] Doruk Cengiz et al. “The effect of minimum wages on low-wage jobs”. In: *The Quarterly Journal of Economics* (2019).

[30] Pew Research Center. “Key findings about online dating in the US”. In: (2023).

[31] Pew Research Center. *Social media use in 2021*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/. 2021.

[32] Pew Research Center. *The State of Online Harassment.* https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. 2021.

[33] Eshwar Chandrasekharan et al. “The Internet’s hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2018.

[34] Eshwar Chandrasekharan et al. “You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2017.

[35] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. “Trajectories of blocked community members: Redemption, recidivism and departure”. In: *The World Wide Web Conference (WWW).* 2019.

[36] Jonathan P. Chang, Charlotte Schluger, and Cristian Danescu-Niculescu-Mizil. “Thread With Caution: Proactively Helping Users Assess and Deescalate Tension in Their Online Discussions”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2022.

[37] Annie Y Chen et al. “Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels”. In: *Science Advances* (2023).

[38] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. “Antisocial behavior in online discussion communities”. In: *Proceedings of the International Conference on Web and Social Media (ICWSM).* 2015.

[39] Justin Cheng et al. “Anyone can become a troll: causes of trolling behavior in online discussions”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2017.

[40] Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. “Norm violation in online communities–A study of Stack Overflow comments”. In: *Coordination, organizations, institutions, norms, and ethics for governance of multi-agent systems XIII.* 2017.

[41] Hyunyoung Choi and Hal Varian. “Predicting the present with google trends”. In: *Economic record* (2012).

[42] Matteo Cinelli et al. “Dynamics of online hate and misinformation”. In: *Scientific reports* (2021).

[43] Katherine Clayton et al. “Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media”. In: *Political Behavior* (2020).

[44] Nick Clegg. *Facebook Newsroom — Welcoming the Oversight Board.* https://bit.ly/2VVgHU7. 2020.

[45] CNBC. *Twitter permanently bans alex jones and infowars accounts.* https://www.cnbc.com/2018/09/06/twitter-permanently-bans-alex-jones-and-infowars-accounts.html. 2018.

[46] Katie Cohen et al. “Detecting linguistic markers for radical violence in social media”. In: *Terrorism and Political Violence* (2014).

[47] The Conversation. *Banning disruptive online groups is a game of Whac-a-Mole that web giants just won’t win.* http://theconversation.com/banning-disruptive-online-groups-is-a-game-of-whac-a-mole-that-web-giants-just-wont-win-154427. 2021.

[48] The Conversation. *Deplatforming online extremists reduces their followers – but there’s a price.* http://theconversation.com/deplatforming-online-extremists-reduces-their-followers-but-theres-a-price-188674. 2022.

[49] The Conversation. *Does ‘deplatforming’ work to curb hate speech and calls for violence? 3 experts in online communications weigh in.* http://theconversation.com/does-deplatforming-work-to-curb-hate-speech-and-calls-for-violence-3-experts-in-online-communications-weigh-in-153177. 2021.

[50] Dan Cosley et al. “How oversight improves member-maintained communities”. In: *Proceedings of the CHI conference on human factors in computing systems.* 2005.

[51] Amanda LL Cullen and Sanjay R Kairam. “Practicing moderation: Community moderation as reflective practice”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2022.

[52] Dipto Das, Carsten Østerlund, and Bryan Semaan. “"Jol" or" Pani"?: How does governance shape a platform’s identity?” In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2021.

[53] Thomas Davidson et al. “Automated hate speech detection and the problem of offensive language”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2017.

[54] Julian Dibbell. *A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society.* http://www.juliandibbell.com/texts/bungle_vv.html. 1994.

[55] Discord. *Automated moderation in Discord.* https://discord.com/safety/auto-moderation-in-discord. 2024.

[56] Bryan Dosono and Bryan Semaan. “Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit”. In: *Proceedings of the CHI conference on Human Factors in Computing Systems.* 2019.

[57] Karen M Douglas and Craig McGarty. “Identifiability and self-presentation: Computer-mediated communication and intergroup interaction”. In: *British journal of social psychology* (2001).

[58] Nicolas Ducheneaut. “Socialization in an Open Source Software Community: A Socio-Technical Analysis”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2005.

[59] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. “Explicit warnings reduce but do not eliminate the continued influence of misinformation”. In: *Memory & Cognition* (2010).

[60] Laura Edelson. “Content moderation in practice”. In: *J. Free Speech L.* (2023).

[61] Sabine A Einwiller and Sora Kim. “How online content providers moderate user-generated content to prevent harmful online communication: An analysis of policies and their implementation”. In: *Policy & Internet* (2020).

[62] Facebook Oversight Board. *Oversight Board upholds former President Trump’s suspension, finds Facebook failed to impose proper penalty Oversight Board.* https://bit.ly/3Py70DQ. 2021.

[63] Ethan Fast, Binbin Chen, and Michael S Bernstein. “Empath: Understanding topic signals in large-scale text”. In: *Proceedings of the CHI conference on Human Factors in Computing Systems.* 2016.

[64] Casey Fiesler et al. “Reddit rules! Characterizing an ecosystem of governance”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2018.

[65] Fernando Filgueiras and Virgílio Almeida. “The digital world and governance structures”. In: *Governance for the digital world.* 2021.

[66] Claudia I. Flores-Saviaga, Brian C. Keegan, and Saiph Savage. “Mobilizing the Trump Train: Understanding collective action in a political trolling community”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2018.

[67] Forbes. *The Extremists, Conspiracy Theorists, And Conservative Stars Banned From Social Media Following The Capitol Takeover.* https://bit.ly/3xPvSLJ. 2020.

[68] Deen Freelon. “Computational research in the post-API age”. In: *Political Communication* (2018).

[69] Deen Freelon. “Computational research in the post-API age”. In: *Political Communication* (2018).

[70] Deen Freelon, Alice Marwick, and Daniel Kreiss. “False equivalencies: Online activism from left to right”. In: *Science* (2020).

[71] Mingkun Gao et al. “To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers’ Selection and Perception of News Articles”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2018.

[72] Katharine Gelber and Luke McNamara. “Evidencing the harms of hate speech”. In: *Social Identities* (2016).

[73] Andrew Gelman. *The secret weapon.* https://statmodeling.stat.columbia.edu/2005/03/07/the_secret_weap/. 2005.

[74] Sarah A Gilbert. “"I run the world’s largest historical outreach project and it’s on a cesspool of a website." moderating a public scholarship site on reddit: A case study of r/AskHistorians”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2020.

[75] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* 2018.

[76] Tarleton Gillespie. “Governance of and by platforms”. In: *SAGE handbook of social media* (2017), pp. 254–278.

[77] Tarleton Gillespie. “Platforms are not intermediaries”. In: *Geo. L. Tech. Rev.* (2017).

[78] Tarleton Gillespie. “The Relevance of Algorithms”. In: *Media Technologies: Essays on Communication, Materiality, and Society* ().

[79] Debbie Ging. “Alphas, betas, and Incels: Theorizing the masculinities of the Manosphere”. In: *Men and Masculinities* (2019).

[80] Andrew Goodman-Bacon. “Difference-in-differences with variation in treatment timing”. In: *Journal of Econometrics* (2021).

[81] Google. *YouTube community guidelines.* https://bit.ly/2ZAyMa9. 2022.

[82] Robert Gorwa, Reuben Binns, and Christian Katzenbach. “Algorithmic content moderation: Technical and political challenges in the automation of platform governance”. In: *Big Data & Society* (2020).

[83] Robert Gorwa, Reuben Binns, and Christian Katzenbach. “Algorithmic content moderation: Technical and political challenges in the automation of platform governance”. In: *Big Data & Society* (2020).

[84] Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass.* 2019.

[85] James Grimmelmann. “The virtues of moderation”. In: *Yale JL & Tech.* (2015).

[86] Nir Grinberg et al. “Fake news on Twitter during the 2016 US presidential election”. In: *Science* (2019).

[87] Ted Grover and Gloria Mark. “Detecting potential warning behaviors of ideological radicalization in an Alt-Right subreddit”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2019.

[88] Hussam Habib et al. *To Act or React: Investigating Proactive Strategies For Online Community Moderation.* 2019.

[89] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. “Identification and estimation of treatment effects with a regression-discontinuity design”. In: *Econometrica* (2001).

[90] Aaron Halfaker, Aniket Kittur, and John Riedl. “Don’t bite the newbies: how reverts affect the quantity and quality of Wikipedia work”. In: *Proceedings of the 7th international symposium on wikis and open collaboration.* 2011.

[91] Soo-Hye Han and LeAnn M Brazeal. “Playing nice: Modeling civility in online political discussions”. In: *Communication Research Reports* (2015).

[92] Hatebase. *Hatebase*. https://www.hatebase.org/. 2018.

[93] Daniel E. Ho et al. “MatchIt: Nonparametric preprocessing for parametric causal inference”. In: *Journal of Statistical Software* (2011).

[94] Bruce Hoffman, Jacob Ware, and Ezra Shapiro. “Assessing the threat of Incel violence”. In: *Studies in Conflict & Terrorism* (2020).

[95] Manoel Horta Ribeiro, Justin Cheng, and Robert West. “Automated Content Moderation Increases Adherence to Community Guidelines”. In: *Proceedings of the Web Conference (WWW)*. 2023.

[96] Manoel Horta Ribeiro, Justin Cheng, and Robert West. “Post approvals in online communities”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. 2022.

[97] Manoel Horta Ribeiro et al. “Auditing radicalization pathways on YouTube”. In: *Proceedings of the ACM conference on fairness, accountability, and transparency*. 2020.

[98] Manoel Horta Ribeiro et al. “Characterizing and detecting hateful users on twitter”. In: *Twelfth international AAAI Conference on Web and Social Media*. 2018.

[99] Manoel Horta Ribeiro et al. “Deplatforming did not decrease Parler users’ activity on fringe social media”. In: *PNAS Nexus* (2023).

[100] Manoel Horta Ribeiro et al. “Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2021.

[101] Manoel Horta Ribeiro et al. “The evolution of the Manosphere across the Web”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2021.

[102] Homa Hosseinmardi et al. “Causally estimating the effect of YouTube’s recommender system using counterfactual bots”. In: *Proceedings of the National Academy of Sciences* (2023).

[103] Homa Hosseinmardi et al. “Examining the consumption of radical content on YouTube”. In: *Proceedings of the National Academy of Sciences* (2021).

[104] Yiqing Hua et al. “Characterizing alternative monetization strategies on YouTube”. In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2022.

[105] Stefano M Iacus, Gary King, and Giuseppe Porro. “Causal inference without balance checking: Coarsened exact matching”. In: *Political analysis* (2012).

[106] Guido Imbens and Karthik Kalyanaraman. “Optimal bandwidth choice for the regression discontinuity estimator”. In: *The Review of Economic Studies* (2012).

[107] Guido W Imbens and Thomas Lemieux. “Regression discontinuity designs: A guide to practice”. In: *Journal of Econometrics* (2008).

[108] incels.co. *Rules*. https://bit.ly/3yY8wF2. 2018.

[109] Instagram. *How We Address Potentially Harmful Content on Feed and Stories*. https://about.instagram.com/blog/announcements/how-we-address-harmful-content-on-feed. 2022.

[110] Robin Jacob et al. "A practical guide to regression discontinuity". In: *MDRC* (2012).

[111] Yvonne Jewkes and Majid Yar. *Handbook of Internet crime*. 2013.

[112] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. "Does transparency in moderation really matter? User behavior after content removal explanations on reddit". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2019.

[113] Shagun Jhaver, Himanshu Rathi, and Koustuv Saha. "Bystanders of online moderation: Examining the effects of witnessing post-removal explanations". In: *arXiv preprint arXiv:2309.08361* (2023).

[114] Shagun Jhaver et al. ""Did You Suspect the Post Would be Removed?": Understanding User Reactions to Content Removals on Reddit". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2019.

[115] Shagun Jhaver et al. "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2021.

[116] Shagun Jhaver et al. "Human-machine collaboration for content regulation: The case of Reddit automoderator". In: *ACM Transactions on Computer-Human Interaction (TOCHI)*. 2019.

[117] Shagun Jhaver et al. "Online harassment and content moderation: the case of blocklists". In: *ACM Transactions on Computer-Human Interaction (TOCHI)*. 2018.

[118] Shan Jiang, Ronald E Robertson, and Christo Wilson. "Reasoning about political bias in content moderation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.

[119] Rafael Jiménez-Durán. "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". In: *George J. Stigler Center for the Study of the Economy & the State (Working Paper)* (2023).

[120] Nicola F Johnson et al. "Hidden resilience and adaptive dynamics of the global online hate ecology". In: *Nature* (2019).

[121] Naama Kates. *Statusmaxxing Admincel*. https://spoti.fi/3TQ09rO. 2019.

[122] Matthew Katsaros, Kathy Yang, and Lauren Fratamico. "Reconsidering Tweets: Intervening during Tweet Creation Decreases Offensive Content". In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2022.

[123] Susie Khamis, Lawrence Ang, and Raymond Welling. "Self-branding, 'micro-celebrity' and the rise of social media influencers". In: *Celebrity studies* (2017).

[124] Douwe Kiela et al. “The hateful memes challenge: Detecting hate speech in multimodal memes”. In: *Advances in Neural Information Processing Systems*. 2020.

[125] Charles Kiene, Andrés Monroy-Hernández, and Benjamin Mako Hill. “Surviving an "Eternal September" how an online community managed a surge of newcomers”. In: *Proceedings of the CHI conference on Human Factors in Computing Systems*. 2016.

[126] Sara Kiesler, Jane Siegel, and Timothy W McGuire. “Social psychological aspects of computer-mediated communication.” In: *American psychologist* (1984).

[127] Sara Kiesler et al. “Regulating behavior in online communities”. In: *Building successful online communities: Evidence-based social design*. 2012.

[128] Jin Woo Kim et al. “The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity”. In: *Journal of Communication* (2021).

[129] Danny Klinenberg. “Does Deplatforming Work?” In: *Journal of Conflict Resolution* (2023).

[130] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. 2012.

[131] Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. “Accurately detecting trolls in slashdot zoo via decluttering”. In: *2014 IEEE/ACM international conference on advances in social networks analysis and mining*. 2014.

[132] Nihal Kumarswamy, Mohit Singhal, and Shirin Nilizadeh. “Impact of Stricter Content Moderation on Parler’s Users’ Discourse”. In: *arXiv preprint arXiv:2310.08844* (2023).

[133] Cliff Lampe and Paul Resnick. “Slash (dot) and burn: distributed moderation in a large online conversation space”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2004.

[134] Kyle Langvardt. “Regulating online content moderation”. In: *Georgetown Law Journal* (2017).

[135] Noam Lapidot-Lefler and Azy Barak. “Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition”. In: *Computers in human behavior* (2012).

[136] David Lazer. “Studying human attention on the Internet”. In: *Proceedings of the National Academy of Sciences* (2020).

[137] David Lazer, David Choffnes, and Christo Wilson. *Observatory for Online Human and Platform Behavior*. https://www.khoury.northeastern.edu/research_projects/observatory-for-online-human-and-platform-behavior/. 2023.

[138] Alex Leavitt and John J Robinson. “Upvote my news: The practices of peer information aggregation for breaking news on reddit. com”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2017.

[139] David S Lee and Thomas Lemieux. “Regression discontinuity designs in economics”. In: *Journal of economic literature* (2010).

[140] Lawrence Lessig. *Code: And other laws of cyberspace*. ReadHowYouWant.com, 2009.

[141] Rebecca Lewis. “Alternative influence: broadcasting the reactionary right on YouTube”. In: *Data & Society* (2018).

[142] Hanlin Li, Brent Hecht, and Stevie Chancellor. “All That’s Happening behind the Scenes: Putting the Spotlight on Volunteer Moderator Labor in Reddit”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2022.

[143] Hanlin Li, Brent Hecht, and Stevie Chancellor. “Measuring the Monetary Value of Online Volunteer Work”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2022.

[144] Mary Lilly. “’The world is not a safe place for men’: The representational politics of the Manosphere”. phd. Université d’Ottawa/University of Ottawa, 2016.

[145] Chen Ling, Krishna P. Gummadi, and Savvas Zannettou. “"Learn the Facts about COVID-19": Analyzing the Use of Warning Labels on TikTok Videos”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2023.

[146] Claudia Lo. “When all you have is a banhammer: the social and communicative work of volunteer moderators”. phd. Massachusetts Institute of Technology, 2018.

[147] Zhifan Luo. ““Why should Facebook (not) ban trump?”: connecting divides in reasoning and morality in public deliberation”. In: *Information, Communication & Society* (2022).

[148] Matthew N Lyons. “Ctrl-alt-delete: the origins and ideology of the alternative right”. In: *Political Research Associates* (2017).

[149] Renkai Ma and Yubo Kou. “"How advertiser-friendly is my video?": YouTuber’s Socioeconomic Interactions with Algorithmic Content Moderation”. In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2021.

[150] Michalis Mamakos and Eli J Finkel. “The social media discourse of engaged partisans is toxic even when politics are irrelevant”. In: *PNAS Nexus* (2023).

[151] Cameron Martel and David G. Rand. “Misinformation warning labels are widely effective: A review of warning effects and their moderating features”. In: *Current Opinion in Psychology* (2023).

[152] Adrienne Massanari. “#Gamergate and The Fappening: How Reddit’s algorithm, governance, and culture support toxic technocultures”. In: *New media & society* (2017).

[153] Binny Mathew et al. “Hate begets hate: a temporal study of hate speech”. In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2020.

[154] J Nathan Matias. “Preventing harassment and increasing group participation through social norms in 2,190 online science discussions”. In: *Proceedings of the National Academy of Sciences* (2019).

[155] J Nathan Matias. “The civic labor of online moderators”. In: *Internet Politics and Policy conference*. 2016.

[156] Tara Matthews et al. “Goals and perceived success of online enterprise communities: what is important to leaders & members?” In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2014.

[157] Justin McCrary. “Manipulation of the running variable in the regression discontinuity design: A density test”. In: *Journal of Econometrics* (2008).

[158] Reid McIlroy-Young and Ashton Anderson. “From “welcome new gabbers” to the Pittsburgh synagogue shooting: The evolution of Gab”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2019.

[159] Connor McMahon, Isaac Johnson, and Brent Hecht. “The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2017.

[160] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. “Design friction”. In: *Proceedings of the 31st european conference on cognitive ergonomics*. 2019.

[161] Paul Mena. “Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook”. In: *Policy & Internet* (2020).

[162] Merriam-Webster. *Deplatform*. https://www.merriam-webster.com/dictionary/deplatform. 2018.

[163] Meta. *How enforcement technology works*. https://bit.ly/3ZSlmCF. 2022.

[164] Meta. *Meta community standards*. https://transparency.fb.com/policies/community-standards/. 2024.

[165] Meta. *Our Approach to Facebook Feed Ranking Transparency Center*. https://transparency.fb.com/features/ranking-and-content/. 2023.

[166] Meta. *Using post approvals in your group*. https://bit.ly/4a5HLkC. 2022.

[167] MIT Technology Review. *Reddit’s automoderator is the future of the internet, and deeply imperfect*. https://www.technologyreview.com/2019/10/30/75229/reddits-automoderator-automod-catches-more-than-any-human-could-but-its-still-imperfect/. 2019.

[168] John C Mittermeier et al. “Using Wikipedia to measure public interest in biodiversity and conservation”. In: *Conservation Biology* (2021).

[169] Tamar Mitts. *Banned: How Deplatforming extremists mobilizes hate in the dark corners of the Internet*. 2021.

[170] Jonathan T Morgan et al. “Tea and sympathy: crafting positive new user experiences on wikipedia”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2013.

[171] Jeremy Wade Morris. “Music platforms and the optimization of culture”. In: *Social Media+ Society* (2020).

[172] Kevin Munger. “Temporal validity as meta-science”. In: *Research & Politics* (2023).

[173] Kevin Munger and Joseph Phillips. “Right-wing YouTube: a supply and demand perspective”. In: *The International Journal of Press/Politics* (2020).

[174] Edward Newell et al. “User migration in online social networks: a case study on Reddit during a period of community unrest”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2016.

[175] AP News. *Meta, TikTok and other social media CEOs testify in heated Senate hearing on child exploitation*. https://apnews.com/article/meta-tiktok-snap-discord-zuckerberg-testify-senate-00754a6bea92aaad62585ed55f219932. 2024.

[176] NPR. *Facebook, Twitter, Google CEOs Testify Before Congress: 4 Things To Know*. https://www.npr.org/2021/03/25/980510388/facebook-twitter-google-ceos-testify-before-congress-4-things-to-know. 2021.

[177] NPR. *Facebook, Twitter, Google CEOs Testify To Senate: What To Watch For*. https://www.npr.org/2020/10/28/928532702/facebook-twitter-google-ceos-testify-to-senate-what-to-watch-for. 2020.

[178] Dawn Carla Nunziato. “Protecting free speech and due process values on dominant social media platforms”. In: *Hastings LJ* (2022).

[179] Sudhakar V Nuti et al. “The use of google trends in health care research: a systematic review”. In: *PloS one* (2014).

[180] NY Times. *Conspiracy Theories Made Alex Jones Very Rich. They May Bring Him Down.* https://www.nytimes.com/2018/09/07/us/politics/alex-jones-business-infowars-conspiracy.html. 2018.

[181] Abby Ohlheiser. *The Washington Post — Fearing yet another witch hunt, reddit bans pizzagate*. https://wapo.st/2Xvbvae. 2016.

[182] Esteban Ortiz-Ospina. “The rise of social media”. In: *Our World in Data* (2019).

[183] Alfonso Palmer et al. “Overdispersion in the Poisson regression model”. In: *Methodology* (2007).

[184] Kostantinos Papadamou et al. “Disturbed Youtube for kids: Characterizing and detecting inappropriate videos targeting young children”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2020.

[185] Joon Sung Park, Joseph Seering, and Michael S Bernstein. “Measuring the prevalence of anti-social behavior in online communities”. In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2022.

[186] Irene V Pasquetto et al. “Tackling misinformation: What researchers could do with social media data”. In: *The Harvard Kennedy School Misinformation Review* (2020).

[187] Dario Pavllo, Tiziano Piccardi, and Robert West. “Quootstrap: Scalable unsupervised extraction of quotation-speaker pairs from large news corpora via bootstrapping”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2018.

[188] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect.* 2018.

[189] Fabian Pedregosa et al. "Scikit-learn: Machine learning in python". In: *the Journal of machine Learning research* (2011).

[190] Thomas Pellissier Tanon et al. "From freebase to wikidata: The great migration". In: *The World Wide Web Conference (WWW)*. 2016.

[191] James W. Pennebaker and Cindy K. Chung. "Computerized text analysis of Al-Qaeda transcripts". In: *A content analysis reader* (2007).

[192] Perspective API. *Jigsaw Perspective*. https://www.perspectiveapi.com/. 2018.

[193] Tiziano Piccardi et al. "On the value of wikipedia as a gateway to the web". In: *Proceedings of the Web Conference (WWW)*. 2021.

[194] Ethan Porter and Thomas J. Wood. "Political Misinformation and Factual Corrections on the Facebook News Feed: Experimental Evidence". In: *The Journal of Politics* (2022).

[195] Project Coral. *Media Coral open source commenting platform*. https://docs.coralproject.net/talk/toxic-comments/. 2020.

[196] r/Incels. *Rules*. https://bit.ly/3iXIA77. 2018.

[197] r/OutOfTheLoop. *Post: "What's up with /r/The_Donald leaving Reddit?"* https://www.reddit.com/r/OutOfTheLoop/comments/6bzv8v/. 2017.

[198] r/The_Donald. *Post: "Bookmark this site"*. https://bit.ly/30brHfm. 2020.

[199] r/The_Donald. *Rules*. https://bit.ly/3k47MIp. 2018.

[200] Rahat Ibn Rafiq et al. "Careful what you share in six seconds: Detecting cyberbullying instances in Vine". In: *2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. 2015.

[201] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. "Quick, community-specific learning: how distinctive toxicity norms are maintained in political subreddits". In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2020.

[202] Adrian Rauchfleisch and Jonas Kaiser. *Deplatforming the Far-right: An Analysis of YouTube and BitChute*. https://papers.ssrn.com/abstract=3867818. 2021.

[203] Reddit. *Content Policy*. https://www.redditinc.com/policies/content-policy. 2023. (Visited on 11/08/2023).

[204] Reddit. *Key facts to understanding reddit's recent API updates*. https://www.redditinc.com/blog/apifacts. 2023.

[205] Reddit. *Moderator code of conduct*. https://www.redditinc.com/policies/moderator-code-of-conduct. 2022.

[206] Reddit. *Reddiquette*. https://support.reddithelp.com/hc/en-us/articles/205926439-Reddiquette. 2023.

[207] Reddit. *What is karma?* https://support.reddithelp.com/hc/en-us/articles/204511829-What-is-karma-. 2023.

[208] Reuters. *Google sued by anti-vax doctor over YouTube ban.* https://www.reuters.com/legal/litigation/google-sued-by-anti-vax-doctor-over-youtube-ban-2022-09-29/. 2022.

[209] Bernhard Rieder et al. “Making a living in the creator economy: A large-scale study of linking on YouTube”. In: *Social Media + Society* (2023).

[210] Hal Roberts et al. “Media cloud: Massive open source collection of global news on the open web”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2021.

[211] Sarah T Roberts. *Behind the screen: content moderation in the shadows of social media.* 2019.

[212] Sarah T Roberts. “Digital detritus: ’Error’ and the logic of opacity in social media content moderation”. In: *First Monday* (2018).

[213] Brishen Rogers. “The social costs of Uber”. In: *U. Chi. L. Rev. Dialogue* (2015).

[214] Richard Rogers. “Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media”. In: *European Journal of Communication* (2020).

[215] Jon Roozenbeek and Adrià Salvador Palau. “I read it on reddit: Exploring the role of online communities in the 2016 US elections news cycle”. In: *Social informatics: 9th international conference, SocInfo.* 2017.

[216] Paul R Rosenbaum and Donald B Rubin. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* (1983).

[217] Leonie Rösner, Stephan Winter, and Nicole C Krämer. “Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior”. In: *Computers in Human Behavior* (2016).

[218] Jonathan Roth et al. “What’s trending in difference-in-differences? A synthesis of the recent econometrics literature”. In: *Journal of Econometrics* (2023).

[219] Minna Ruckenstein and Linda Lisa Maria Turunen. “Re-humanizing the platform: Content moderators and the logic of care”. In: *New media & society* (2020).

[220] Giuseppe Russo, Manoel Horta Ribeiro, and Robert West. “Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2024.

[221] Giuseppe Russo et al. “Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM).* 2023.

[222] Giuseppe Russo et al. "Understanding Online Migration Decisions Following the Banning of Radical Communities". In: *Proceedings of the ACM Web Science Conference*. 2023.

[223] Koustuv Saha et al. "A social media study on the effects of psychiatric medication use". In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2019.

[224] Koustuv Saha et al. "Understanding moderation in online mental health communities". In: *Social computing and social media*. 2020.

[225] Punyajoy Saha et al. "On the rise of fear speech in online social media". In: *Proceedings of the National Academy of Sciences* (2023).

[226] Punyajoy Saha et al. "Short is the Road that Leads from Fear to Hate: Fear Speech in Indian WhatsApp Groups". In: *The World Wide Web Conference (WWW)*. 2021.

[227] Haji Mohammad Saleem and Derek Ruths. "The aftermath of disbanding an online hateful community". In: *arXiv preprint arXiv:1804.07354* (2018).

[228] Matthew J Salganik. *Bit by bit: Social research in the digital age*. 2019.

[229] Arthur D Santana. "Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards". In: *Journalism practice* 8.1 (2014), pp. 18–33.

[230] James J Schlesselman. *Case-control studies: design, conduct, analysis*. 1982.

[231] Anna Schmidt and Michael Wiegand. "A survey on hate speech detection using natural language processing". In: *Proceedings of the fifth international workshop on natural language processing for social media*. 2019.

[232] Angela M Schöpke-Gonzalez et al. "Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors". In: *New Media & Society* (2022).

[233] David W Scott. "On optimal and data-based histograms". In: *Biometrika* (1979).

[234] Joseph Seering. "Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation". In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2020.

[235] Joseph Seering and Sanjay R Kairam. "Who moderates on Twitch and what do they do? Quantifying practices in community moderation on Twitch". In: *ACM Transactions on Computer-Human Interaction (GROUP)*. 2023.

[236] Joseph Seering, Robert Kraut, and Laura Dabbish. "Shaping pro and anti-social behavior on twitch through moderation and example-setting". In: *Proceedings of the ACM on human-computer interaction (CSCW)*. 2017.

[237] Joseph Seering et al. "Moderator engagement and community development in the age of algorithms". In: *New Media & Society* (2019).

[238] Farhana Shahid and Aditya Vashistha. “Decolonizing content moderation: Does uniform global community standard resemble utopian equality or western power hegemony?” In: *Proceedings of the CHI conference on Human Factors in Computing Systems.* 2023.

[239] Qinlan Shen and Carolyn Rose. “The discourse of online content moderation: Investigating polarized user responses to changes in reddit’s quarantine policy”. In: *Proceedings of the third workshop on abusive language online.* 2019.

[240] Craig Silverman. “Viral fake election news outperformed real news on facebook in final months of the us election”. In: *BuzzFeed News* (2016).

[241] Philipp Singer et al. “Why we read Wikipedia”. In: *The World Wide Web Conference (WWW).* 2017.

[242] Shamika N. Sirimanne. “COVID-19 and e-commerce: A global review”. In: *UNCTAD* (2021).

[243] Southern Poverty Law Center. *Alt-right celebrity @Ricky_Vaughn99 suspended from twitter.* https://www.splcenter.org/hatewatch/2016/10/06/alt-right-celebrity-rickyvaughn99-suspended-twitter. 2016.

[244] Dhanya Sridhar and Lise Getoor. “Estimating causal effects of tone in online debates”. In: *International joint conference on artificial intelligence (IJCAI).* 2019.

[245] Karthik Srinivasan. *Paying attention.* Tech. rep. Technical Report 2023, University of Chicago, 2023.

[246] Kumar Bhargav Srinivasan et al. “Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2019.

[247] Wessel Stoop et al. “Detecting harassment in real-time as conversations develop”. In: *Proceedings of the third workshop on abusive language online.* 2019.

[248] John Suler. “The online disinhibition effect”. In: *Cyberpsychology & behavior* 7.3 (2004), pp. 321–326.

[249] John R Suler and Wende L Phillips. “The bad boys of cyberspace: Deviant behavior in a multimedia chat community”. In: *CyberPsychology & Behavior* (1998).

[250] Nicolas P Suzor et al. “What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation”. In: *International Journal of Communication* (2019).

[251] Yla R Tausczik and James W Pennebaker. “The psychological meaning of words: LIWC and computerized text analysis methods”. In: *Journal of language and social psychology* (2010).

[252] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. “Effects of algorithmic flagging on fairness: quasi-experimental evidence from Wikipedia”. In: *Proceedings of the ACM on Human-Computer Interaction (CSCW).* 2021.

[253] The Cut. *How many bones would you break to get laid?* https://bit.ly/2VSTrpD. 2019.

[254] The Daily Dot. *4chan and Reddit users set out to prove Seth Rich murder conspiracy.* https://bit.ly/2VWJ4B8. 2017.

[255] The Guardian. *Facebook bans Alex Jones, Milo Yiannopoulos and other far-right figures.* https://www.theguardian.com/technology/2019/may/02/facebook-ban-alex-jones-milo-yiannopoulos. 2019.

[256] The Guardian. *The Guardian — Reddit bans misogynist men's group blaming women for their celibacy.* https://bit.ly/2W2M0fq. 2017.

[257] The Markup. *Google Has a Secret Blocklist that Hides YouTube Hate Videos from Advertisers—But It's Full of Holes.* https://themarkup.org/google-the-giant/2021/04/08/google-youtube-hate-videos-ad-keywords-blocklist-failures. 2021.

[258] The Washington Post. *Transcript of Mark Zuckerberg's Senate hearing.* https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/. 2018.

[259] thedonald.win. *Post: "I hope if you came from T_D you reserved your reddit username even if you don't plan to use it".* 2020.

[260] thedonald.win. *Rules.* https://bit.ly/3g7IKH9. 2018.

[261] Daniel Robert Thomas and Laila A. Wahedi. "Disrupting hate: The effect of deplatforming hate organizations on their online audience". In: *Proceedings of the National Academy of Sciences* (2023).

[262] Kurt Thomas et al. "SoK: Hate, harassment, and the changing landscape of online abuse". In: *2021 IEEE symposium on security and privacy (SP)*. 2021.

[263] Xu Tian and Joseph Sarkis. "Emission burden concerns for online shopping returns". In: *Nature Climate Change* (2022).

[264] TikTok. *Community guidelines enforcement report.* https://bit.ly/3QVeAb2. 2022.

[265] Amaury Trujillo and Stefano Cresci. "Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald". In: *Proceedings of the ACM on Human-Computer Interaction (CSCW)*. 2022.

[266] Amaury Trujillo and Stefano Cresci. "One of many: Assessing user-level effects of moderation interventions on r/The_Donald". In: *Proceedings of the ACM Web Science Conference*. 2023.

[267] Twitter. *Rules enforcement.* https://bit.ly/406aIZh. 2022.

[268] Twitter. *Twitter rules and policies.* https://help.twitter.com/en/rules-and-policies/. 2024.

[269] Tom Tyler et al. "Social media governance: can social media companies motivate voluntary rule following behavior among their users?" In: *Journal of Experimental Criminology* (2021).

[270] Aleksandra Urman and Stefan Katz. “What they do in the shadows: examining the far-right networks on Telegram”. In: *Information, communication & society* (2022).

[271] David VanDyke. “Coded data: Tracking discursive trends in the January 6 Parler data”. In: *The capitol riots*. 2022.

[272] Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. “Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks”. In: *arXiv preprint arXiv:2306.07899* (2023).

[273] Nicholas Vincent et al. “Measuring the importance of user-generated content to search engines”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2019.

[274] Vox. *YouTube, Facebook, and Apple's ban on Alex Jones, explained.* https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories. 2018.

[275] Vox. *YouTube, facebook, and apple's ban on alex jones, explained.* https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories. 2022.

[276] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Communications of the ACM* (2014).

[277] Michael W Wagner. “Independence by permission”. In: *Science* (2023).

[278] Yixue Wang and Nicholas Diakopoulos. “Highlighting High-quality Content as a Moderation Strategy: The Role of New York Times Picks in Comment Quality and Engagement”. In: *ACM Transactions on Social Computing* (2022).

[279] Washington Post. *Reddit closes long-running forum supporting President Trump after years of policy violations.* https://wapo.st/3ySp2Gv. 2020.

[280] Washington Post. *These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them.* https://wapo.st/3AO7pbl. 2016.

[281] Galen Weld, Amy X Zhang, and Tim Althoff. “What makes online communities ‘better’? Measuring values, consensus, and conflict across thousands of subreddits”. In: *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 2022.

[282] Robert West. “Calibration of google trends time series”. In: *Proceedings of the 29th ACM international conference on information & knowledge management*. 2020.

[283] WHO. *Managing the COVID-19 infodemic.* https://bit.ly/4cqyDsr. 2020.

[284] David Wiener. “Negligent publication of statements posted on electronic bulletin boards: Is there any liability left after Zeran”. In: *Santa Clara L. Rev.* (1998).

[285] Wikidata. *Organization (Q43229).* https://www.wikidata.org/wiki/Q43229. 2024.

[286] Wikipedia. *Manual of style.* https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section.

[287] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing.* 2011.

[288] Matthew Williams. *Virtually criminal: Crime, deviance and regulation online*. 2006.

[289] Kevin Wise, Brian Hamman, and Kjerstin Thorson. "Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate". In: *Journal of Computer-Mediated Communication* (2006).

[290] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. "Ex machina: Personal attacks seen at scale". In: *The World Wide Web Conference (WWW)*. 2017.

[291] Longqi Yang et al. "The effects of remote work on collaboration among information workers". In: *Nature human behaviour* (2022).

[292] Taha Yasseri and Jonathan Bright. "Wikipedia traffic data and electoral prediction: towards theoretically informed models". In: *EPJ Data Science* (2016).

[293] YouTube. *Learn about comment settings*. https://support.google.com/youtube/answer/9483359?hl=en. 2024.

[294] YouTube. *YouTube channel monetization policies*. https://support.google.com/youtube/answer/1311392?hl=en#zippy=. 2021.

[295] Savvas Zannettou. ""I Won the Election!": An empirical analysis of soft moderation interventions on twitter". In: *Proceedings of the International Conference on Web and Social Media (ICWSM)*. 2021.

[296] Savvas Zannettou et al. "Measuring and characterizing hate speech on news websites". In: *Proceedings of the ACM Web Science Conference*. 2020.

[297] Savvas Zannettou et al. "On the origins of memes by means of fringe web communities". In: *Proceedings of the Internet Measurement Conference*. 2018.

[298] Savvas Zannettou et al. "The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources". In: *Proceedings of the Internet Measurement Conference*. 2017.

[299] John Zarocostas. "How to fight an infodemic". In: *The Lancet* (2020).

[300] Justine Zhang et al. "Conversations Gone Awry: Detecting Early Signs of Conversational Failure". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018.

[301] Ethan Zuckerman and Chand Rajendra-Nicolucci. "Deplatforming our way to the alt-tech ecosystem". In: *Knight First Amendment Institute at Columbia University, January* (2021).

**MANOEL HORTA RIBEIRO** manoel.hortaribeiro@epfl.ch

## Education

**Ph.D. in Computer Science** 2019 – present
École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
*Advisor*: Robert West

**M.S. in Computer Science** 2017 – 2019
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
*Advisors*: Wagner Meira Jr, Virgílio Almeida

**B.S. in Computer Science** 2013 – 2017
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil

## Awards and Honors

| | |
|---|---|
| Forbes 30 under 30 (Europe) | 2023 |
| Best Reviewer Award (ICWSM) | 2022, 2023 |
| EPFL Best Teaching Assistant Award | 2022 |
| Facebook Computational Social Science Fellowship ($84,000) | 2021 – 2022 |
| Best Paper Honorable Mention CSCW | 2021 |
| Best Dissertation Brazilian Computing Association (3rd Place) | 2020 |
| Featured in Altmetric's 100 most influential papers | 2019 |
| Google Latin America Research Award ($9,000) | 2018 |

—

Media Coverage: My work has received extensive media coverage from NBC News, WIRED, DER SPIEGEL, Rolling Stone, El País, and various other venues. See go.epfl.ch/manoel-media for more.

## Work Experience

**Microsoft Research**, Redmond, WA, USA 2023
*Research intern* under Siddharth Suri and Teny Shapiro

**Reddit**, San Francisco, CA, USA 2023
*Research contractor* under Sanjay Kairam

**Meta**, San Francisco, CA, USA 2021–2022
*Research contractor* under Justin Cheng

**Facebook**, San Francisco, CA, USA 2021
*Research intern* under Justin Cheng

**EPFL**, Lausanne, Switzerland 2018
*Research intern* under Robert West

**Simula**, Oslo, Norway 2017
*Research intern* under Sagar Sen

## Service

**Area chair** for CHI 2023 (Understanding People – Quantitative Methods).

**PC member** for TTO 2020, TTO 2022, AAAI 2021, ICWSM 2020, ICWSM 2021, ICWSM 2022, ICWSM 2023. IC2S2 2021, IC2S2 2022, IC2S2 2023.

**Ad-hoc reviewer** for ACM Transactions in Social Computing, ACM Transactions on the Web, PLOS One, EPJ Data Science, CSCW 2021, CSCW 2022, CSCW 2023.

## Teaching

**Applied Data Analysis**, CS-401, EPFL 2020 – 2023
Head teaching assistant — 617 students in the 2023 edition.

**Data Visualization**, COM-480, EPFL 2021 – 2022
Teaching assistant — 108 students in the 2022 edition.

**Algorithms and Data Structures III**, DCC005, UFMG 2018
Teaching assistant — Around 100 students.

## Invited Talks

**Does Enterprise Social Media Shape Communication Networks?**
*Microsoft*, Redmond 2023

**Measuring the Impact of Content Moderation?**
*UCL*, London (remote) 2023
*ETH Chair of Systems Design*, Zürich 2022
*Meta Economics Interest Group*, Menlo Park (remote) 2022
*Google*, New York (remote) 2023

**Do Platform Migrations Harm the Effectiveness of Content Moderation?**
*Google*, Zürich 2022
*Facebook Core Data Science*, Menlo Park (remote) 2021
*ETH Immigration Policy Lab*, Zürich 2022

**Auditing Radicalization Pathways on YouTube**
*Raditube Workshop*, Tec de Monterrey, Monterrey (remote) 2023
*Hochschule der Künste Bern*, Bern 2022
*Google*, Zürich 2020
*MILA*, Montréal (remote) 2020

## Selected Publications

For full list, see Google Scholar. Citations: 2166, h-index: 15
⋆: Papers with students I mentored †: Shared authorship

(21) **Causally estimating the effect of YouTube's recommender system using counterfactual bots**
Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, Duncan J Watts
*Proceedings of the National Academy of Sciences of the United States of America*, 2024

(20) **Automated content moderation increases adherence to community guidelines**
Manoel Horta Ribeiro, Justin Cheng, Robert West
*Proceedings of the ACM Web Conference*, 2023

(19) **Deplatforming Parler did not decrease consumption of fringe social media**
Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, Duncan J Watts
*PNAS Nexus*, 2023

(18) **The Amplification Paradox in Recommender Systems⋆**
Manoel Horta Ribeiro, Veniamin Veselovsky, Robert West
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2023

(17) **Quotatives Indicate Decline in Objectivity in US Political News⋆**
Tiancheng Hu, Manoel Horta Ribeiro, Robert West, Andreas Spitz
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2023

(16) **Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning⋆**
Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, Giona Casiraghi
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2023

(15) **Understanding Online Migration Decisions Following the Banning of Radical Communities⋆**
Giuseppe Russo, Manoel Horta Ribeiro, Giona Casiraghi, Luca Verginer
*Proceedings of the ACM Web Science Conference*, 2023

(14) **Post approvals in online communities**
Manoel Horta Ribeiro, Justin Cheng, Robert West
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2022

(13) **Characterizing Alternative Monetization Strategies on YouTube**
Yiqing Hua,† Manoel Horta Ribeiro,† Thomas Ristenpart, Robert West, Mor Naaman
*Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2022

(12) **Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions⋆**
Thorsten Ruprechter, Manoel Horta Ribeiro, Tiago Santos, Florian Lemmerich, Markus Strohmaier, Robert West, Denis Helic
*Nature Scientific Reports*, 2022

(11) **Analyzing the Sleeping Giants Activism Model in Brazil⋆**
Bárbara Gomes Ribeiro, Manoel Horta Ribeiro, Virgílio Almeida, Wagner Meira Jr
*Proceedings of the ACM Web Science Conference*, 2022

(10) **Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels**
Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Robert West *Proceedings of the ACM on Human-Computer Interaction (CSCW)*, 2021

(9) **The Evolution of the Manosphere across the Web**
Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, Savvas Zannettou
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2021

(8) **Sudden Attention Shifts on Wikipedia During the COVID-19 Crisis**
Manoel Horta Ribeiro,† Kristina Gligorić,† Maxime Peyrard, Florian Lemmerich, Markus Strohmaier, Robert West
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2021

(7) **Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube⋆**
Robin Mamié, Manoel Horta Ribeiro, and Robert West
*Proceedings of the ACM Web Science Conference*, 2021.

(6) **YouNiverse: Large-scale channel and video metadata from English-speaking YouTube**
Manoel Horta Ribeiro, Robert West
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2021

(5) **Deep neural network estimated electrocardiographic-age as a mortality predictor**
Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, Sandhi M Barreto, Wagner Meira, Thomas B Schön, Antonio Luiz P Ribeiro
*Nature Communications*, 2021

(4) **Auditing radicalization pathways on YouTube**
Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio Almeida, Wagner Meira Jr
*Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 2020.

(3) **Automatic diagnosis of the 12-lead ECG using a deep neural network**
Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, Thomas B Schön, Antonio Luiz P Ribeiro
*Nature Communications*, 2020

(2) **Message distortion in information cascades**
Manoel Horta Ribeiro, Kristina Gligorić, Robert West
*Proceedings of the ACM Web Conference*, 2019

(1) **Characterizing and detecting hateful users on Twitter**
Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, Wagner Meira Jr
*Proceedings of the International Conference on the Web and Social Media (ICWSM)*, 2018

*— preprints —*

(C) *Prevalence and prevention of large language model use in crowd work*
Veniamin Veselovsky,† Manoel Horta Ribeiro,† Phillip Cozzolino, Andrew Gordon, David Rothschild, Robert West
*Under submission: Communications of the ACM*, 2023

(B) *Protection from evil and good: The differential effects of page protection on Wikipedia article quality⋆*
Thorsten Ruprechter, Manoel Horta Ribeiro, Robert West, Denis Helic
arXiv:2310.12696, *under submission (double-blind)*, 2023

(A) *Cross-community interactions with fringe users increase the growth of fringe communities on Reddit⋆*
Giuseppe Russo, Manoel Horta Ribeiro, Robert West
arXiv:2310.12186, *under submission (double-blind)*, 2023

## Research Mentoring

**Semester projects (EPFL)**

2019: Thomas Spilsbury (Ph.D. at Aalto University)
2020: Olivier Lam (Nestlé), Robin Mamie (Fed. Sup. Court), Stanislas Jouven (Firmenich)
2021: Deniz Ira (Swisscom), Marie Reignier-Tayar (Greenly), Leopaul Boessinger (Oracle)
2022: Francis Murray, Marie Biolková (Swisscom), Jordi Martinell (SBB), Mihaela Berezantev (ELCA)
2023: Mateo Echeverry Hoyos

**Master theses (EPFL)**

| | |
|---|---|
| Julia Majkowska (Google) | 2022 |
| Tiancheng Hu (Ph.D. at Cambridge) | 2022 |
| Maciej Styczen (Uber) | 2023 |
| Leopaul Boessinger (Oracle) | 2023 |
| Francesco Salvi (ongoing) | 2023 |
| Paula Rescala (ongoing) | 2023 |

**Others**

| | |
|---|---|
| Barbara Gomes Ribeiro (undergraduate student; Google) | 2020 |
| Breno Matos (summer intern) | 2020 |
| Veniamin Veselovsky (research assistant) | 2022–2023 |
| Giuseppe Russo (visiting Ph.D. student) | 2023 |
| Thorsten Ruprechter (visiting Ph.D. student) | 2023 |