When Do Social Media Platforms Intervene in Online Hate Communities?

September 24, 2024

Abstract

The practice of social media platforms intervening to ban or suspend online hate communities has received a great deal of recent attention. However, very little scholarly focus has sought to understand the conditions under which social media platforms choose to intervene in these communities. This paper uses a massive dataset of over one thousand online hate communities across four social media platforms to assess the predictors of platform intervention. We find that salient violent offline events, such as the Capitol Riot, are more reliable predictors of deplatforming than the content of online hate community. In fact we did not find evidence that online hate speech was associated with similar interventions, although we find that toxic speech does predict intervention. Overall, our findings emphasize the role of offline political events in producing social media intervention and question the extent to which hate speech is a target for proactive intervention.

1 Introduction

The question of how to handle uncivil and hostile forms of political communication is critical to promoting a healthy democratic society. Many seek to foster a dialogue where individuals can disagree while still ascribing to fundamental norms of civility and tolerance [Gutmann and Thompson, 2004]. Many scholars have warned that hostile rhetoric produces concerning political implications including declining political trust and increased hostility towards political opponents [Lorenz-Spreen et al., 2023, Goovaerts and Marien, 2020, Skytte, 2020].

A great deal of contemporary debate on this topic has centered on ways to handle uncivil and hostile forms of political communication on social media [Tucker et al., 2017, Gillespie, 2018, Mitts, 2021, Mitts et al., 2022]. Some of the most concerning forms of hostile political speech on social media emerge from communities of individuals who openly discuss and bond over their mutual hatred and distrust of other groups [Kim, 2022, Lupu et al., 2022, Long, 2022], using these online spaces to recruit new members and organize violent offline activities [Lupu et al., 2022, Mitts, 2019, Kim, 2022, Gillespie, 2018]. Such groups, whether they are neo-Nazis, conspiracy theorists, or anti-government militia groups consistently rely on social media to organize and plan offline violence, raising the stakes of intervention considerably. These issues lead to questions of how and when social media platforms should intervene in online discourse, as well as the effects of these interventions. The most controversial such interventions have centered around the removal of health misinformation during the COVID-19 pandemic, efforts to ban hate speech, and the deplatforming of former U.S. President Donald Trump.

The effectiveness of these strategies has been questioned [Chandrasekharan et al., 2017, Gillespie, 2018, Jardine, 2019, Thomas and Wahedi, 2023, Klinenberg, 2024]. Social media users adapt, self-censor, and migrate across online platforms in order to circumvent these

efforts [Jardine, 2019, Mitts, 2021]. When mainstream platforms remove content, users often transfer their online communities to other platforms that promise relatively un-moderated spaces [Mitts, 2021, Velásquez et al., 2019].

Yet much less is known about the conditions under which platforms are likely to conduct these interventions [Young, 2022, Pradel et al., 2024, Gillespie, 2018]. Mainstream platforms, such as Facebook, have policies restricting the use of hateful and harmful speech, yet such speech has been shown to persist on those platforms [Lupu et al., 2022]. Do platforms consistently intervene when online communities use hateful and harmful speech? Do increases in the use of such speech predict a greater likelihood of platform intervention? The use of such speech tends to increase in response to offline trigger events [Lupu et al., 2022], and the pressure on platforms to intervene increases when online extremism triggers offline violence. Are platforms more likely conduct these interventions when online discourse and coordination lead to offline violent mobilization, which can increase the salience of online discourse and calls for deplatforming? Are there meaningful cross-platform differences in the predictors of intervention?

Here we analyze these questions using a longitudinal, cross-platform study of 1049 online communities. We tracked these communities across four social media platforms from 2019 to 2021, collecting a corpus of over 23 million public posts and comments. All of these communities regularly post hateful and harmful speech as well as other extremist content. Many of the communities were removed either temporarily or permanently during the study period, while others remained, allowing us make comparisons across online communities and time. We analyze three sets of predictors of intervention. The first is whether intervention efforts are more likely after salient violent offline events, after which platforms often publicly announce their intent to remove certain communities. Second, we look at whether interventions are more likely in response hateful and harmful speech, as both have been at the core of content moderation historically [Gillespie, 2018]. Finally, because our data includes both

mainstream and fringe platforms, we also compare the predictors of platform intervention along this dimension.

We find that highly salient and violent offline events are consistently stronger predictors of platform intervention than the extent to which users post hateful and harmful speech. While certain types of harmful language moderately increase the risk of intervention, hate speech was not generally associated with statistically significant increases in the risk of intervention. Comparing across platforms, we find that Facebook has been consistently more likely to engage in such interventions, but that fringe platforms nonetheless regularly engage in similar intervention efforts.

Ours is the first study we are aware of that compares these differing predictors of platform intervention and does so across a wide scope of online spaces and time. While further work is needed to continue understanding when platforms intervene, our findings have several key implications. Most importantly, we find that the strongest predictors of online interventions are highly salient and violent offline events. While examining the processes by which such events result in pressure on online platforms to intervene is outside the scope of our analysis, our findings suggest that such processes may be the most effective in terms of leading platforms to intervene on a large scale. In the conclusions, we offer suggestions for future research on these processes. Second, we find that many different types of hateful and harmful speech are not consistent predictors of intervention. This suggests that, while some platforms do censor and regulate such speech, their ability (and perhaps will) to proactively and consistently respond to it may be weaker than their reaction to high-profile events. Finally, we observe that platforms that promise users a generally non-interventionist approach are less likely to engage in deplatforming, showing that this disinclination to intervene is consistent and results in a more tolerant approach to the presence of online hate communities. We nonetheless find that even platforms typically labeled as non-mainstream or un-moderated, have engaged in important deplatformings.

2 Social Media Platforms, Online Hate Communities, and Intervention

While popular platforms that are provided by Meta and Google dominate, the social media ecology also includes many other platforms, such as Telegram, Gab, and VKontakte. One feature these platforms have in common is they allow users to form communities in which like-minded individuals share a dedicated online space. Such communities are typically organized around a central theme and allow users to post and comment on each other's posts. This includes Telegram channels and Facebook/Gab groups, but it also includes "fan pages" that center around influential individuals and allowed users to comment and discuss them or related issues.

We focus on online hate communities, which are united by users' mutual antipathy toward some target (often along racial, religious, gender, or national lines) and feature the regular use of harmful and hate speech. Discussions in these communities often focus on negativity toward the out-group target and the perceived superiority of some in-group, thus intensifying feelings of belonging and cohesion [Vollhardt et al., 2006]. By contributing to these discussions, individuals gain a sense of identity, often by using slurs, insider symbols, and euphemisms [Hine et al., 2017, Klein et al., 2007]. Mainstream platforms have consistently publicized policies for restricting the use of certain hateful and harmful speech. On Facebook, for example, certain slurs are automatically censored out such that users cannot include them in their posts and comments.

Scrutiny of these online hate communities and calls for platform intervention have been particularly intense during violent offline events. Examples include the increased mobilization of a militia movement known as the Boogaloos in 2020 and the January 6, 2021 assault on the U.S. Capitol. In response, some social media companies, such as Google and Meta, openly pledged to curb harmful and hate-based content [Jhaver et al., 2021, Ribeiro et al.,

2023]. Interventions occur in many forms, including community removal, individual bans, and the deletion of specific posts or certain words. In some cases, platforms only suspend certain users or communities, allowing their return to the platform after a certain time. When a community is suspended or banned, members can gravitate to different communities and/or different platforms. Thus, while such a a step may discourage user participation on a given platform [Chandrasekharan et al., 2017], users may simply migrate to other platforms and re-engage [Jardine, 2019].

Despite these efforts, online hate communities, hateful speech, and other harmful content remain prevalent online [Lupu et al., 2022, Chandrasekharan et al., 2017, Jhaver et al., 2021]. Gillespie [2018] emphasizes the difficulty social media companies face in effectively policing speech. Content moderation often depends on the users reporting content to the platform, which can then decide on a course of action. Yet this approach is costly, and users may not be particularly inclined to report harmful content, suggesting that many interventions ultimately originate with the platform Pradel et al. [2024]. Platforms also use algorithmic interventions and attempt to manage content automatically, yet these approaches are limited by their ability to identifying tone, intent, and context [Gillespie, 2018, Young, 2022].

Scholars have begun to assess the efficacy of these intervention efforts [Chandrasekharan et al., 2017, Jardine, 2019]. Some have asked whether intervention events manage to reduce the sort of speech in question and successfully ban communities, or whether they simply lead members of such communities to change their terminology or gravitate to less interventionist platforms [Gibson, 2019, Grondahl et al., 2018]. Others have looked at the process of content moderation [Young, 2022, Pradel et al., 2024, Gillespie, 2018].

Much less is known about the conditions under which platforms actually intervene in practice. Some interventions follow public announcements by platforms, such as Facebook's announcement in June of 2020 that it would remove content and communities associated

with the extremist Boogaloo movement [1]. In Meta's own discussion of efforts to police violent content, the company discussed its approach of "strategic network disruption (SND)", which involved direct interventions beyond the "routine content enforcement." The stated goal of these SNDs is to "ban groups that proclaim a hateful and violent mission" and to "remove content that represents, praises or supports them" [2]. When platforms announce interventions, they often state that their intention is to remove hateful and/or harmful content. For instance, Meta pointed to the planned violence of ongoing militia movements in 2020, citing company policy against threats of violence or calls to violent action [3]. Meta offered a similar explanation for bans after the Capitol Riot, including Trump's account suspension [4]. Yet platforms' policies are often opaque and change over time while many interventions are not publicly announced (including many of the interventions found in the data we describe below). Such a lack of transparency may be justified as clear guidelines could serve as a blueprint that enables online hate communities to circumvent them.

We thus do not not which types of factors tend to predict platform interventions. After violent offline events, do platforms follow through on announced efforts to remove online hate communities? Do these interventions consistently occur when these communities post hateful and harmful speech? Understanding these phenomena is important to understanding the broader process of online mobilization and coordination by potentially violent actors. Past literature on ethnic conflict and terrorism has already pointed to the ways that violent actors often organize online. This includes Mitts [2019]'s discussion of Islamic State organizing through social media, especially in response to major newsworthy events, as well as Weidmann [2015]'s observations about the ability of interpersonal communication to spread ethnic conflicts around the world. Similarly, online hate communities with neo-Nazi or White

[1] https://about.fb.com/news/2020/06/banning-a-violent-network-in-the-us/
[2] https://about.fb.com/news/2020/05/combating-hate-and-dangerous-organizations/
[3] https://transparency.fb.com/policies/community-standards/violence-incitement
[4] https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/

supremacist leanings have been found to organize online, especially in response to major moments such as the 2020 Black Lives Matter protests [Lupu et al., 2022] or the 2016 election of Donald Trump [Long, 2022]. While social media companies do not consistently ban every online hate community, understanding how and when they act provides further insight into changes that may be necessary to make intervention more effective, ideally mitigating the effects of hateful political speech and corresponding offline violence.

3 When Do Platforms Intervene?

3.1 Offline Events

During our study period, we considered two clear violent offline events that involved online organization on social media, including the mobilization of the Boogaloo movement during 2020 and the Capitol Riot in 2021 [5]. Each of these incidents were associated with public statements by social media companies intending to intervene and widespread calls for intervention in response to offline violence. Mainstream reporting has pointed to the ways that social media companies did, in fact, ban individuals and online communities in the wake of these events [6] [7]. Not only do these events serve as instances where we would expect social media intervention, but also allow us to explore the relationship between offline violence and online content moderation.

As a consequence, a key predictor of deplatforming may be instances of offline violence organized on social media platforms within online hate communities. If such offline events serve as central predictors of deplatforming, we stress the importance of considering social media moderation in the context of violence organized online, and the susceptibility of platforms to the pressure of activists and media after these events.[8]

- H_1 Intervention events become more common during or following salient violent offline

[5] https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html
[6] https://www.businessinsider.com/people-banned-social-media-platforms-after-capitol-riots-2021-1
[7] https://www.nbcnews.com/tech/internet/facebook-banned-boogaloo-groups-new-research-rcna93424
[8] https://www.cnn.com/2023/01/26/tech/jan-6-committee-social-media-companies/index.html

events.

To test this hypothesis, we consider the organizing of the "Boogaloo" movement during 2020 and the Capitol Riot on January 6, 2021. While such a choice is not intended as an exhaustive list of typical offline events, but rather inidcates two examples during our study period that can test platform sensitivity to salient violent offline events organized at least in part on social media.

The first of these events refers to the rise of the so-called "Boogaloos" during 2020, capturing the widespread organizing of a heavily armed anti-government militia movement that often attended protests with multiple members eventually arrested for violent crimes. Their discourse heavily features jokes and memes, while their members became well known for attending protests with Hawaiian shirts.[9] Specificially, Meta announced a ban on Boogaloo organizing on June 30, 2020, referencing efforts that began months previously. [10] However, it remains unclear whether their announcement was accompanied by actual interventions or if other social media companies were responding to the movement in similar fashion.

The second of these events refers to the Capitol Riot on January 6, 2021, when thousands of pro-Trump protesters gathered in Washington, D.C. in response to debunked allegations of election fraud surrounding the 2020 Presidential election. Many of these protesters proceeded to break through barricades, entering the U.S. Capitol, and calling for violence against elected officials, resulting in several deaths.[11][12]. After these events, there were widespread calls for social media companies to investigate whether participants organized on their platforms and to take action against online hate communities.[13][14] In response, both Meta[15] and Telegram[16]

[9]https://www.adl.org/resources/backgrounder/boogaloo-movement
[10]https://about.fb.com/news/2020/06/banning-a-violent-network-in-the-us/
[11]https://www.cnn.com/interactive/2021/06/us/capitol-riot-paths-to-insurrection/
[12]https://www.govinfo.gov/collection/january-6th-committee-final-report
[13]https://www.washingtonpost.com/technology/2023/01/17/jan6-committee-report-social-media/
[14]https://www.cnn.com/2022/01/06/tech/social-media-january-6-anniversary/index.html
[15]https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/
[16]https://www.motherjones.com/politics/2021/01/telegram-nazi-ban/

announced efforts to ban online hate communities and intervene in their organization and emergence that may lead to violence, while alternative sites like Gab have deflected blame for hate communities forming on their platforms.[17]

Both of these events serve as highly prominent salient violent offline events during our study period. The Capitol Riot captured widespread attention as news stations devoted extensive coverage.[18] As Google Trends data reveals (displayed in Appendix A), public attention was particularly fixed on this event during its immediate aftermath, pointing to the salience of the event, while violence is evident both in the scores of injuries inflicted on police officers[19] and verbal threats to murder politicians during the riot.[20]

Similarly, the Boogaloo movement captured public attention during 2020, as indicated by the surging interest in the Boogaloos as of Google trends data (Appendix A) and numerous articles attempting to situate and explain the suddenly salient significance of the Boogaloos.[21] Such rising significance during May and June 2020, around when Meta announced that they had been trying to ban Boogaloo content, was marked by prominent acts of violence by Boogaloo members[22], including a plot to kidnap the governor of Michigan.[23] Thus, both the mobilization of the Boogaloos and the Capitol Riot serve as high profile incidents that were well known to the public and as violent activities organized through online hate communities.

Therefore, we will test our first hypothesis by looking at the impact of both events:

- $H_1(a)$ Intervention events increased during the Spring and Summer of 2020 when the

[17]https://www.forbes.com/sites/jemimamcevoy/2021/01/14/gab-ceo-denies-responsibility-for-capitol-attack-amid-increased-scrutiny/?sh=5507591c6c84

[18]https://www.hollywoodreporter.com/tv/tv-news/capitol-insurrection-drives-huge-audience-to-tv-news-4113264/, https://www.bbc.com/news/world-us-canada-55574780

[19]https://www.cnn.com/2022/03/07/politics/capitol-police-injuries/index.html

[20]https://www.cnbc.com/2022/06/28/jan-6-hearing-trump-thought-pence-deserved-chants-to-hang-him-aide-says.html, https://apnews.com/article/capitol-riot-pelosi-death-threat-pauline-bauer-18673508efe43ac9e70c3e47cb840d6b

[21]https://www.bbc.com/news/blogs-trending-53018201, https://abcnews.go.com/Politics/boogaloo-movement-recent-violent-attacks/story?id=71295536

[22]https://abcnews.go.com/Politics/boogaloo-movement-recent-violent-attacks/story?id=71295536

[23]https://www.bridgemi.com/michigan-government/star-witness-whitmer-kidnap-plotters-wanted-boogaloo-war-stop-biden

Boogaloo movement was organizing online.

- $H_1(b)$ Intervention events increased following the Capitol Riot on January 6, 2021.

3.2 Differences Between Platforms

Our second hypothesis considers the differences between platforms and their likelihood to intervene in online hate communities. Specifically, we expect a clear difference between mainstream platforms like Facebook and alternative platforms, such as Gab and Telegram. In particular, mainstream platforms boast millions of users and regularly discuss their content moderation strategies, while alternative platforms position themselves as non-interventionist alternatives to mainstream platforms.[24] Alternative sites tacitly or explicitly[25] invite members of online hate communities to join their sites and have a reputation for refraining from intervention in those communities.[26] In turn, mainstream platforms have a prominent social profile, announced interventions in the past and regularly highlight their community guidelines and willingness to take action against violating groups. Therefore, we expect a clear difference in the likelihood to intervene when contrasting mainstream platforms with alternative ones:

- H_2 Facebook is more likely to intervene in online hate communities than alternative platforms, such as Gab and Telegram.

3.3 Harmful Speech and Hate Speech

Hateful speech, otherwise harmful speech, and violent rhetoric are often at the core of demands for and platforms' announced interventions Gillespie [2018]. Specifically, hate speech includes slurs and specific attacks on a variety of marginalized populations, such as racial and ethnic minorities or members of the LGBTQ community. Harmful speech more

[24] https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201

[25] https://www.washingtonpost.com/technology/2021/01/11/gab-social-network/

[26] https://www.pewresearch.org/journalism/2022/10/06/alternative-social-media-sites-frequently-identify-as-free-speech-advocates/

generally refers to a variety of aggressive, toxic, or insulting language, including profanity, irrespective of the inclusion of specific instances of hate speech. Meta points to efforts to police hate speech in its regular Community Standards Enforcement Reports[27] and claims to have been very proactive in policing hate speech, including "violent and dehumanizing language" and slurs.[28]

Given the centrality of hate speech and harmful speech in both demands for platform intervention and major platforms' policies, we expect that online hate communities' risk of intervention increases with these types of speech, leading to the following hypotheses:

- H_3 Social media platforms are more likely to intervene in online hate communities that use greater levels of harmful speech.
- H_4 Social media platforms are more likely to intervene in online hate communities that use greater levels of hate speech.

4 Data and Methods

4.1 Online Hate Communities

We used and expanded on the online hate community data presented in Lupu et al. [2022] and Velásquez et al. [2019]. Data collection began in 2019 and involved members of the research team locating public pages on social media platforms that featured community members engaged in hate speech. Candidate communities were found by searching terms associated with ADL's database of hate symbols.[29] The results of these searches were examined qualitatively to determine whether two or more of the last twenty posts included hate speech, pointing to an online hate community that was included in our data set. Hate speech was defined as either (a) content that would fall under the provisions of the United States Code regarding hate crimes or hate speech according to Department of Justice guidelines;

[27]https://transparency.fb.com/reports/community-standards-enforcement/dangerous-organizations/facebook/

[28]https://transparency.fb.com/reports/community-standards-enforcement/hate-speech/facebook/

[29]https://www.adl.org/resources/hate-symbols/search

or (b) content that supports or promotes Fascist ideologies or regime types (i.e., extreme nationalism and/or racial identitarianism). With such online hate communities identified, we used a snowball method, transitioning to other similar groups and including/excluding them from our data according to the same criteria. We removed non-English content using Google's Compact Language Detector 2. We collected data on these online hate communities from June 2, 2019 through December 31, 2021.

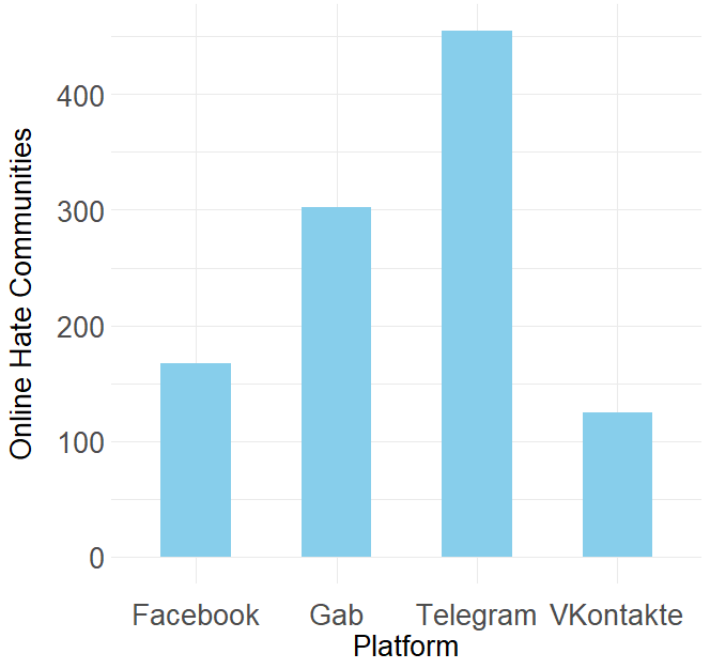

Figure 1: Online Hate Communities by Platform

The resulting data include 1049 online hate communities across Facebook, Telegram, Gab, and VKontakte (VK) (Figure 1). While Facebook is a popular mainstream platform, the other outlets are sometimes considered as alternative sites that host many online hate communities. Telegram and Gab both resemble more mainstream platforms like Facebook, but are known to have more permissive content policies. VK serves as a social networking site primarily for non-American audiences, although there is still a host of English-language content. In total, we collected over 23 million public posts and comments from these com-

munities. This is a particularly broad sample of online communities, advancing prior work that tends to draw data from relatively few communities on moderated platforms.

All data we collected were publicly posted on the Internet, and we had no contact with the users who posted the data. We did not collect any user data, including personally identifying information, user names, and geographic locations. We make no assumptions about users' geographic locations. We obtained IRB approval prior to collecting these data.

4.2 Intervention Events

We focus on intervention events at the community level, such as the removal of entire communities, including permanent bans or temporary suspensions. It is challenging to measure exactly when, where, and how platforms intervene within online hate communities because, as discussed above, platforms do not provide complete information on which online communities they have removed or suspended. We therefore measure platform intervention using the data we collected. Because we sought to continuously collect all posts and comments from all of the online hate communities we identified, we can infer platform intervention from our data. Doing so creates some ambiguity, so we created three measures alternative measures to enhance robustness. In some cases, our ongoing efforts to collect data from an online hate community found that no such data were available after a certain data. Measure 1 codes a social media platform intervention as occurring on the last date such data were collected. This measure raises the possibility that an online hate community was temporarily suspended and re-emerged on the same platform using a different URL other identifier. Therefore, in November 2023, we manually searched for each such community within the applicable social media platform search engine. Measure 2 codes a social media platform intervention as occurring on the last date data were collected from a community if, based on our manual search, the community was no longer online. Both measures raise the possibility that members of a community simply stopping posting, resulting in a lack of data collection that is not due to platform intervention. Measure 3 therefore codes a platform intervention

as occurring on the date in which our data collection software returned error messages indicating that the community was no longer available. The following is a summary of our three measures:

- Measure 1: A platform intervention has occurred when our data collection software no longer detects new content in a community.
- Measure 2: A platform intervention has occurred when our data collection software no longer detects new content in a community and a manual search does not locate the community.
- Measure 3: A platform intervention has occurred when our data collection software returns an error message indicating the community was no longer available.

Each analysis below was replicated with each measure, but the main analyses of this paper focus on the second measure. Intervention events may include outright bans or suspensions, but the second method most closely reflects final bans on these communities. These are of course the most stringent form of intervention and thus the most important to understand in the context of hate communities. We also have the most confidence in the validity of this measure, because it is based on a combination of automated and human-coded information. We only included communities with at least 6 days of collected content, thus looking at only communities where we would expect continued posting. This helps give us confidence that when a community stopped posting, it was because they were intervened in, rather than simply a relatively inactive community.

4.3 Harmful Speech and Hate Speech

There are many possible measures of both hate speech and harmful speech. Past literature has established that these are distinct, with harmful or uncivil speech potentially occurring during healthy political debates and hate speech thriving in homogeneous and prejudiced communities [Rossini, 2020]. We use the supervised machine-learning classifiers created by Lupu et al. [2022]. These classifiers have two key advantages. First, they classify

seven distinct types of hate speech, including speech targeting race, gender, religion, gender identity/sexual orientation (GI/SO), immigration, and ethnicity/identitarian/nationalism (E/I/N), as well as anti-Semitism. Using this measure therefore allows us to determine not only whether more hate speech predicts intervention, but which types of hate speech are more likely to do so, if any. Second, the classifiers were trained using data drawn from both moderated platforms, including Facebook, and less-moderated platforms like Gab, Telegram, and 4Chan. We used these classifiers to measure how many uses of each of the 7 types of hate speech were posted on each of the 1049 online hate communities on each day. The unit of analysis for these estimations are individual posts.

Harmful speech includes aggressive or threatening language that does not contain overt hate speech. We utilize the supervised machine learning approach, presented in Lees et al. [2022] which uses Google's Perspective API. This approach has the advantage of measuring several types of harmful speech, including toxicity, identity-based attacks, insults, profanity, and threats. Toxicity measures the overall hostility of the speech present, looking at "rude, disrespectful, or unreasonable" posts, while identity-based attacks refer to harmful language that is targeted at someone because of their identity. The last three forms of harmful speech are straightforward, measuring insulting language, curse words, and threats of violence against other individuals, respectively.[30] Previous research has used the Perspective API to better understand how to detect harmful speech [Smith, 2022, Jain et al., 2022]. For each online hate community and for each day, Perspective API provides a probability that a person reading that day's content would find the content to meet a given category, e.g., would find the content toxic, profane, etc. The unit of analysis for these estimations are community-days, with all posts on a given day within a given community combined.

[30]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en$_U S$

4.4 Estimation

In order to assess predictors of intervention events, we estimated a series of Cox proportional hazard models. These were conducted on each of the possible intervention effects measures described earlier, although only the one selected in that section will be displayed in the body of the paper, with the full results in Appendix B. Multiple sets of hazard models are estimated on different covariates. The first set were performed looking at the impact of offline events with the second set dedicated to comparing platforms. The next were conducted by including levels of hate speech or harmful language in each online hate communities. The next set combined each of those, with subsequent models including all of these variables. Additional models were estimated looking at each platform individually as needed.

The basic unit of analysis is a given day in a given online hate community, or community-day, for each online hate community that we selected. The data ranged from June 2, 2019 to December 31, 2021. The dependent variable is a binary indicator coded as 1 on the last community-day before an intervention event occurred, and 0 otherwise. Because these models are all estimated on data over time, we also include a count of days for which a given community appears in the data. Initially, we include each set of covariates separately, and the subsequent five subsections each includes models using one set of covariates. This is intended to look specifically at each of these hypothesized factors to determine both if they appear significantly predictive alone. We then describe the results of a series of models combining all variables to determine whether they significantly predict intervention.

Some of the major possible predictors of intervention events are actual periods where social media companies appeared to take concerted efforts to ban certain types of content in response to offline events. Therefore, the first set of models relies on ways to operationalize two major interventions discussed in the hypotheses: efforts to ban Boogaloo content in 2020 and efforts to ban insurrectionist content following January 6, 2021, which have been

well documented across multiple platforms.[31] First, we estimated a model using whether the community-day was between March 1, 2020 through September 1, 2020 as a dummy variable to capture the period that Meta stated an effort to intervene in Boogaloo content. Thus, a community-day between those dates is assigned a 1 and a community-day outside of those daes a 0. In a similar fashion, the second model looks at the period between January 6, 2021 and February 1, 2021 to assess the immediate impact of post-Riot intervention events, while another similar model estimates lasting impacts after January 6, 2021 with a dummy variable for before and after that event. When theoretically necessary, such as with Facebook and the Boogaloo bans, additional models were estimated subset to just one platform.

5 Results

5.1 Preliminary Examination of Intervention Events

We begin with a descriptive analysis of our data. Figures 2 and 3 show the distribution of intervention events across platforms and over time, respectively. Figure 2 shows the number of online hate communities by platform that were subjected to intervention events, displaying each of our measures of intervention. In terms of percentage, this indicates that Facebook and Telegram intervened the most in online hate communities. It is not surprising that Facebook should be likely to intervene in these communities, as it has announced such interventions and is considered a mainstram platform. It may be more surprising to see Telegram's high levels of interventions, as that platform has a reputation for tolerating online hate communities and other extremists [Walther and McCoy, 2021]. Nonetheless, after the events of January 6, 2021, Telegram engaged in a series of bans of neo-Nazi and pro-violent channels, with the company's CEO describing this intervention as "unprecedented."[32] This was in part driven by pressures by external actors, including activists and companies like Apple and Google.

Figure 2 shows interventions over time, displaying the total weekly intervention events

[31] https://www.npr.org/2022/01/06/1070763913/kicked-off-facebook-and-twitter-far-right-groups-lose-online-clout

[32] https://www.motherjones.com/politics/2021/01/telegram-nazi-ban/

throughout our study period. There appear to be multiple periods when intervention events became noticeably more common. The first of these was during the summer of 2020, when the Boogaloo movement was surging. Given that Facebook announced a concerted effort to intervene in Boogaloo organizing efforts during that period, it is possible that this increase is reflective of those efforts.[33] When examining data from the Facebook communities in our data, there is some initial evidence that Facebook increased interventions to target pro-violent Boogaloo communities during the summer of 2020. Two clearly pro-violent communities were banned during that period: "Defend Europa" in July and "Feel the Liberty" in May. The former included clear pro-violent posts referring to shootings or lynchings while the latter espoused calls for revolution, urging members to prepare for violence.

While bans on these and similar communities suggest Facebook was particularly active during this period, many Boogaloo-related communities that expressed clear violent rhetoric survived this period. For instance, the community "Antifa Public Watch North" was not banned until very late in the year (December 2020), despite engaging in repeated pro-violent rhetoric earlier that year, including explicit references to killing enemies.

Figure 3 indicates a notably drastic increase in intervention events following the Capitol Riot. While this is sizable for that entire month, there are also considerable spikes in intervention events throughout the rest of 2021. This may indicates interventions following the Capitol Riot beyond the immediate aftermath of that event. Some Telegram communities were banned very shortly after January 6. Multiple users in communities like "MiloOfficial" and "AtomFront" expressed excitement for upcoming violence in the days leading up to the riot. Both of those communities were banned shortly after January 6, on the 21st and 29th of January, respectively. This period also saw an apparent rise in instances of Telegram displaying a message stating that "This channel can't be displayed because it violated Telegram's Terms of Service", indicating intervention to remove specific posts [Morgia et al., 2023].

[33] https://about.fb.com/news/2020/06/banning-a-violent-network-in-the-us/

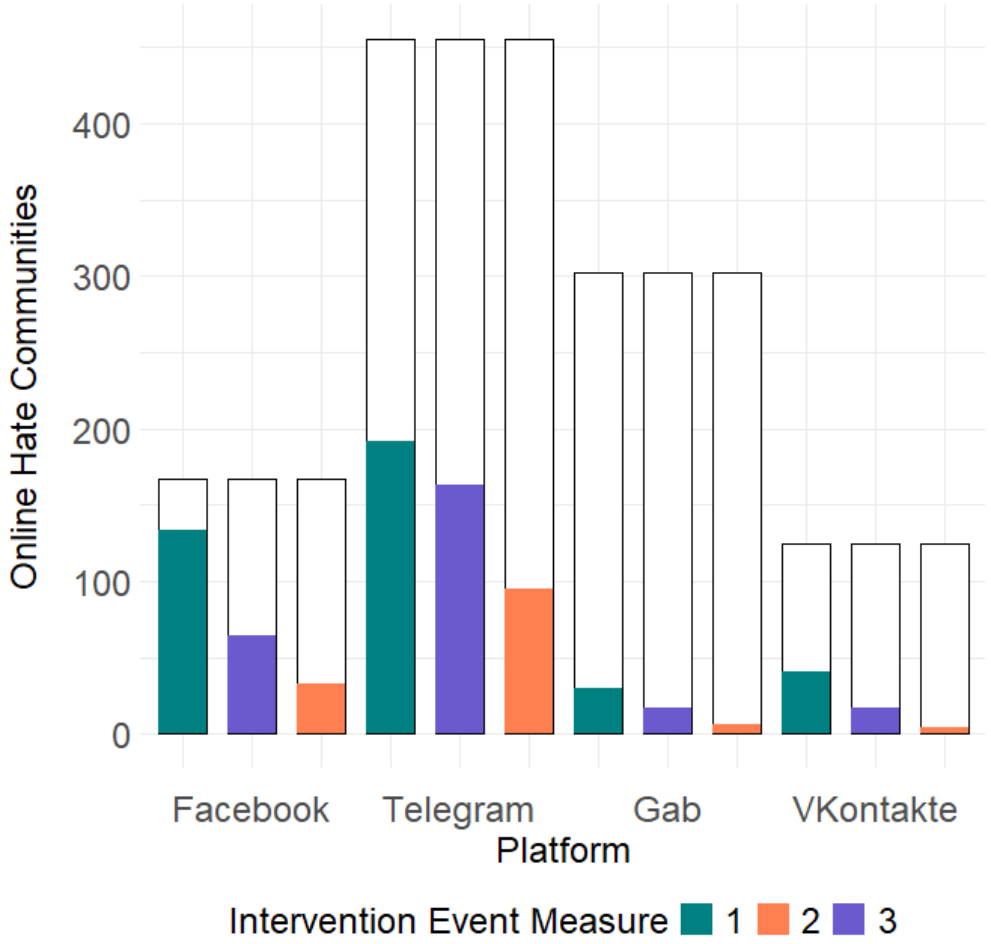

Figure 2: Interventions by Platform

Many other communities, especially Telegram communities that expressed pro-violence and insurrectionist sentiments after the Capitol Riot were banned months or even a year later. This includes a community called "AwakeningGlobal" where a user incorrectly expressed confidence on January 12, 2021 that President-elect Biden would be blocked from becoming president, but the community was not banned until January 2022. Similarly, another community called "patriotpartyflorida" contained fraudulent posts alleging that participants were framed for participating in the Capitol Riot. That community was not banned until March 17, 2021. Thus, there is some initial evidence that pro-insurrectionist sentiment

around the Capitol Riot was related to intervention events, but there is also a great deal of similar content that resulted in no intervention.

5.2 Boogaloo Ban

As discussed, there are two major periods during which we expect to see increases in intervention events. The first was the summer of 2020 when Facebook announced a concerted effort to ban Boogaloo content. We tested this with a model that included an indicator for the period between March 1 and September 1, 2020.

This first model includes only Facebook. Because it is limited to Facebook, this model includes just 30,689 community-days and 64 intervention events across 167 communities. The results of this model are presented in Table 1 and indicate that, at least for Facebook, there was a sizable increase in intervention events during that period. This impact is sizable, with intervention about three times more likely during this period than it was otherwise.[34]

Table 1: Boogaloo Ban (March 1 - Sept 1 2020)

	Facebook	All Platforms
Boogaloo Ban	1.168***	0.121
	(0.269)	(0.163)
Observations	30,689	241,884
Log Likelihood	−281.074	−1,649.891
Note:	*p<0.1; **p<0.05; ***p<0.01	

Next, we test the impact of that period on all four platforms with a similar indicator for the period of March through September 2020. By including Gab, Telegram, and VKontakte, we can have a better sense on whether the mobilization of Boogaloo groups inspired interventions on platforms that did not announce planned interventions. This model includes 241,884 community-days and 261 intervention events across 1049 communities.

[34]This is calculated by exp(1.17), converting the Cox model coefficient to hazard ratios

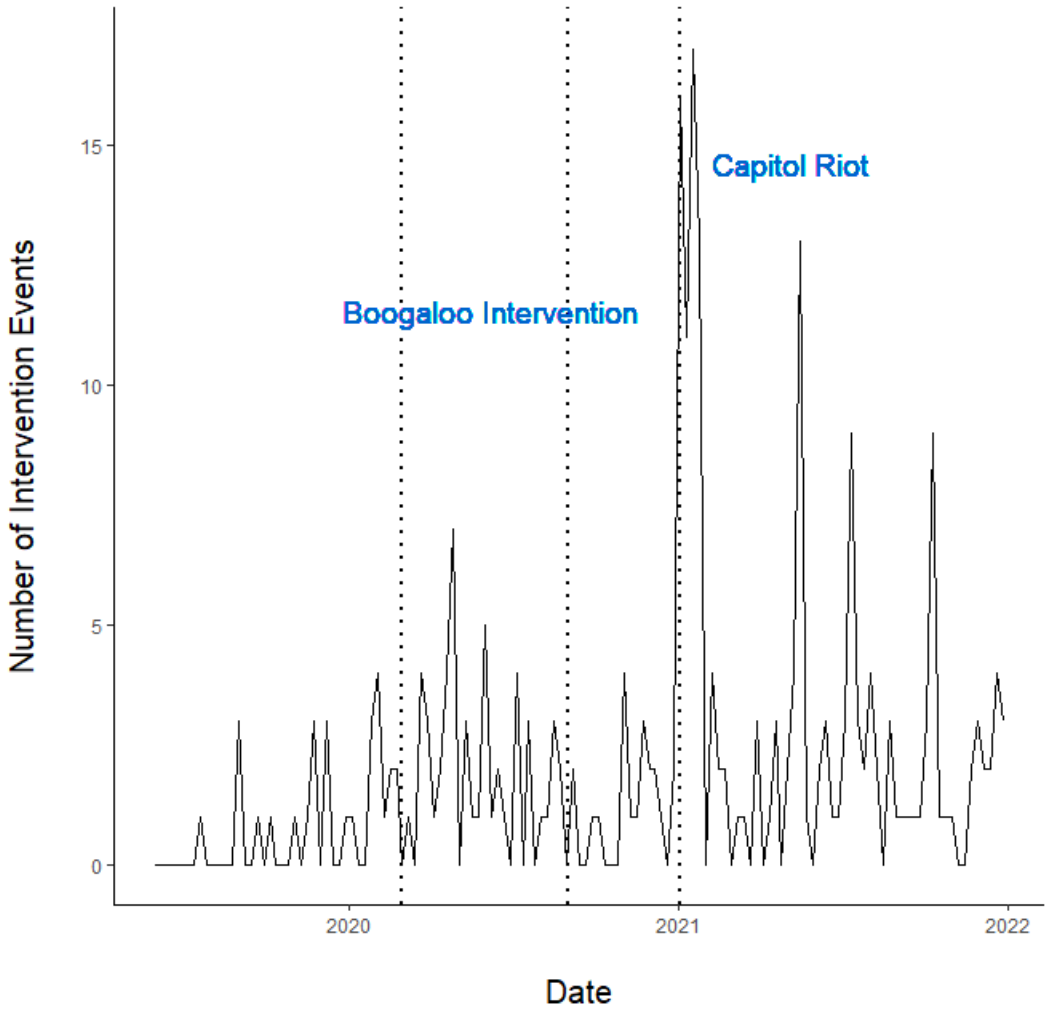

Figure 3: Intervention Events over Time

This second model, displayed in **??** does not find evidence that intervention events increased during that period. This is robust among variations in our measurement of intervention events. We therefore do not find an overall increase in intervention events between March and September 2020 when including social media platforms that did not announce increased intervention.

5.3 Capitol Riot

The next set of models, presented in Table 2, looks both at the immediate and longer-term impact of the interventions announced and discussed following the Capitol Riot. These models include an indicator variable for either the immediate impact (January 6 to January 31, 2021) or the lasting impact (January 6, 2021 through the end of the study period) of the Capitol Riot. Both models include 241,884 community-days and 261 intervention events across 1049 communities. Each of these models provides evidence that the Capitol Riot had a profound impact on increased intervention effects. This was true of the immediate impact which had a consistent and significant effect with intervention being 1.9 times as likely as before,[35] but it was especially true of the lasting impact of this event where intervention became nine times more likely than before the riot.[36]

Table 2: Capitol Riot (All Platforms)

	Immediate Impact	Lasting Impact
Capitol Riot	0.639^{***}	2.208^{***}
	(0.129)	(0.271)
Observations	241,884	241,884
Log Likelihood	−1,637.910	−1,591.600
Note:		*p<0.1; **p<0.05; ***p<0.01

5.4 Differences Between Platforms

Are there important differences between platforms' likelihood of intervention? We analyze this by estimating models using indicator variables for the individual platforms, with Facebook as the baseline. This model utilizes the full 241,884 community-days and 261 intervention events. As the results of the model in Figure 4 indicates, Facebook appears more likely to intervene than any other platform analyzed, although Telegram appears only

[35]This is calculated by exp(0.64), converting the Cox model coefficient to hazard ratios
[36]This is calculated by exp(2.2), converting the Cox model coefficient to hazard ratios

slightly less likely to intervene, between 0.5 and 0.89 times as likely. This is surprising given Telegram's reputation, but as will be discussed, this may be due to Telegram's interventions following the Capitol Riot. Gab appears much less likely to intervene than either Telegram or Facebook, with interventions only 0.14 times as likely as Facebook.

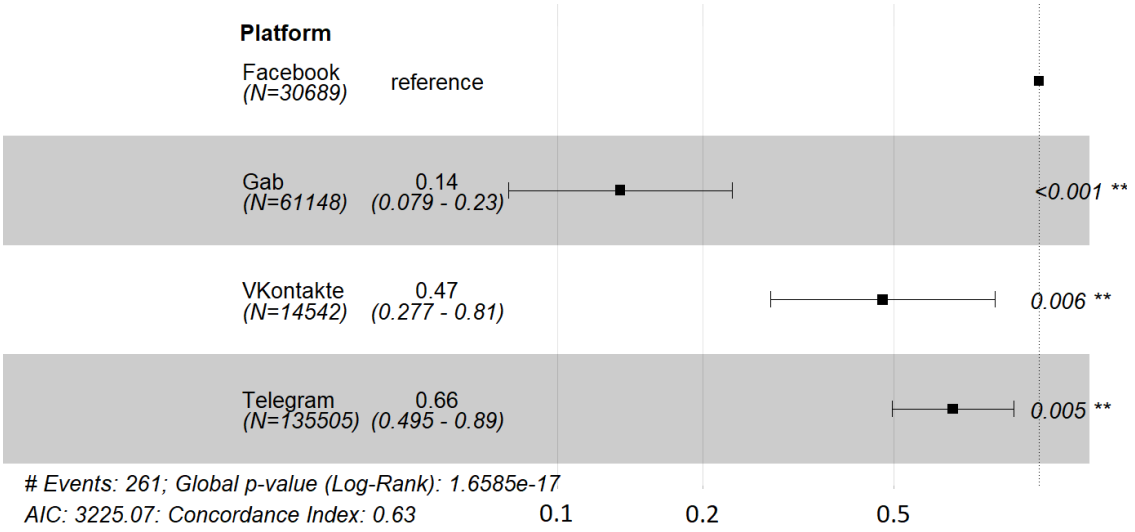

Figure 4: Intervention Risk by Platform, with Facebook as Reference Category

5.5 Harmful Speech

The next set of models examine whether different measures of harmful speech, such as toxicity, contributed to shorter lifespans for the hate communities under discussion with results displayed in Figure 5. This first model only includes our various measures of harmful speech, namely toxicity, identity-based attacks, insults, profanity.

Based on these models, there appears to be consistent evidence that toxicity, as one type of harmful speech, is a predictor of intervention. It appears that moving from 0 to full toxicity results in a dramatic 33 times increase ($B = 3.50$)[37] in the risk of intervention effects for a given community. However, given the large confidence intervals, this increase could be considerably larger or smaller, ranging from a 3.45 times increase to a 278.6 times increase.

[37]This is calculated by exp(3.50), converting the Cox model coefficient to hazard ratios

Regardless of the estimated hazard ratio, it still appears clearly associated with an increased risk of intervention. As shown in Appendix B, this is supported by each of our additional measures of intervention events with one (measure 3) indicating an even higher measure of risk.

Interestingly, across models, including platform subsets for Facebook and Telegram, there also appears to be a consistent and considerable effect for the presence of insults on a given community-day, yet the effect is a smaller likelihood of intervention. Insults are not distinctive markers of online hate communities and are quite common across various types of online communities. In some online hate communities, especially those affiliated with offline organizations, insulting behavior is discouraged as it is considered antithetical to the goals of those organizations. While such online hate communities use fewer insults, they are also more likely to trigger intervention because they are often affiliated with offline groups that engage in violence.

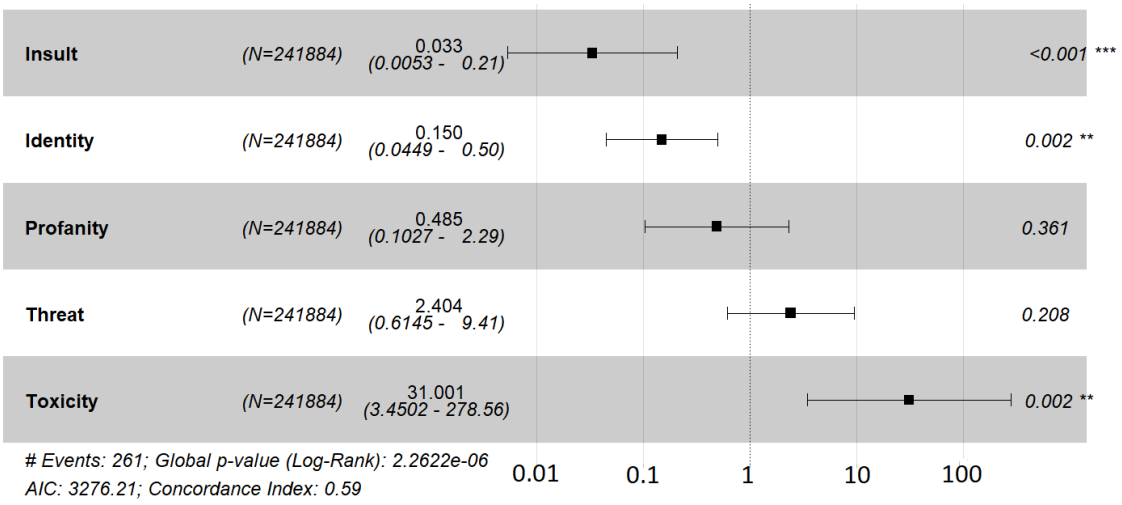

Figure 5: Intervention Risk for Types of Harmful Speech

We also looked at similar models subset only to Facebook, Gab, and Telegram as presented Tables A7, A8, and A9 in Appendix B. Here, the role of harmful speech appears to be significantly predict intervention only for Telegram. While this might indicate that Telegram

is more active in responding to toxic speech, it is likely the case that as a mainstream and more actively moderated platform Facebook automatically removes highly toxic speech as it is posted. In addition, it may be the case that Facebook users are more likely to self-censor because they anticipate moderation by the platform.

5.6 Hate Speech

Figure 6 shows the results of a model that estimates the impact of hate speech. This model relied on 231,947 community days across 260 intervention events.[38] For each community-day, we include a count of the number of posts that contain each type of hate speech. Somewhat surprisingly, the volume of hate speech does not generally appear significant. In fact, there is some evidence of a reduction in intervention risk for hate speech based on gender, immigration, and gender identity/sexual orientation. Nonetheless, the standard errors are quite large and these findings are not replicated across intervention measures.

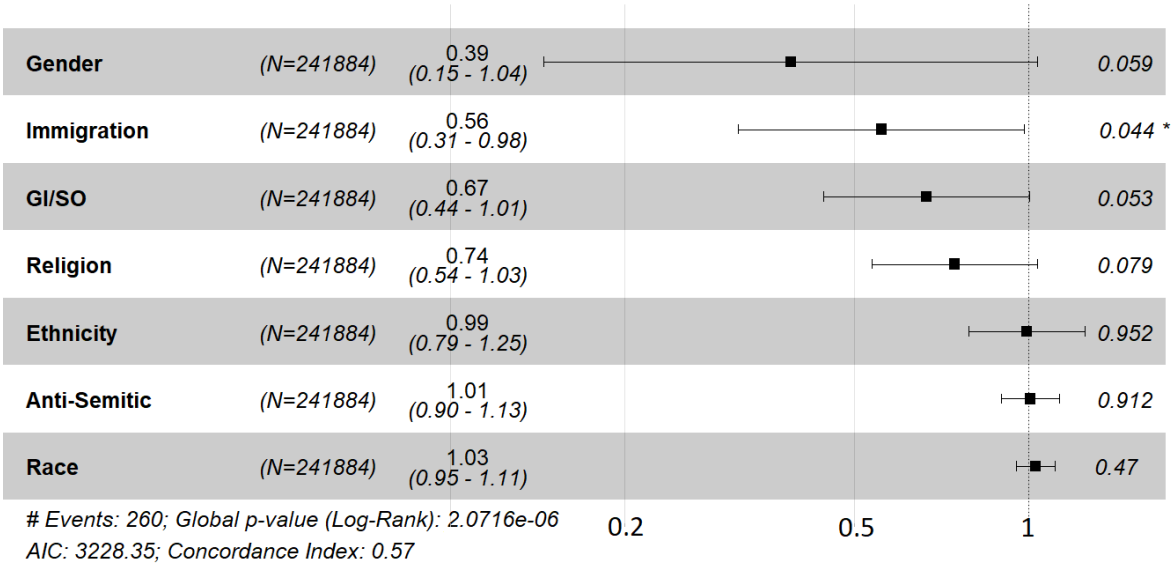

Figure 6: Intervention Risk by Type of Hate Speech

Estimating the effects of hate speech by platform may reveal important nuance. The

[38]Using the classifier provided by Lupu et al. [2022], we were unable to measure hate speech for some observations because they contained text that was too short to classify.

results of these models are displayed in Tables A11, A12, and A13 in Appendix B. The Facebook subset included 30,000 community-days and 64 intervention events across 167 communities, while the Gab subset included 54,416 community-days and 16 intervention events across 302 communities, and the Telegram subset included 133,522 community-days and 163 intervention events across 455 communities. We do not provide results for religion on Facebook, because there were no instances of community-days with both an intervention (under our main operationalization) and religious hate speech. There were instances with other variations of our intervention measure which provided insignificant as shown in Appendix B.

Generally, the volume of most types of hate speech does not significantly predict intervention on these individual platforms. On Facebook, none of the types of hate speech are significant. As with harmful speech, while one might interpret this as indicating that Facebook does not moderate hate speech, it is more likely that Facebook does so proactively. Thus, our interpretation of this result is that the hate speech Facebook does not automatically filter out is not significantly likely to predict intervention against an online hate community. For example, users cannot post on Facebook many of the slurs found on other platforms because Facebook automatically censors them. Interestingly, Gab does appear to be more interventionist in response to racist hate speech (B = 0.81), but it should be noted that this is only significant when using two of our three intervention measures as shown in Table A11.

5.7 Full Results

This section will disentangle the impact of offline events, platform effects, and hate/harmful speech. These models include an intervention measure, either that for the Boogaloo Ban or the Capitol Riot, and each of the other covariates, including each hate speech measure, each harmful speech measure, and platform indicators. Across platforms, this results in models estimated on 231,947 community-days and 260 intervention events.

Table 3 initially reports the results of a model that re-examines the impact of the Facebook's anti-Boogaloo intervention, while including all measures of harmful and hate speech. Religion was again not included in this model as there were no instances of a community-day with both an intervention (under our main operationalization) and religious hate speech. This full model upholds the initial findings across measures of intervention events. This model does not find evidence that volumes of hate speech increase the likelihood of intervention, and this is consistent across platforms and model specifications.

Table 3 also shows results of models estimating the impact of the Capitol Riot, harmful speech, hate speech, and individual platforms. Facebook is the platform baseline. These models support our earlier result that the Capitol Riot had a lasting impact, with communities being over 10 times more likely to experience intervention after January 6, 2021 than before.[39] Notably, moreover, there appears to be a noticeably initial increase during January of that year as reflected in the "immediate impact" of the Capitol Riot, across all three of our intervention models. Communities were 1.6 times more likely[40] to experience intervention during that month than either before or after. Moreover, higher levels of toxicity are associated with a larger risk of intervention, while most types of hate speech are not statistically significant. There is some evidence that anti-immigrant hate speech is associated with smaller risk of intervention. Facebook is significantly more likely to intervene in online hate communities than Telegram, Gab, and VKontakte while controlling for these other factors.

In order to disentangle the effects of the Capitol Riot further, we also examine the lasting effects of the Capitol Riot on Facebook, Gab, and Telegram individually. The results of these models indicate that intervention risk increased following the Capitol Riot on both Facebook and Telegram, and that it increased more so on Telegram. The results in Table 3 not only shows a demonstrative increase during January of that year, and they also suggest that this

[39]This is calculated by exp(2.32), converting the Cox model coefficient to hazard ratios
[40]This is calculated by exp(0.48), converting the Cox model coefficient to hazard ratios

increase lasted, with the Capitol Riot initiating a longer-term effort by different platforms to intervene in online hate communities. Full results are presented in Appendix B.

5.8 Comparing Predictive Power

Appendix C provides full results comparing the predictive value of each model, based on the concordance measures provided in the Cox models. Overall, models that include both measures of harmful speech and salient violent offline events are the most predictive. While the previous analysis emphasized which were most important, these predictive measures suggest that each covariate under discussion has at least some role to play in predicting the likelihood of intervention events.

6 Discussion

Overall, our most consistently robust finding is that salient, violent offline events raise the risk of intervention considerably, whereas most measures of harmful and hateful speech are not statistically significant. After Meta announced a ban on Boogaloo content, intervention events increased significantly on Facebook, and the Capitol Riot appeared to have profound and lasting impacts on both Facebook and Telegram. Telegram's impact is particularly interesting here, as it is known as a very friendly space to online hate communities, as reflected in the large number of online hate communities detected in Figure 1, but appears to have grown increasingly punitive following the events of January 6, 2021. Apparently, even some "alternative" social media sites can still be brought to intervention in such extreme and violent circumstances.

In terms of differences between platforms, our results generally support our hypothesis that platforms like Facebook would be more active interventionists than platforms like Gab. However, counter to that expectation, Telegram was only slightly less interventionist than Facebook. This was clear from Figure 3 but received considerable support from subsequent models, with results holding across models, including our full models in Table 3.

Table 3: Offline Events, Harmful Speech, and Hate Speech

	Boogaloo Ban (FB)	Capitol Riot (Immediate)	Capitol Riot (Lasting)
Offline Event	0.937***	0.483***	2.321***
	(0.270)	(0.131)	(0.273)
GI/SO	−0.340	−0.338*	−0.329
	(0.702)	(0.205)	(0.205)
Race	−0.132	0.022	0.019
	(0.291)	(0.038)	(0.038)
Gender	−6.600	−0.809	−0.780
	(52.341)	(0.497)	(0.499)
Immigration	−0.526	−0.561*	−0.526*
	(0.816)	(0.288)	(0.287)
Ethnicity	−8.168	0.019	0.026
	(61.856)	(0.108)	(0.104)
Anti-Semitic	−9.249	0.042	0.041
	(108.218)	(0.042)	(0.039)
Religion		−0.288*	−0.201
		(0.168)	(0.162)
Toxicity	3.358	3.248***	3.326***
	(2.380)	(1.106)	(1.109)
Identity	−1.902	−1.040*	−0.858
	(1.480)	(0.620)	(0.616)
Insult	−3.425*	−3.318***	−3.277***
	(1.832)	(0.932)	(0.941)
Profanity	−1.129	−0.867	−1.013
	(1.564)	(0.784)	(0.785)
Threat	−0.570	0.697	0.720
	(1.553)	(0.681)	(0.682)
Telegram		−0.515***	−0.776***
		(0.150)	(0.150)
Gab		−1.972***	−2.394***
		(0.282)	(0.282)
VKontakte		−0.904***	−1.044***
		(0.277)	(0.277)
Observations	30,000	231,947	231,947
Log Likelihood	−260.589	−1,551.494	−1,494.860

Note: *p<0.1; **p<0.05; ***p<0.01

Toxic speech is a consistent predictor for intervention events, even when controlling for the differences between platforms. There is some evidence that this is especially true on Telegram, as subsetting to individual platforms provides inconsistent evidence for the impact of toxicity on Facebook and Gab, whereas it continues to have a very large impact on Telegram.

Hate speech and other forms of harmful speech, on the other hand, are not statistically significant predictors of intervention events. Instead, some types of hate speech may even reduce the risk of intervention, and these remain significant when controlling for the platform in question, with hate speech related to gender identity/sexual orientation (GI/SO), immigration, and religion statistically significant in these models.[41] These results should be treated with caution, however, because when we analyze platforms individually we do not find that hate speech is statistically significant.

Overall, our measures were generally reliable at predicting intervention events, with concordance measures close to 0.80 for models incorporating the impact of the Capitol Riot. This indicates that our overall measures do a decent job measuring the likelihood of intervention, and thus presumably capturing how likely a community is to be sufficiently extreme or pro-violent to garner attention. The significant difference in concordance when assessing the post-Capitol Riot period, moreover, underscores how significant that violent offline event was in driving intervention by social media companies, especially Facebook and Telegram. This responsiveness to such offline instances provides a valuable component to understanding what drives social media intervention in online hate communities.

[41] Many online hate communities actively self-censor so as to avoid attention Gibson [2019], and this tendency may be the strongest among those that are the most virulent and pro-violence. Other communities that are less focused on violence and hence perhaps less likely to be targeted by intervention have less need to self censor gratuitous hate speech. Similarly, as discussed in Young [2022], many online platforms rely on users to report material in order to intervene. Many users sensitive to hate speech may be systematically less likely to frequent the sort of online hate communities where it is particularly common, thus less likely to report content and attract intervention.

7 Conclusion

This paper attempted to provide a comprehensive analysis of social media intervention events and their predictors between June 2019 and December 2021. Overall, we do not find that content-related factors such as hate speech or harmful speech are consistent predictors of platform intervention, a particularly striking result given the growing concern over this content. One exception is our measure of toxic speech, which is associated with a larger intervention risk. By contrast, we find that offline violent events that trigger broad calls for platform intervention do predict actions by both mainstream and some fringe platforms.

Our findings help provide a broader understanding of social media platform intervention. These results indicate both that some social media platforms engage in almost no intervention that we detected and that social media platforms primarily act when the salience of violent offline events compels or induces them to act. Telegram, especially, appears to have been driven into action by the Capitol Riot. The aftermath of the Capitol Riot included widespread concern over online hate communities and demands that companies intervene more heavily.

While the question of moderating online political communication is at the core of this paper, there is mixed evidence that social media companies tackle hostile or uncivil communication on an ongoing basis. While some forms of harmful speech do increase the risk of intervention, we were unable to find evidence that any of our studied social media groups took action against communities rife with hate speech. Instead, the primary driver to social media action is the aftermath of violent offline events that lead to pressure on platforms.

This raises pressing questions over whether increased moderation of online hate communities, on an ongoing basis, would have concrete and profound impacts in reducing offline political violence. With the Capitol Riot serving as one of the most notorious events in recent American politics, one questions raised by our results is whether similar events could be

prevented by more proactive and ongoing efforts to curb online hateful and harmful speech.

We hope this paper will serve as an early effort to better understand the conditions under which social media platforms intervene. Given the importance of online political communications, and the growing politicization of platforms' moderation of such communication, it is important for political scientists to continue to research these processes. Under what conditions do offline events leads to the types of pressure on social media platforms that results in intervention? We have explored two such events, but further research is needed on this point. Along similar lines, if there is outside pressure on platforms to regular hateful and harmful speech, why has such pressure not had effects similar to those that we observe after high-profile events? To answer these and other questions, we hope future work will continue to research the relationships between offline events, online content, and platform interventions.

References

E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *PACM on Human-Computer Interaction*, 2017.

A. Gibson. Free speech and safe spaces: How moderation policies shape online discussion spaces. *Social Media + Society*, 2019.

T. Gillespie. *Custodians of the Internet*. Yale University Press, New Haven, 2018.

I. Goovaerts and S. Marien. Uncivil communication and simplistic argumentation: Decreasing political trust, increasing persuasive power? *Political Communications*, 37(6):768–788, 2020.

T. Grondahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is "love": Evading hate speech detection. *ArXiv*, 2018.

A. Gutmann and D. F. Thompson. *Why Deliberative Democracy?* Chicago University Press, Princeton, 2004.

G. E. Hine, J. Onaolapo, E. D. Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Eleventh International AAAI Conference on Web and Social Media*, 2017.

E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan. Adversarial text generation for google's perspective api. *International Conference on Computational Science and Computational Intelligence*, 2022.

E. Jardine. Online content moderation and the dark web: Policy responses to radicalizing hate speech and malicious content on the darknet. *First Monday*, 2019.

S. Jhaver, C. B. D. Yang, and A. Bruckman. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(381):1–30, 2021.

T. Kim. Violent political rhetoric on twitter. *Political Science Research and Methods*, 11: 673–695, 2022.

O. Klein, R. Spears, and S. Reicher. Social identity performance: Extending the strategic side of side. *Personality and Social Psychology Review*, 11(1), 2007.

D. Klinenberg. Does deplatforming work? *Journal of Conflict Resolution*, 68(6):1199–1225, 2024.

A. Lees, V. Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, and L. Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

S. Long. White identity, trump, and the mobilization of extremism. *Politics, Groups, and Identities*, 2022.

P. Lorenz-Spreen, L. Oswald, S. Lewandowsky, and R. Hertwig. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature*

Human Behaviour, 7:74–101, 2023.

Y. Lupu, R. Sear, N. Velasquez, R. Leahy, N. J. Restrepo, B. Goldberg, and N. Johnson. Offline events and online hate. *PLoS ONE*, 2022.

T. Mitts. From isolation to radicalization: Anti-muslim hostility and support for isis in the west. *American Political Science Review*, 113:173–194, 2019.

T. Mitts. Countering violent extremism and radical rhetoric. *International Organization*, 76:251–272, 2021.

T. Mitts, N. Pisharody, and J. Shapiro. Removal of anti-vaccine content impacts social media discourse. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 319–326, 2022.

M. L. Morgia, A. Mei, and A. M. Mongardini. Tgdataset: a collection of over one hundred thousand telegram channels. *Proceedings of the Association for Computing Machinery*, 2023.

F. Pradel, J. Zilinsky, S. Kosmidis, and Y. Theocharis. Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 2024.

M. H. Ribeiro, H. Hosseinmardi, R. West, and D. J. Watts. Deplatforming did not decrease parler users' activity on fringe social media. *Proceedings of the National Academy of Sciences*, 2(3), 2023.

P. Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49:399–425, 2020.

R. Skytte. Dimensions of elite partisan polarization: Disentangling the effects of incivility and issue polarization. *British Journal of Political Science*, 51(4):1457–1475, 2020.

D. Smith. Effectiveness of google's perspective api in self-contained communities. *Hal Open Science*, 2022.

D. R. Thomas and L. A. Wahedi. Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120

(24):e2214080120, 2023.

J. A. Tucker, Y. Theocharis, M. E. Roberts, and P. Barberá. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Journal of Democracy*, 28:46–59, 2017.

N. Velásquez, R. Leahy, N. J. Restrepo, Y. Lupu, R. Sear, N. Gabriel, O. K. Jha, B. Goldberg, and N. F. Johnson. Online hate network spreads malicious covid-19 content outside the control of individual social media platforms. *Scientific Reports*, 11(11549):93–117, 2019.

J. Vollhardt, J. V. M. Coutin, E. Staub, G. Weiss, and J. Deflander. Deconstructing hate speech in the drc: A psychological media sensitization campaign. *Journal of Hate Studies*, 5(1), 2006.

S. Walther and A. McCoy. Us extremism on telegram: Fueling disinformation, conspiracy theories, and accelerationism. *International Review of Law, Computers Technology*, 15: 100–124, 2021.

N. B. Weidmann. Communication networks and the transnational spread of ethnic conflict. *Journal of Peace Research*, 52:285–296, 2015.

G. Young. How much is too much: the difficulties of social media content moderation. *Information Communications Technology Law*, 2022.