The Interconnected Nature of Online Harm and Moderation: Investigating the Cross-Platform Spread of Harmful Content between YouTube and Twitter

VALERIO LA GATTA, Information Sciences Institute, University of Southern California, Los Angeles, California, USA and University of Naples Federico II, Italy

LUCA LUCERI, Information Sciences Institute, University of Southern California, USA

FRANCESCO FABBRI, Pompeu Fabra University, Spain

EMILIO FERRARA, Information Sciences Institute, University of Southern California, USA

The proliferation of harmful content shared online poses a threat to online information integrity and the integrity of discussion across platforms. Despite various moderation interventions adopted by social media platforms, researchers and policymakers are calling for holistic solutions. This study explores how a *target platform* could leverage content that has been deemed harmful on a *source platform* by investigating the behavior and characteristics of Twitter users responsible for sharing moderated YouTube videos. Using a large-scale dataset of 600M tweets related to the 2020 U.S. election, we find that moderated Youtube videos are extensively shared on Twitter and that users who share these videos also endorse extreme and conspiratorial ideologies. A fraction of these users are eventually suspended by Twitter, but they do not appear to be involved in state-backed information operations. The findings of this study highlight the complex and interconnected nature of harmful cross-platform information diffusion, raising the need for cross-platform moderation strategies.

1 INTRODUCTION

Social media platforms play a significant role in shaping the modern digital information ecosystem by allowing users to contribute to discussions on a wide range of topics, including public health, information technology, and socio-political issues. However, the freedom of expression offered by these platforms, combined with lax moderation policies, can potentially threaten the integrity of these information ecosystems when harmful content, such as fake news, propaganda and inappropriate or violent content, is shared and propagates across the digital population. Mainstream social media platforms like Facebook and Twitter attempt to preserve the integrity of their environments by enforcing conduct policies and deploying various moderation interventions to target both harmful content and the users responsible for

spreading it. These interventions can include flagging, demotion, or deletion of content, as well as a temporary or permanent suspension of users.

However, these moderation efforts are typically enacted in a siloed fashion, largely overlooking other platforms' interventions on harmful content that has migrated to their spaces. This approach poses risks as any inappropriate content originating on a *source platform* can migrate to other *target platforms*, gaining traction with specific communities and reaching a wider audience. For example, cross-platform diffusion of anti-vaccine content on YouTube and Twitter has led to extensive amplification and virality on multiple platforms [8, 14]. Additionally, research has shown that moderation efforts on a source platform can foster the proliferation of harmful content on target platforms [26]. For example, the removal of anti-vaccine groups on Facebook has been found to increase engagement with anti-vaccine content on Twitter [22]. Similarly, Buntain et al. [5] found that when YouTube decided to demote conspiratorial content, some Reddit communities pushed demoted videos on the platform making them go viral, effectively nullifying YouTube's strategy. Futhermore, Ali et al. [3] found that users who got banned on Twitter or Reddit exhibit an increased level of toxicity on Gab.

Overall, this highlights the need for cross-platform moderation strategies that consider the interconnected nature of the digital information ecosystem, and that the removal of content or suspension of users on one platform may not be sufficient in addressing the spread of inappropriate content across multiple platforms. Cooperation among social media platforms is therefore desirable but also practically valuable: knowing what content has been deemed inappropriate on another platform can inform moderation strategies, or help with the early detection of similarly harmful, or related content. Recent work demonstrated how cross-platform strategies can help with the moderation of radical content or inauthentic activities [11, 25], e.g., by tracking users' activity on multiple platforms [3, 8, 22].

Contributions of this work

In this paper, we approach this problem from a different perspective. We consider YouTube (YT) and Twitter as the *source* and *target* platforms, respectively, and investigate the prevalence of moderated YT videos on Twitter, i.e., videos that are shared on Twitter but are eventually removed from YouTube. Also, we characterize Twitter users responsible for sharing YT videos—hereafter, YouTube *mobilizers*, as defined by [7]—across several dimensions, including their political ideology and potential engagement with *fringe* platforms.

In particular, we aim to answer the following research questions (RQs):

RQ1: *What is the prevalence, lifespan, and reach of moderated YT videos that are shared on Twitter?*
RQ2: *What are the characteristics of the mobilizers of moderated YT videos? And, are there any differences with the mobilizers of non-moderated YT videos?*
RQ3: *Do the mobilizers of moderated YT videos receive significant engagement from the Twitter population?*

Leveraging a large-scale dataset related to the 2020 U.S. election [6], we observed that YouTube is the most shared mainstream social media platform on Twitter. By using the YouTube API to retrieve videos' metadata, we found that 24.7% of the videos shared in the election discussion were moderated. We also found that these moderated videos spread significantly more than non-moderated videos and were shared more than content from other mainstream and fringe social media platforms, such as Gab and 4chan. When examining Twitter users sharing YT videos, we discovered that the users sharing moderated videos mostly engaged with YT content via retweets, while the users sharing non-moderated videos actively shared YT content in their original tweets or replies. We found that more than half of the users in the former group were suspended by Twitter, but surprisingly, there were more accounts involved

in information operations—as identified by Twitter—in the latter group. Additionally, we found that the users sharing moderated YT videos supported Trump and promoted election fraud claims, while the users sharing non-moderated videos explicitly denounced Trump and had a more uniform political leaning, including users supporting both Biden and Republican representatives who did not endorse Trump's political campaign.

Finally, we found that users sharing moderated and non-moderated YT videos tend to interact within their group and have similar interaction patterns in terms of retweets, suggesting the formation of fragmented communities resembling echo chambers. Overall, our findings provide insights into the complex dynamics of cross-platform information diffusion, highlighting the need for a more holistic approach to moderation.

2 RELATED WORK

2.1 Cross-platform moderation

To preserve the integrity of their own environment, most *mainstream* social media platforms deploy diverse intervention strategies, targeting both inappropriate content (e.g., through flagging, demotion, or deletion) and the users (e.g., through temporary or permanent suspension) who share it. However, the effectiveness of these interventions is increasingly questioned by researchers and policymakers who demand for a proactive and holistic effort rather than the current siloed and retroactive solutions [9, 31]. Indeed, even if the moderation intervention is effective on an isolated platform, it might trigger harmful activities on other platforms. For instance, Ali et al. [3] showed that, following the suspension of radical communities on Reddit, users migrated to alternative platforms becoming more active and sharing more toxic content. A similar pattern emerged following the deplatforming of Parler, where users migrated to other fringe social media platforms such as Gab and Rumble [17]. In addition, Russo et al. [26] have found that the antisocial behaviors of migrated users may spill over onto the mainstream one through other (non-radical) users active across platforms. Accordingly, Mitts et al. [22] discovered that when Facebook banned some anti-vaccine groups, the toxic content promoted by these groups resonated on Twitter. Also, during the 2020 U.S. election, videos removed from *mainstream* platforms were republished on the (low-moderated) BitChute platform [28].

The above-mentioned studies raise the need for proactive and collaborative moderation approaches to guarantee the integrity of the whole digital information ecosystem. In this paper, we investigate the potential benefits of social media platforms to share information concerning their moderation interventions by studying the users who post moderated YT videos on Twitter.

2.2 Cross-platform spread of YouTube content

Cross-platform multimodal information (e.g., images, videos) diffusion is particularly threatening as multimedia content is proven to be much more attractive and credible than only-textual posts [16]. In particular, the cross-posting of (harmful) video content across multiple social media platforms is a well-documented problem in the scientific literature [15, 31]. For example, a considerable number of suspicious YT videos were shared on Twitter to raise skepticism regarding the COVID-19 vaccination campaign [24]. Cinelli et al. [8] show that anti-vaccine YT videos that were shared on Twitter experienced an increased level of visibility and dissemination on YouTube, whereas Nogara et al. [23] recognized YouTube as one of the most prolific channels used by the infamous *Disinformation Dozen* to spread Covid-19-related conspiracies on Twitter. Similarly, Golovchenko et al. [15] have shown that the Internet Research Agency (IRA) leveraged content from YouTube in their 2016 propaganda campaign on Twitter, and, more recently,
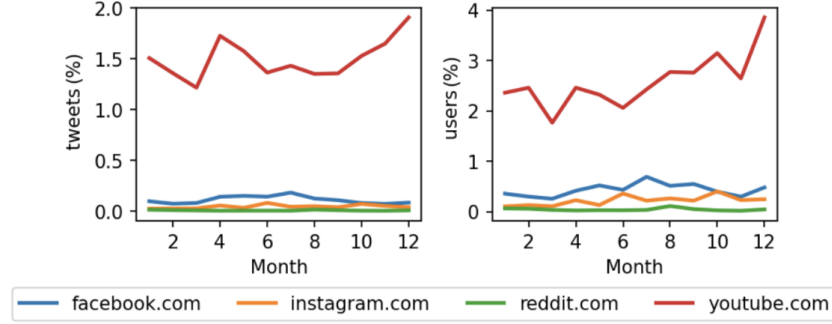
Fig. 1. The monthly percentage of tweets (left) and users (right) who shared a link to a *mainstream* social media platforms in 2020

Wilson and Starbird [31] have also reported the adoption of YT content to support anti-White Helmet operations in 2020.

Overall, the above-mentioned studies demonstrate that YT harmful content is not only engaged within the *source* platform but can flourish across other *target* platforms, often with the intent of influencing vulnerable and fringe communities. In such a scenario, the entities who mobilize and disseminate harmful content can take on various forms, i.e., bots, sockpuppets, influential elites, or even information consumers susceptible to misinformation and conspiracies. In this paper, we study the characteristics of these entities focusing on Twitter users who share moderated YT videos and investigating their similarities and differences with the users who share non-moderated YT content.

3 METHODOLOGY

In this section, we describe the data employed in the analysis and we detail the methodology used to understand the prevalence of YT moderated content on Twitter (*RQ1*) and to characterize the users sharing moderated YT videos (*RQ2* and *RQ3*).

3.1 Data Collection

We employ a dataset of election-related tweets collected using Twitter's streaming API service in the run-up to the 2020 U.S. election [6]. In particular, we focus on the six months, from June 2020 to December 2020, covering the last part of the electoral campaign as well as the aftermath of the election, which was characterized by the widespread diffusion of misleading claims and conspiracies on the integrity of the election results [12, 29]. In this period of observation, we collect more than 600M tweets (including original tweets, replies, retweets, and quotes) shared by 7.5M unique users. In particular, tweets including YT videos account for the 0.65% (3.9M) of the collected messages. Note that we do not consider URLs to YT channels. Fig. 1 shows that the fractions of tweets (resp. users) sharing YT content are consistently larger than tweets (resp. users) pointing to other mainstream social media, which is in line with [1, 7]. In addition, in both cases, we observe an increasing trend towards our observation period (the second half of 2020) possibly because of the approaching election (November 3rd, 2020).

In total, 527k YT videos were shared on Twitter by 830k users. Through the YouTube API, we could retrieve various video metadata, which includes the ability to determine if a particular video was removed from the platform. We find that, among all YT videos, 24.7% (130k out of 527k) were moderated. Besides, before the intervention, those videos
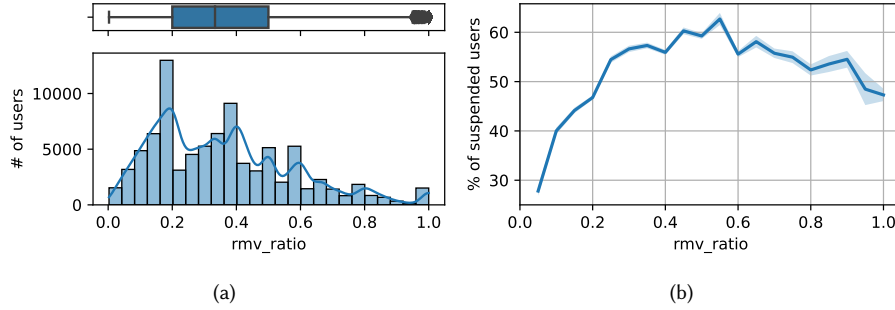
Fig. 2. YT Mobiliziers characteristics with: (a) Distribution of the *rmv(u)*; (b) Percentage of the suspended users with respect to their *rmv(u)*. The shaded area is the 95% confidence interval.

were shared on Twitter by 34.5% (287k out of 830k) users. Finally, we collect YT video metadata of non-moderated videos, including the video title, description, tags, and the channel that published the video. Interestingly, YouTube does not allow to collect metadata for moderated videos, including the date when the intervention occurred as well as the reason(s) for the moderation intervention. However, we can safely assume that a video is still online when shared in an original tweet as this requires the user to report the video URL in the Twitter post.

3.2 Identifying mobilizers of moderated YouTube videos

To identify the mobilizers of moderated YT videos, we first consider the most active YT mobilizers, as we aim to investigate the characteristics and behaviors of users who repeatedly (rather than occasionally) post YT content. For this reason, we consider users that shared at least 5 YT videos on Twitter, which results in a set of 113k Twitter users.

Then, we partition this list of YT mobilizers into two groups based on the volume of shared *moderated* YT videos.

For every user u, we define the *ratio of moderated videos* (*rmv(u)*) as the proportion of moderated YT videos out of the total number of YT videos shared by u during our observation period. Based on this metric, we define non-moderated YT videos mobilizers (**NMYT**) as the users with an $rmv(u) = 0$, which results in 25.4k Twitter accounts. Then, observing the distribution of the *ratio of moderated videos* (Fig. 2a), we define the mobilizers of moderated YT videos (**MYT**) the users with a $rmv(u) \geq 0.5$, i.e., all the users with a $rmv(u)$ higher than the 75-th percentile of the distribution, which results in 14.5k Twitter accounts. This threshold allows us to focus on users particularly prone to sharing moderated videos, thus excluding from the analysis the ones sporadically sharing moderated YT content.

To validate our choice, we examine whether the accounts sharing moderated videos are still active on Twitter or were suspended. Fig. 2b shows the percentage of suspended users as a function of the *rmv(u)*. We note that the percentage of suspended users does not increase after they share more than 50% of moderated videos. Also, the probability of being suspended by Twitter is positively correlated with the value of the *rmv(u)* (Spearman correlation = 0.451).

4 RESULTS

4.1 Prevalence of moderated YouTube videos on Twitter (RQ1)

To answer RQ1, we perform an analysis of the consumption of YT videos on Twitter comparing moderated vs. non-moderated YT videos. It is worth noting that moderated videos tend to be shared for a limited number of days (20 days on average), while non-moderated videos have a longer lifespan (50 days on average), likely due to YT moderation
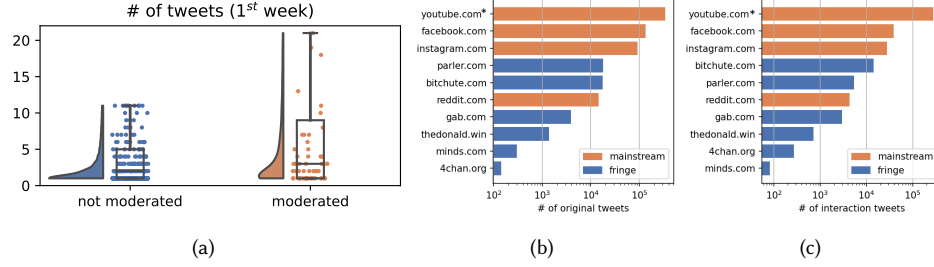
Fig. 3. The prevalence of moderated YT videos with: (a) The distribution of the number of tweets sharing each video during the week after its first share; (b) Number of original tweets containing a link to each social media platform (Log-scale); (c) Number of retweets containing a link to each social media platform (Log-scale)

interventions. Therefore, to perform a fair comparison, we examine the number of tweets including the YT videos during the first week after its first share on Twitter.

Fig. 3a shows the distributions of the number of tweets for non-moderated and moderated videos. It can be noted how moderated videos distribution is characterized by a right heavy-tail, meaning that moderated videos when posted for the first time on Twitter, generate a higher volume of sharing activity with respect to non-moderated videos[1].

To further explore the prevalence of moderated YT content on Twitter, we compare the volume of interactions with content originating from other social media platforms. Specifically, Fig. 3b and Fig. 3c shows the volume of tweets and retweets of moderated YT videos, URLs pointing to *mainstream* online social networks, and URLs redirecting to fringe platforms [30]. We observe that the volume of tweets linking *moderated* YT videos alone is greater than the volume of tweets pointing to any other social media platform. In line with [7], we find that *fringe* content supplied by *Parler* and *BitChute* is outnumbered by the content supplied by mainstream platforms.

4.2 YouTube Mobilizers (RQ2)

To address RQ2, we characterize YT mobilizers that share moderated (MYT) and non-moderated videos (NMYT) and investigate whether these users show significantly different behaviors and characteristics across three dimensions:

- *Cross-Posting Activity*: we explore the sharing activity that users perform on Twitter, including their cross-posting of content originating from other *mainstream* and *fringe* platforms;
- *Trustworthiness of the account*: we examine whether MYT and NMYT users are verified accounts or bots, also looking at their account status (active vs. suspended) and potential involvement in information operations;
- *User Interests*: we investigate the political leaning and topics of interest of mobilizers, both on Twitter and YT.

4.2.1 Cross-Posting. We consider all the possible sharing activities that a user can perform on Twitter, i.e., posting an *original* tweet, commenting on a tweet with a *reply*, re-sharing a tweet with or without a comment via *retweet* or a *quote*, respectively.

From Fig. 4 it can be noticed that, while exhibiting similar behaviors in posting *original* tweets, MYT users retweet more and reply less than NMYT users. To further characterize this discrepancy in the use of retweets, we examine the proportion of re-shared tweets embedding links to other web domains (beyond YouTube). We find that MYT users often retweet content linking to external sources, i.e., 50% of their retweet on average contain URLs with respect to 28% for

---

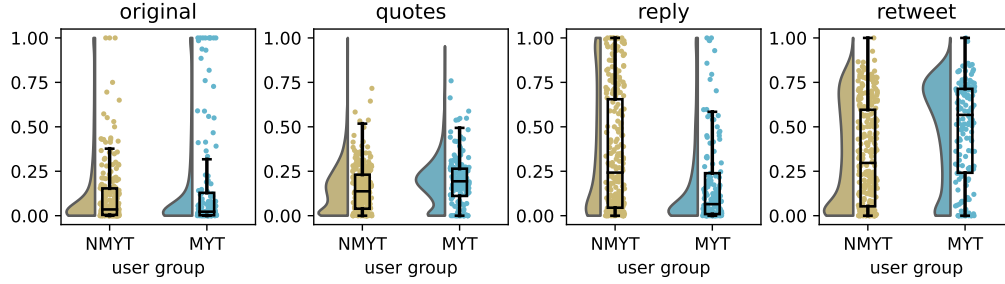[1]A Mann–Whitney test (*p*-value< 0.01) was performed to validate this finding.

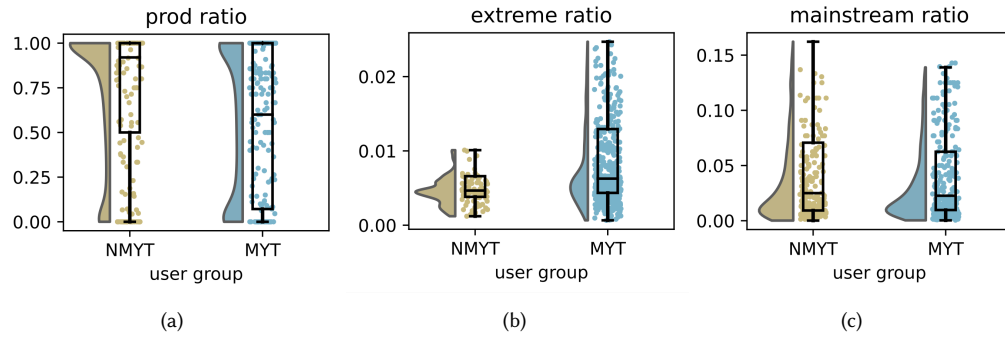Fig. 4. The distribution of original tweets, replies, retweets and quotes for NMYT and MYT mobilizers

Fig. 5. The distribution of *prod_ratio*, *extreme ratio* and *mainstream ratio* for NMYT and MYT mobilizers

NMYT mobilizers. This analysis suggests that NMYT users' activity is more diversified on Twitter while MYT users tend to passively consume and spread (through retweets) the content they see on Twitter, especially if such content links to external resources.

To further characterize the two groups of mobilizers, we consider how they engage with YT videos on Twitter. Specifically, we aim at understanding whether users passively interact with YT videos (e.g., through retweets) or they pro-actively share them (e.g., in their original tweets). To this end, we define users as a *producers* (resp. *consumers*) of a YT video if their first share of that video is an original tweet or a reply (resp. retweet). Then, we define the *prod_ratio* as the proportion of YT videos *produced* by a user out of the total number of videos that he/she engage with. It is worth to note that with "video producer" we do not imply that the user is the publisher of the video on YT.

Interestingly, we find that the mobilizers in both groups are either mostly producers (*prod_ratio* > 80%), or mostly consumers (*prod_ratio* < 20%) of YT videos. Specifically, we count 15,761 (62.1%) producers and 4,439 (17.4%) consumers in the NMYT group, while there are 4,630 (31.9%) producers and 3,742 (25.8%) consumers belong to the MYT group. Fig. 5a shows the distribution of the *prod_ratio* in both groups. On the one hand, we find that NMYT mobilizers are mostly producers, i.e., the 25-percentile of the *prod_ratio* distribution is 0.50. On the other hand, the MYT group includes the same amount of producers and consumers accounts. This result is consistent with our previous finding and confirms MYT users' tendency to passively retweet content if compared to NMYT accounts, who actively participate in the discussion through replies or quotes.

Furthermore, we investigate how mobilizers engage with content from other social media platforms. In particular, we target *mainstream* platforms, i.e., Facebook, Instagram, and Reddit, and the *fringe* platforms outlined in Fig. 3b. We define the *extreme ratio* as the fraction of extreme URLs out of the total number of tweets shared by each user.

Fig. 5c shows that MYT and NMYT mobilizers have similar distributions in terms of *mainstream ratio* (p-value > 0.01 with a Mann-Whitney test). Altogether, MYT and NMYT mobilizers share the same percentage (3.3%) of *mainstream* content on average, and no user is posting more than 15% of tweets linking other *mainstream* social media platforms. However, Figure 5b shows that MYT users engages with *fringe* platforms much more than NMYT mobilizers. Indeed, the distribution of their *extreme ratio* are different according to a Mann-Whitney test (p-value< 0.01), even if they have similar mean values, i.e., 0.5% and 0.7% for NMYT and MYT mobilizers, respectively. Overall, this analysis highlights that the two groups of mobilizers do not exhibit any difference when interacting with *mainstream* social media platforms, whereas MYT mobilizers share more content from low-moderated online spaces than NMYT users, suggesting a form of endorsement towards the extreme ideas pushed on *fringe* platforms [30].

4.2.2   Trustworthiness. Here, we investigate the nature and status of the accounts in the two groups of mobilizers. Given the pivotal role that political elites, bot accounts, state-backed trolls have in orchestrated campaigns and (mis)information operations [15, 21, 23, 32], we aim at identifying the entities pushing moderated or non-moderated YT videos on Twitter. To this end, we leverage Botometer [33] and the Twitter API to assess whether our mobilizers are automated or verified accounts, respectively. As shown in Figure 6, we find that both groups include very few verified accounts (268 and 19 in the NMYT and MYT groups, respectively) and bots (2, 234 and 586 in the NMYT and MYT groups, respectively).

Next, we examine whether users in the two groups of mobilizers were suspended by means of Twitter moderation interventions. Indeed, during the 2020 U.S. election, the platform made an increased effort to guarantee the integrity of discussion by adding warnings to suspicious or misleading content, as well as suspending accounts involved into information operations [20, 27]. From our perspective, we are interested in understanding whether and to what extent the accounts of the MYT and NMYT mobilizers were moderated by these actions. We find that accounts are suspended in both groups but in different proportion: Fig. 6 shows that 53.8% of MYT mobilizers (7,793 accounts) are moderated by Twitter, while 31.4% of NMYT users (7,984 accounts) are suspended. In addition, we further explore whether the (suspended) accounts are involved in state-backed information operations (in short, InfoOps) on Twitter. As reported in Figure 6, we find that a minority of mobilizers (599 accounts in total) are involved in those campaigns and, interestingly, NMYT users are more involved than MYT users (2.2% of MYT against just 0.2% of NMYT). Overall, this analysis suggests that even though MYT mobilizers violated Twitter policies, they are not involved in state-backed, orchestrated campaigns during the election.

4.2.3   User Interests. We now turn our attention to the content shared by MYT and NMYT mobilizers. As the data under analysis relate to the U.S. 2020 Presidential election, we expect that the discussion resonate political topics, especially regarding the electoral campaign of the candidates, as well as the heated discussion around the alleged evidence of fraud during the aftermath of the election. Also, we focus on the political orientation of the users under analysis, and on their potential connection with conspiratorial and fringe theories.

To explore the general interests of MYT and NMYT mobilizers, we compare the hashtags of their tweets, as well as the descriptions of the YT videos they share on Twitter. To this aim, we apply SAGE [10] to find the most distinctive hashtags and keywords for the tweets and the video descriptions, respectively, shared by the two groups. It is worth to note that for MYT users we only took into account the non-moderated videos as the YouTube API cannot retrieve metadata related to moderated content. Table 1 reports the keywords and hashtags extracted by SAGE. We observe that

	Total Accounts	Total Videos	Verified Accounts	Bot Accounts	Suspended Accounts	InfoOps Accounts
NMYT Mobilizers	25396	88451	268	2234	7984	569
MYT Mobilizers	14481	61884	19	586	7793	30

Fig. 6. The number of accounts in each mobilizer group that were verified, bots or suspended. The columns are as follows: "Total Accounts" is the total number of accounts in each group. "Total Videos" is the number of unique YT videos shared by each group. "Verified Accounts" is the number of verified accounts in each group. "Bot Accounts" is the number of accounts labeled as a bot by the Botometer API in each group. "Suspended Accounts" is the number of accounts in each group that were later suspended by Twitter. "InfoOps Accounts" is the number of (suspended) accounts involved into information operation in each group

MYT users are Trump's supporters, as showed by the hashtags *#bestpresidentever45, #demonrats*, and they also sustain his allegation of voter fraud after the electoral count, as noticeable by the hashtags *#krakenteam, #trumpwon*. On the contrary, NMYT mobilizers explicitly despise Trump, as it can be noticed by the hashtags *#trumpvirus, #traitorinchief*, but their political orientation is not as clear as for MYT users. Indeed, *#gojoe* is the only pro-Biden hashtag in the NMYT's top-50 hashtags.

Overall, the keywords extracted from YT videos descriptions are aligned with the Twitter hashtags of the two groups. However, they do not communicate positive or negative sentiment but usually refer to (groups of) people who publicly stated their political preference. For instance, NMYT Mobilizers shared several videos related to the *lincoln* project and its ad starring *Barkhuff Dan* explicitly saying "I can see Trump for what he is — a coward. We need to send this draft-dodger back to his golf courses". It is worth to note that the Lincoln project is run by Republicans opposing Trump, which supports our intuition that NMYT users are not always Biden supporters. On the contrary, the MYT mobilizers support for Trump is clear also from their shared videos mentioning *Christina Bobb*, who has been close to the Trump's legal team that tried to overturn the result of the Presidential election, and *Tucker Carlson*, who has been recently nicknamed as Trump's heir[2].

To further investigate users' political orientation, we leverage the political leaning score assigned by MediaBias-FactCheck to several news outlets and, consistently with previous works [4, 13], we measure users' political orientation by averaging the scores of the domains they share on Twitter during the observation period. Fig. 7a and 7b show the top-10 domains shared by YT mobilizers and the political leaning distribution of MYT and NMYT users. The former group includes several far-right users (the median of the distribution is 0.47) who mostly share news from *breitbart.com* and *thegatewaypundit.com*, which are known to promote conspiracy theories and publish extreme conservative content. On the contrary, the latter group includes less extreme and more liberal users. However, the political leaning distribution of the NMYT group is bimodal (the larger mode is −0.27 and the smaller one is 0.41) and a small subset of these mobilizers shows an extreme, conservative ideology. This result is further confirmed by looking at the top-10 domains of MYT mobilizers, which include *foxnews.com*, which is also shared frequently by the NMYT group, and *forbes.com*, which is a center-right news outlet.

---

[2]https://www.theguardian.com/media/2020/jul/12/tucker-carlson-trump-fox-news-republicans

Table 1. Most shared hashtags and YT video keywords by NMYT and MYT mobilizers

	NMYT Mobilizers	**MYT Mobilizers**
Twitter hashtags	#putinspuppet #trumpvirus #resignnowtrump #trumplies #traitorinchief #gojoe #trumpkillus #weirdotrump	#krakenteam #chinebitchbiden #demonrats #evidenceoffraud #bestpresidentever45 #arrestfauci #trumpwon #trumppatriots
YouTube keywords	barkhuff dan bernie sander lincoln rainbow loyalty cnn incompetence	rsbn bobulinski censored christina bobb fitton spoiled tucker carlson

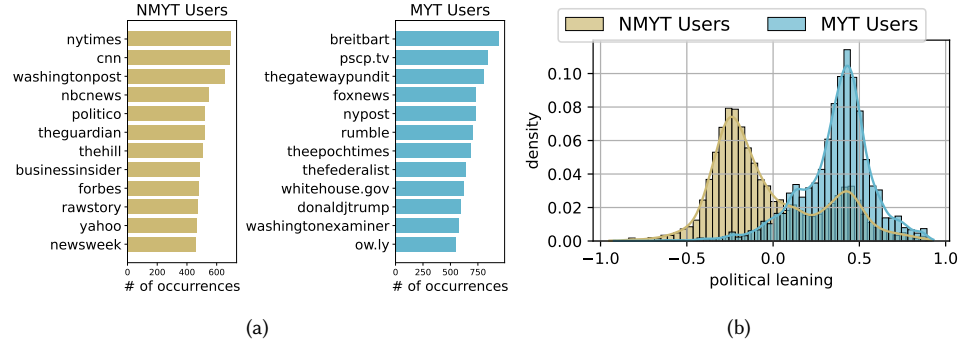(a)   (b)

Fig. 7. (a) The news outlet shared by each group of mobilizers (we omit the *.com* extension for brevity) ; (b) the distribution of the political leaning within the two groups of mobilizers

5 ENGAGEMENT TOWARDS MOBILIZERS OF MODERATED YOUTUBE VIDEOS (RQ3)

To respond to RQ3, we analyze the interaction patterns enacted by MYT and NMYT mobilizers by looking at the intra- and inter-group retweets exchanged by the users. In addition, we also consider the group of all *other* users (750k Twitter accounts) who shared at least one YT video and, by definition, do not fall in the category of YT mobilizers.

As the number of users in the three groups is different, we do not compare the absolute numbers of intra- and inter-group retweets. For this reason, we normalize the number of interactions by source (i.e., the total number of retweets that the group performs, see Fig. 8a) or by target (i.e., the total number of retweets that the group receives, see Fig. 8b). On the one hand, we observe that MYT and NMYT mobilizers generate the same relative amount of intra-group retweets, 13.2% and 11.7%, respectively, and inter-group retweets, 3.4% and 2.9%, respectively. On the other hand, Fig. 8b highlights that the two groups are engaged differently when considering the proportion of retweets they receive, i.e., 28.3% of the retweets received by MYT users are from accounts of the same group and only 2.1% are from the NMYT mobilizers. However, NMYT mobilizers tend to retweet both groups with the same frequency (11.9% NMYT to NMYT
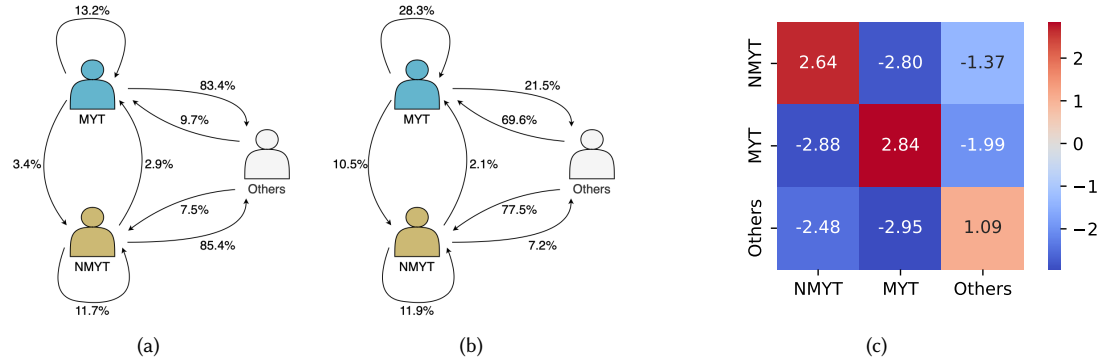
Fig. 8. Interaction patterns enacted by NMYT and MYT accounts: (a) Proportion of interactions between YouTube mobilizers normalized by the source; (b) Proportion of interactions between YouTube mobilizers normalized by the destination; (c) Z-scores of observed retweets between YouTube mobilizers (p-value < 0.01)

and 10.5% MYT to NMYT). This result indicates that MYT users tend to retweet and be retweeted more within their group than between groups. Finally, we can also notice that the users in the *Others* category retweets MYT and NMYT mobilizers almost in the same way. This observation emphasizes that the level of user activity is not informative to characterize interactions between NMYT users and the rest.

To further validate our results, we compare the observed number of retweets to a null model, which assumes that interactions occur by chance. Specifically, we randomize the users' assignment to the three groups (i.e., NMYT, MYT, and Others) and compute the mean and standard deviation of the interactions between the groups for 100 iterations. We then compute the z-scores to compare the observed retweets with the expected number of retweets from the null model. Fig. 8c indicates that the observed pattern of retweeting behavior among the mobilizers is consistent with the principle of homophily. Specifically, both groups of mobilizers were found to have a greater number of retweets within their respective groups and a lower number of retweets across groups than what would be expected by chance, i.e., $z > 2.5$ and $z < -1.5$, respectively.

6 DISCUSSIONS

6.1 Contributions

In this paper, we studied the Twitter discussion around the (video) content that is deemed harmful on YouTube. Leveraging an unprecedented large-scale dataset of 600M tweets shared by over 7.5M users, we discovered an unexpectedly high number of *moderated* YT videos shared on Twitter during the run-up to and aftermath of the 2020 U.S. election. Overall, moderated videos were shared more than non-moderated ones and received far more attention than content from *fringe* social media platforms. Moving beyond previous work, we investigated the characteristics of the Twitter users responsible for sharing both moderated and non-moderated YT videos. On the one hand, we found that users sharing moderated content tend to passively retweet what they see on Twitter rather than actively posting original tweets or replies. On the other hand, the mobilizers of non-moderated videos actively share YT videos in their original tweets. Overall, most of the users were regular Twitter accounts rather than bots or state-sponsored actors, and, even if we did not find any involvement in information operations, more than half of the mobilizers of moderated videos were suspended by Twitter. Furthermore, we found that the mobilizers of moderated YT videos are far-right supporters and

sustained Trump during the 2020 U.S. election. By contrast, the political preference of the mobilizers of non-moderated YT videos is more diverse since users in this group range from Biden supporters to other republican representatives who did not endorse Trump's political campaign. Finally, we studied the interactions between the mobilizers of moderated and non-moderated videos and discovered that both groups exhibit strong group cohesion and are engaged by the general Twitter audience in a similar manner.

6.2 Limitations

There is a number of limitations in our study. First, neither Twitter nor YouTube provides any supplementary information about account suspension and video moderation, as well as we do not know when these interventions occurred. As a result, there is no guarantee that YT videos were still online when re-shared through retweets on Twitter, but we can confidently assume they were not moderated yet when shared in an original tweet. In addition, we acknowledge our analyses, like several previous works [18, 19], might be biased towards moderated YT content which includes not only videos violating YouTube policies but also those removed by their publishers for any reason.

Second, we overlooked YT channels shared on Twitter to safeguard our analysis from Twitter users who just advertise their own (or others) YouTube channel [2]. However, this choice might prevent us to consider another potential source of harmful YT content on Twitter.

Third, the partition strategy to define the two groups of mobilizers is quite conservative as we considered Twitter users who *never share* moderated videos (NMYT mobilizers) and those who *mostly share* moderated videos (MYT mobilizers). The investigation of the users sharing a limited number of moderated videos is left for future research.

6.3 Conclusions and Future Works

Our study has two major takeaways: first, moderated YT videos are extensively shared on Twitter, and users who (passively) share those, endorse extreme and conspiratorial ideas. From a broader perspective, we have shown how harmful content originating in a *source* platform significantly pollute discussion on a *target* platform. Although more research is still needed, we conjecture that sharing information about the taken interventions would improve our understanding of cross-platform harmful content diffusion and benefit all entities within the information ecosystem. For instance, in the YouTube-Twitter cross-posting scenario considered in this paper, YouTube moderation activity can benefit both parties of the cooperation: on the one hand, Twitter has the opportunity to (early-)detect intra-platform harmful activities; on the other hand, YouTube can further improve its moderation based on the cross-platform signals tied with harmful YT content diffusion on Twitter.

Second, the mobilizers of moderated YT videos were rarely bot accounts or involved in information operations, but instead appeared to be regular Twitter users who do not necessarily share content from *fringe* platforms. This suggests that the cross-posting of (harmful) cross-platform content is participatory [20] and research in this field should not only target bots and trolls but, instead, consider the role of online crowds and more complex social structures across different social media platforms.

Future work might build upon our findings to design algorithms to automatically identify or predict whether a YT video will be moderated based on the engagement it receives on Twitter, as well as detect early-signals of radicalization. In addition, we aim to investigate whether our results generalize to other highly-moderated social media (e.g. Facebook, Instagram), including case studies considering content originating from low-moderation, fringe spaces.

REFERENCES

[1] Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 901–912. https://doi.org/10.1609/icwsm.v15i1.18113

[2] Adiya Abisheva, Venkata Rama Kiran Garimella, David Garcia, and Ingmar Weber. 2014. Who Watches (and Shares) What on Youtube? And When? Using Twitter to Understand Youtube Viewership. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14)*. Association for Computing Machinery, New York, NY, USA, 593–602. https://doi.org/10.1145/2556195.2566588

[3] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 187–195. https://doi.org/10.1145/3447535.3462637

[4] Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Who Falls for Online Political Manipulation?. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 162–168. https://doi.org/10.1145/3308560.3316494

[5] Cody Buntain, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. YouTube Recommendations and Effects on Sharing Across Online Social Platforms. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 11 (apr 2021), 26 pages. https://doi.org/10.1145/3449085

[6] Emily Chen, Ashok Deb, and Emilio Ferrara. 2020. #Election2020: The First Public Twitter Dataset on the 2020 US Presidential Election. *CoRR* abs/2010.00600 (2020). arXiv:2010.00600 https://arxiv.org/abs/2010.00600

[7] Matthew Childs, Cody Buntain, Milo Z. Trujillo, and Benjamin D. Horne. 2022. Characterizing YouTube and BitChute Content and Mobilizers During U.S. Election Fraud Discussions on Twitter. In *14th ACM Web Science Conference 2022* (Barcelona, Spain) *(WebSci '22)*. Association for Computing Machinery, New York, NY, USA, 250–259. https://doi.org/10.1145/3501247.3531571

[8] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Valensise, Emanuele Brugnoli, Ana Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 social media infodemic. *Scientific reports* 10 (10 2020). https://doi.org/10.1038/s41598-020-73510-5

[9] Evelyn Douek. 2020. The rise of content cartels. *Knight First Amendment Institute at Columbia* (2020).

[10] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse Additive Generative Models of Text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 1041–1048.

[11] Francesco Fabbri, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. 2022. Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 2719–2728. https://doi.org/10.1145/3485447.3512143

[12] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 US presidential election. *First Monday* (2020).

[13] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. 2020. Characterizing social media manipulation in the 2020 U.S. presidential election. *First Monday* 25, 11 (Oct. 2020). https://doi.org/10.5210/fm.v25i11.11431

[14] Tamar Ginossar, Iain J. Cruickshank, Elena Zheleva, Jason Sulskis, and Tanya Berger-Wolf. 2022. Cross-platform spread: vaccine-related content, sources, and conspiracy theories in YouTube videos shared in early Twitter COVID-19 conversations. *Human Vaccines & Immunotherapeutics* 18, 1 (2022), 1–13. https://doi.org/10.1080/21645515.2021.2003647 arXiv:https://doi.org/10.1080/21645515.2021.2003647 PMID: 35061560.

[15] Yevgeniy Golovchenko, Cody Buntain, Gregory Eady, Megan A. Brown, and Joshua A. Tucker. 2020. Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. *The International Journal of Press/Politics* 25, 3 (2020), 357–389. https://doi.org/10.1177/1940161220912682 arXiv:https://doi.org/10.1177/1940161220912682

[16] Michael Hameleers, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. 2020. A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication* 37, 2 (2020), 281–301. https://doi.org/10.1080/10584609.2019.1674979

[17] Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. 2023. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus* 2, 3 (03 2023). https://doi.org/10.1093/pnasnexus/pgad035 arXiv:https://academic.oup.com/pnasnexus/article-pdf/2/3/pgad035/49703177/pgad035.pdf pgad035.

[18] Maram Kurdi, Nuha Albadi, and Shivakant Mishra. 2020. "Video Unavailable": Analysis and Prediction of Deleted and Moderated YouTube Videos. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 166–173. https://doi.org/10.1109/ASONAM49781.2020.9381310

[19] Marcelo Sartori Locatelli, Josemar Caetano, Wagner Meira Jr., and Virgilio Almeida. 2022. Characterizing Vaccination Movements on YouTube in the United States and Brazil. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media* (Barcelona, Spain) *(HT '22)*. Association for Computing Machinery, New York, NY, USA, 80–90. https://doi.org/10.1145/3511095.3531283

[20] Luca Luceri, Stefano Cresci, and Silvia Giordano. 2021. Social Media against Society. *The Internet and the 2020 Campaign* (2021).

[21] Luca Luceri, Silvia Giordano, and Emilio Ferrara. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 417–427.

[22] Tamar Mitts, Nilima Pisharody, and Jacob Shapiro. 2022. Removal of Anti-Vaccine Content Impacts Social Media Discourse. In *14th ACM Web Science Conference 2022* (Barcelona, Spain) *(WebSci '22)*. Association for Computing Machinery, New York, NY, USA, 319–326. https://doi.org/10.1145/3501247.3531548

[23] Gianluca Nogara, Padinjaredath Suresh Vishnuprasad, Felipe Cardoso, Omran Ayoub, Silvia Giordano, and Luca Luceri. 2022. The Disinformation Dozen: An Exploratory Analysis of Covid-19 Disinformation Proliferation on Twitter. In *14th ACM Web Science Conference 2022* (Barcelona, Spain) *(WebSci '22).* Association for Computing Machinery, New York, NY, USA, 348–358. https://doi.org/10.1145/3501247.3531573

[24] Francesco Pierri, Matthew R. DeVerna, Kai-Cheng Yang, David Axelrod, John Bryden, and Filippo Menczer. 2022. One year of COVID-19 vaccine misinformation on Twitter. https://doi.org/10.48550/ARXIV.2209.01675

[25] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr. 2019. Auditing Radicalization Pathways on YouTube. *CoRR* abs/1908.08313 (2019). arXiv:1908.08313 http://arxiv.org/abs/1908.08313

[26] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2022. Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning. https://doi.org/10.48550/ARXIV.2209.09803

[27] Zeve Sanderson, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review* (2021).

[28] Digital Forensic Research Lab Stanford Internet Observatory, Center for an Informed Public. 2021. The Long Fuse: Misinformation and the 2020 Election. Stanford Digital Repository: Election Integrity.

[29] Vishnuprasad Padinjaredath Suresh, Gianluca Nogara, Felipe Cardoso, Stefano Cresci, Silvia Giordano, and Luca Luceri. 2023. Tracking Fringe and Coordinated Activity on Twitter Leading Up To the US Capitol Attack. *arXiv preprint arXiv:2302.04450* (2023).

[30] Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. 2021. A Multi-Platform Analysis of Political News Discussion and Sharing on Web Communities. In *2021 IEEE International Conference on Big Data (Big Data).* 1481–1492. https://doi.org/10.1109/BigData52589.2021.9671843

[31] Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).

[32] Yiping Xia, Josephine Lukito, Yini Zhang, Chris Wells, Sang Jung Kim, and Chau Tong. 2019. Disinformation, performed: self-presentation of a Russian IRA account on Twitter. *Information, Communication & Society* 22, 11 (2019), 1646–1664. https://doi.org/10.1080/1369118X.2019.1621921 arXiv:https://doi.org/10.1080/1369118X.2019.1621921

[33] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. 2022. Botometer 101: Social bot practicum for computational social scientists. *CoRR* abs/2201.01608 (2022). arXiv:2201.01608 https://arxiv.org/abs/2201.01608