RESEARCH-ARTICLE

Auditing radicalization pathways on YouTube

MANOEL HORTA RIBEIRO, EPFL, Lausanne, Switzerland

RAPHAEL OTTONI, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

ROBERT WEST, EPFL, Lausanne, Switzerland

VÍRGILIO AUGUSTO FERNANDES ALMEIDA, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

WAGNER MEIRA, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Open Access Support provided by:

EPFL

Federal University of Minas Gerais

Published: 27 January 2020

Citation in BibTeX format

FAT* '20: Conference on Fairness, Accountability, and Transparency
January 27 - 30, 2020
Barcelona, Spain

Conference Sponsors:
ACM

https://doi.org/10.1145/3351095.3372879
ISBN: 9781450369367

Auditing Radicalization Pathways on YouTube

Manoel Horta Ribeiro*
EPFL
manoel.hortaribeiro@epfl.ch

Raphael Ottoni
UFMG
rapha@dcc.ufmg.br

Robert West
EPFL
robert.west@epfl.ch

Virgílio A. F. Almeida
UFMG, Berkman Klein Center
virgilio@dcc.ufmg.br

Wagner Meira Jr.
UFMG
meira@dcc.ufmg.br

ABSTRACT

Non-profits, as well as the media, have hypothesized the existence of a radicalization pipeline on YouTube, claiming that users systematically progress towards more extreme content on the platform. Yet, there is to date no substantial quantitative evidence of this alleged pipeline. To close this gap, we conduct a large-scale audit of user radicalization on YouTube. We analyze 330,925 videos posted on 349 channels, which we broadly classified into four types: Media, the Alt-lite, the Intellectual Dark Web (I.D.W.), and the Alt-right. According to the aforementioned radicalization hypothesis, channels in the I.D.W. and the Alt-lite serve as gateways to fringe far-right ideology, here represented by Alt-right channels. Processing 72M+ comments, we show that the three channel types indeed increasingly share the same user base; that users consistently migrate from milder to more extreme content; and that a large percentage of users who consume Alt-right content now consumed Alt-lite and I.D.W. content in the past. We also probe YouTube's recommendation algorithm, looking at more than 2M video and channel recommendations between May/July 2019. We find that Alt-lite content is easily reachable from I.D.W. channels, while Alt-right videos are reachable only through channel recommendations. Overall, we paint a comprehensive picture of user radicalization on YouTube.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

Radicalization, hate speech, extremism, algorithmic auditing

ACM Reference Format:
Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira Jr.. 2020. Auditing Radicalization Pathways on YouTube. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3351095.3372879

*Work done mostly while at UFMG.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372879

1 INTRODUCTION

Video channels that discuss social, political and cultural subjects have flourished on YouTube. Frequently, the videos posted in such channels focus on highly controversial topics such as race, gender, and religion. The users who create and post such videos span a wide spectrum of political orientation, from prolific podcast hosts like Joe Rogan to outspoken advocates of white supremacy like Richard Spencer. These individuals not only share the same platform but often publicly engage in debates and conversations with each other on the website [24]. This way, even distant personalities can be linked in chains of pairwise co-appearances. For instance, Joe Rogan interviewed YouTuber Carl Benjamin [35], who debated with white supremacist Richard Spencer [6].

According to Lewis [24], this proximity may create "radicalization pathways" for audience members and content creators. Examples of these journeys are plenty, including content creator Roosh V's trajectory from pick-up artist to Alt-right supporter [23, 37] and Caleb Cain's testimony of his YouTube-driven radicalization [36].

The claim that there is a "radicalization pipeline" on YouTube should be considered in the context of decreasing trust in mainstream media and increasing influence of social networks. Across the globe, individuals are skeptical of traditional media vehicles and growingly consume news and opinion content on social media [21, 31]. In this setting, recent research has shown that fringe websites (*e.g.*, *4chan*) and subreddits (*e.g.*, */r/TheDonald*) have great influence over which memes [43] and news [44] are shared in large social networks, such as Twitter. YouTube is extremely popular, especially among children and teenagers [5], and if the streaming website is actually radicalizing individuals this could push fringe ideologies like white supremacy further into the mainstream [41].

A key issue in dealing with topics like radicalization and hate speech is the lack of agreement over what is "hateful" or "extreme" [38]. A workaround is to perform analyses based on *communities*, large sets of loosely associated content creators (here represented by their YouTube channels). For the purpose of this work, we consider three "communities" that have been associated with user radicalization [24, 36, 42] and that differ in the extremity of their content: the "Intellectual Dark Web" (I.D.W.), the "Alt-lite" and the "Alt-right". While users in the I.D.W. discuss controversial subjects like race and I.Q. [42] without necessarily endorsing extreme views, members of the Alt-right sponsor fringe ideas like that of a white ethnostate [18]. Somewhere in the middle, individuals of the Alt-lite deny to embrace white supremacist ideology, although they frequently flirt with concepts associated with it (*e.g.*, the "Great Replacement", globalist conspiracies).

Present work. In this paper, we audit whether users are indeed becoming radicalized on YouTube and whether the recommendation algorithms contribute towards this radicalization. We do so by examining three prominent communities: the Intellectual Dark Web, the Alt-lite and the Alt-right. More specifically, considering Alt-right content as a proxy for extreme content, we ask:

RQ1 How have these channels grown on YouTube in the last decade?
RQ2 To which extent do users systematically gravitate towards more extreme content?
RQ3 Do algorithmic recommendations steer users towards more extreme content?

We develop a data collection process where we *(i)* acquire a large pool of relevant channels from these communities; *(ii)* collect metadata and comments for each of the videos in the channels; *(iii)* annotate channels as belonging to several different communities; and *(iv)* collect YouTube video and channel recommendations. We also collect traditional and alternative media channels for additional comparisons. We use these as a sanity check to capture the growth of other content on YouTube, rather than trying to obtain similar users in other channels. These efforts resulted in a dataset with more than 72M comments in 330,925 videos of 349 channels and with more than 2M video and 10K channel recommendations. Importantly, our recommendations do not account for personalization. We analyze this large dataset extensively:

- We look at the growth of the I.D.W., the Alt-lite and the Alt-right throughout the last decade in terms of videos, likes and views, finding a steep rise in activity and engagement in the communities of interest when compared with the media channels. Moreover, comments per view seem to be particularly high in more extreme content, reaching near to 1 comment for every 5 views in Alt-right channels in 2018 (Sec. 4).
- We inspect the intersection of commenting users across the communities, finding they increasingly share the same user base. Analyzing the overlap between the sets of commenting users, we find that approximately half of the users who commented on Alt-right channels in 2018 also comment on Alt-lite and on I.D.W. channels (Sec. 5).
- We also find that the intersection is not only growing due to new users but that there is significant user migration among the communities being studied. Users that initially comment only on content from the I.D.W. or the Alt-lite throughout the years consistently start to comment on Alt-right content. These users are a significant fraction of the Alt-right commenting user base. This effect is much stronger than for the large traditional and alternative media channels we collected (Sec. 6).
- Lastly, we take a look at the impact of YouTube's recommendation algorithms, running simulations on recommendation graphs. Our analyses show that, particularly through the channel recommender system, Alt-lite channels are easily discovered from I.D.W. channels, and that Alt-right channels may be reached from the two other communities (Sec. 7).

This is, to our best knowledge, the first large scale quantitative audit of user radicalization on YouTube. We find strong evidence for radicalization among YouTube users, and that YouTube's recommender system enables Alt-right channels to be discovered, even in a scenario without personalization. We discuss our findings and our limitations further in Sec. 8. We argue that commenting users are a good enough proxy to measure the user radicalization, as more extreme content seems to beget more comments. Moreover, regardless of the degree of influence of the recommender system in the process of radicalizing users, there is significant evidence that users are reaching content sponsoring fringe ideologies from the Alt-lite and the Intellectual Dark Web.

2 BACKGROUND

Contrarian communities. We discuss three of YouTube's prominent communities: the Intellectual Dark Web, the Alt-lite and the Alt-right. We argue that all of them are *contrarians*, in the sense that they often oppose mainstream views or attitudes. According to Nagle, these communities flourished in the wave of "anti-PC" culture of the 2010s, where social-political movements (e.g. the transgender rights movement, the anti-sexual assault movement) were portrayed as hysterical, and their claims, as absurd [30].

According to the Anti Defamation League [3], the Alt-Right is a loose segment of the white supremacist movement consisting of individuals who reject mainstream conservatism in favor of politics that embrace racist, anti-Semitic and white supremacist ideology. The Alt-right skews younger than other far-right groups, and has a big online presence, particularly on fringe web sites like *4chan*, *8chan* and certain corners of *Reddit* [2].

The term Alt-lite was created to differentiate right-wing activists who deny embracing white supremacist ideology. Atkison argues that the Unite the Rally in Charlottesville was deeply related to this change, as participants of the rally revealed the movement's white supremacist leanings and affiliations [8]. Alt-right writer and white supremacist Greg Johnson [3] describes the difference between Alt-right and Alt-lite by the origin of its nationalism: *"The Alt-lite is defined by civic nationalism as opposed to racial nationalism, which is a defining characteristic of the Alt-right"*. This distinction was also highlighted in [28]. Yet it is important to point out that the line between the Alt-right and the Alt-lite is blurry [3], as many Alt-liters are accused of dog-whistling: attenuating their real beliefs to appeal to a more general public and to prevent getting banned [22, 25]. To address this problem, in this paper we take a conservative approach to our labeling, naming only the most extreme content creators as Alt-right.

The "Intellectual Dark Web" (I.D.W.) is a term coined by Eric Weinstein to refer to a group of academics and podcast hosts [42]. The neologism was popularized in a New York Times opinion article [42], where it is used to describe "iconoclastic thinkers, academic renegades and media personalities who are having a rolling conversation about all sorts of subjects, [...] touching on controversial issues such as abortion, biological differences between men and women, identity politics, religion, immigration, etc."

The group described in the NYT piece includes, among others, Sam Harris, Jordan Peterson, Ben Shapiro, Dave Rubin, and Joe

Rogan, and also mentions a website with an unofficial list of members [7]. Members of the so-called I.D.W. have been accused of espousing politically incorrect ideas [9, 15, 26]. Moreover, a recent report by the Data & Society Research Institute has claimed these channels are "pathways to radicalization" [24], acting as entry points to more radical channels, such as those in Alt-right. Broadly, members of this loosely defined movement see these criticisms as a consequence of discussing controversial subjects [42], and some have explicitly dismissed the report [40]. Similarly to what happens between Alt-right and Alt-lite, there are also blurry lines between the I.D.W. and the Alt-lite, especially for non-core members, such as those listed on the aforementioned website [7]. To break ties, we label borderline cases as Alt-lite.

Radicalization. We consider the definition given by McCauley and Moskalenko [29]: ("Functionally, political radicalization is increased preparation for and commitment to intergroup conflict. Descriptively, radicalization means change in beliefs, feelings, and behaviors in directions that increasingly justify intergroup violence and demand sacrifice in defense of the ingroup.") and use increased consumption of Alt-right content as a proxy for radicalization. This is reasonable since the Alt-right's rhetoric has been invoked by the perpetrators of some recent terrorist attacks (*e.g.* the Christchurch mosque shooting [27]), and since it champions ideas promoting intergroup conflict (*e.g.* a white ethnostate [18]). Our conservative strategy when labeling channels is of particular importance here: Alt-right channels are closely related to these ideas, while the Alt-lite/I.D.W. are given the benefit of doubt.

Auditing the web. As algorithms play an ever-larger role in our lives, it is increasingly important for researchers and society at large to reverse engineer algorithms' input-output relationships [13]. Previous large scale algorithmic auditing include measuring discrimination on AirBnB [14], personalization on web search [19] and price discrimination on e-commerce web sites [20]. We argue this work is an audit in the sense that it measures a troublesome phenomenon (user radicalization) in a content-sharing social environment heavily influenced by algorithms (YouTube). Unfortunately, it is not possible to obtain the entire history of YouTube recommendation, so we must limit the algorithmic analyses to a time slice of a constantly changing black-box. Although comments may give us insight into the past, it is challenging to tease apart the influence of the algorithm in previous times. Another limitation of our auditing is that we do not account for user personalization. Despite these flaws, we argue that: *(i)* our analyses provide answers to important questions related with impactful societal processes that are allegedly happening in YouTube (regardless of the impact of the recommender system), and *(ii)* our framework for auditing user radicalization can be replicated through time and expanded to handle personalization.

Previous research from/on YouTube. Previous work by Google sheds light into some of the high-level technicalities of YouTube's recommender system [11, 12]. Their latest paper indicates they use embeddings for video searches and video histories as inputs for a dense neural network [12]. There also exists a large body of work studying violent [16], hateful or extremist [4, 39] and disturbing content [34] on the platform. Much of the existing work focuses on creating detection algorithms for these types of content using features of the comments, the commenting users and the videos [4, 16]. Sureka et al. [39] use a seed-expanding methodology to track extremist user communities, which yielded high precision in including relevant users. This is somewhat analogous to what we do, although we use YouTube's recommender system while they use user friends, subscriptions and favorites. Ottoni et al. perform an in-depth textual analysis of 23 channels (13 broadly defined as Alt-right), finding significantly different topics across the two groups [32]. O'Callegan et al. [33] simulate a recommender system with channels tweeted in an extreme right dataset. They show that a simple non-negative matrix factorization metadata-based recommender system would cluster extreme right topics together.

3 DATA COLLECTION

We are interested in three communities on YouTube: the I.D.W., the Alt-lite, and the Alt-right. Identifying such communities and the channels which belong to them is no easy task: the membership of channels to these communities is volatile and fuzzy, and there is disagreement between how members of these communities view themselves, and how they are considered by scholars and the media. These particularities make our challenge multi-faceted: on one hand, we want to study user radicalization, and determine, for example, if users who start watching videos by communities like the I.D.W. eventually go on to consume Alt-right content. On the other, there is often no clear agreement on who belongs to which community.

Due to these nuances, we devise a careful methodology to **(a)** collect a large pool of relevant channels; **(b)** collect data and the recommendations given by YouTube for these channels; **(c)** manually labeling these channels according to the communities of interest.

(a) For each community, we create a pool of channels as follows. We refer to channels obtained in the i-th step as *Type i* channels.

(1) We choose a set of *seed channels*. Seeds were extracted from the I.D.W. unofficial website [7], Anti Defamation League's report on the Alt-lite/the Alt-right [3] and Data & Society's report on YouTube Radicalization [24]. We pick popular channels that are representative of the community we are interested in. Each seed was independently annotated two times and discarded in case there was any disagreement. We further detail the annotation process later in this section.
(2) We choose a set of *keywords* related to the sub-communities. For each keyword, we use YouTube's search functionality and consider the first 200 results in English. We then add channels that broadly relate in topic to the community in question. For example, for the Alt-right, keywords included both terms associated with their narratives, such as *The Jewish Question* and *White Genocide,* as well as the names or nicknames of famous Alt-righters, such as *weev* and *Christopher Cantwell.*
(3) We iteratively search the related and featured channels collected in steps (1) and (2), adding relevant channels (as defined in 2). Note that these are two ways channel can link to each other. Featured channels may be chosen by YouTube content creators: if your friend has a channel and you want to support it, you can put it on your "Featured Channels" tab. Related channels are created by YouTube's recommender system.
(4) We repeat step (3), iteratively collecting another hop of featured/recommended channels from those obtained in (3).

Table 1: Top 16 YouTube channels with the most views per each community and for media channels.

	Alt-right	Views	Alt-lite	View	Intellectual Dark Web	Views	Media	Views
1	James Allsup	62M	StevenCrowder	727M	PowerfulJRE	1B	vox	1B
2	Black Pigeon Speaks	50M	Rebel Media	405M	JRE Clips	717M	gq magazine	1B
3	ThuleanPerspective	45M	Paul Joseph Watson	356M	PragerUniversity	635M	vice news	1B
4	Red Ice TV	42M	MarkDice	334M	The Daily Wire	247M	wired magazine	1B
5	The Golden One	12M	SargonofAkkad100	258M	The Rubin Report	206M	vanity fair	639M
6	AmRenVideos	9M	Stefan Molyneux	193M	ReasonTV	138M	the verge	636M
7	NeatoBurrito Productions	7M	hOrnsticles3	145M	JordanPetersonVideos	90M	glamour magazine	620M
8	The Last Stand	7M	MILO	133M	Bite-sized Philosophy	62M	business insider	523M
9	MillennialWoes	6M	Styxhexenhammer666	132M	Owen Benjamin	35M	huffington post	329M
10	Mark Collett	6M	OneTruth4Life	112M	AgatanFoundation	33M	today i found out	328M
11	AustralianRealist	5M	No Bullshit	104M	Essential Truth	32M	cbc news	324M
12	Jean-François Gariépy	5M	SJWCentral	90M	Ben Shapiro	30M	the guardian	300M
13	Prince of Zimbabwe	5M	Computing Forever	87M	YAFTV	30M	people magazine	287M
14	The Alternative Hypothesis	5M	The Thinkery	86M	joerogandotnet	25M	big think	258M
15	Matthew North	4M	Bearing	81M	TheArchangel911	24M	cosmopolitan	256M
16	Faith J Goldy	4M	RobinHoodUKIP	64M	Clash of Ideas	24M	global news	252M

The annotation process done here followed the same instructions as the one explained in detail for data collection step **(c)**. Steps (2)—(4), were done by a co-author with more than 50 hours of watch-time of the communities of interest. Notice that, in steps (2)—(4), we are not labeling the channels, but creating a pool of channels to be further inspected and labeled in subsequent steps. The complete list of seeds obtained from (1) and of keywords used in (2) may be found in Appendix A. A clear distinction between featured and recommended channels may be found in Appendix B.

(b) For each channel, we collect the number of subscribers and views, and for their videos, all the comments and captions. Video and channel recommendations were collected separately using custom-made crawlers. We collected multiple "rounds" of recommendations, 22 for channel recommendations and 19 for video recommendations. Each "round" consists of collecting all recommended channels (on the channel web page) and all recommended videos (on the video web page). To circumvent possible location bias in the data we collected we used VPNs from 7 different locations: 3 in the USA, 2 in Canada, 1 in Switzerland and 1 in Brazil. Moreover, channels were always visited in random order, to prevent any biases from arising from session-based recommendations. As we extensively discuss throughout the paper, this does not include personalization, as we do not log in into any account.

(c) Channel labeling was done in multiple steps. All channels are either seeds (*Type 1*) or obtained through YouTube's recommendation/search engine (*Types 2 and 3*). Notice that *Type 1* channels were assigned labels at the time of their collection. For the others, we had 2 of the authors annotate them carefully. They both had significant experience with the communities being studied, and were given the following instructions:

> Carefully inspect each one of the channels in this table, taking a look at the most popular videos, and watching, altogether, at least 5 minutes of content from that channel. Then you should decide if the channel belongs to the Alt-right, the Alt-lite, the Intellectual Dark Web (I.D.W.), or whether you think it doesn't fit any of the communities. To get a grasp on who belongs to the I.D.W., read [42], and check out the website with some of the alleged members of the group [7]. Yet, we ask you to consider the label holistically, including channels that have content from these creators and with a similar spirit to also belong in this category. To distinguish between the Alt-right and the Alt-lite, read [3] and [28]. It is important to stress the difference between civic nationalism and racial nationalism in that case. Please consider the Alt-right label only to the most extreme content. You are encouraged to search on the internet for the name of the content creator to help you make your decision.

The annotation process lasted for 3 weeks. In case they disagreed, they had to discuss the cases individually until a conclusion was reached. Interanotator agreement was of 75.57% (95% CI [67.5, 82.5]). We ended up with 85 I.D.W., 112 Alt-lite and 84 Alt-right channels.

Media. We also collect popular media channels. These were obtained from the *mediabiasfactcheck.com* [1]. For each media source of the categories on the website (*Left, Left-Center, Center, Right-Center, Right*) we search for its name on YouTube and consider it if there is a match in the first page of results [1]. Some of the channels were not considered because they had too many videos (15, 000+) and we were not able to retrieve them all (which is important, because our analyses are temporal). In total, we collect 68 channels that way. We use these media channels as a sanity check to capture general trends among more mainstream YouTube channels.

We summarize the dataset collected in the Tab. 2. Data collection was performed during the 19-30th of May 2019, and the collection of the recommendations between May-July 2019.

Table 2: Overview of our dataset.

Channels	349	Video Recs rounds	19
Videos	330,925	Video Recs	2,474,044
Comments	72,069,878	Channel Recs Rounds	22
Commenting users	5,980,709	Channel Recs	14,283

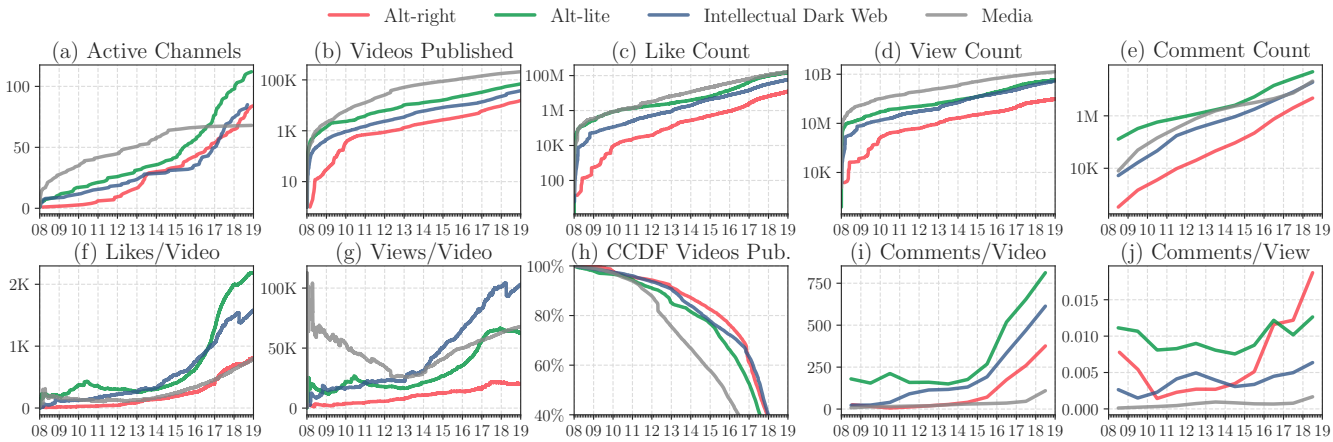

Figure 1: On the top row figures (a)—(e), for each community and media channels, we have the cumulative number of active channels (that posted at least one video), of videos published, of likes, views and of comments. In the bottom row, we have engagement metrics (accumulated over time), (figures (f), (g), (i) and (j)) and the CCDF of videos published, zoomed in the range $[40\%, 100\%]$ on the y-axis (figure (h)). Notice that for comments, we know only the year when they were published, and thus the CDFs granularity is coarser (years rather than seconds). The raw numbers of views, likes, videos published and more are shown in Appendix C

4 THE RISE OF CONTRARIANS

We present an overview of the channels in the communities of interest, and show results about their growth in the last years, setting the stage to more in-depth analyses in later sections. Tab. 1 shows the 16 most viewed YouTubers for each of the communities and for the media channels, and Figure 1 shows information on the number of videos published, channels created, likes, views, and comments per year, as well as several engagement metrics.

Recent rise in activity. Figs. 1(a)—(e) show the rise in channel creation, video publishing, likes, views, and comments in the last decade. The four latter are growing exponentially for all the communities of interest and for the media channels. Noticeably, the rise in the number of active channels is much more recent for the communities of interest than for media channels, as shown in Fig. 1(a). In mid 2015, for example, 66 out of the 68 of the media channels were active (posted their first video), while less than 50% of the Alt-lite, Alt-right and I.D.W. channels had done so. This growth in the communities of interest during 2015 may also be noted in Fig. 1(i), which shows the CDF of number comments per videos, and can also be seen between early 2014 and late 2016 in Figs. 1(f)—(g), which show the number of likes and views per video, respectively. Notice that the number of likes and views is obtained during data collection, and thus, it might be that older videos from those channels became popular later. Altogether, our data corroborates with the narrative that these communities gained traction in (and fortified) Donald Trump's campaign during the 2016 presidential elections [10, 17].

Engagement. A key difference between the communities of interest and the media channels is the level of engagement with the videos, as portrayed by the number of likes per video, comments per video and comments per view, shown in Figs. 1 (f), (i), and (j), respectively. For all these metrics, the communities of interest have more engagement than the media channels: Although media channels have more views per video, as shown in Figs. 1(g), these views are less often converted into likes and comments. Notably, Alt-right channels have, since 2017, become the ones with the highest number of comments per view, with nearly 1 comment per 5 views by 2018.

Dormant Alt-right Channels. Although by 2013, approximately the same number of channels of all three communities had become active ($\sim$ 30), as it can be seen in Fig. 1(a), the number of videos they published by the Alt-right was low before 2016. This can be seen in the CCDF in Fig. 1(h): while media and Alt-lite channels had published nearly 40% of their content, the Alt-right had published a bit more than 20%. This is not because the most popular channels did not yet exist: 4 out of the 5 current top Alt-right channels (accumulating approximately 150M views) had already been created by 2013. Moreover, it is noteworthy that many of the channels now dedicated to Alt-right content have initial videos related to other subjects. Take for example the channel "The Golden One", number 5 in Tab. 1. Most of the initial videos in the channel are about working out or video-games, with politics related videos becoming increasingly occurring. The growth in engagement metrics such as likes per video and comments per video of the Alt-right succeeds that of the I.D.W. and of the Alt-lite, resonating with the narrative that the rise of Alt-Lite and I.D.W. channels created fertile grounds for individuals with fringe ideas to prosper [24, 30].

Although our data-driven analysis sheds light on existing narratives on the communities of interest, it is still impossible to determine, from these simple CDFs, whether there is a radicalization pipeline. To do so, in the following two sections, we dig deeper into the relationship between these communities looking closely at the users who commented on them.

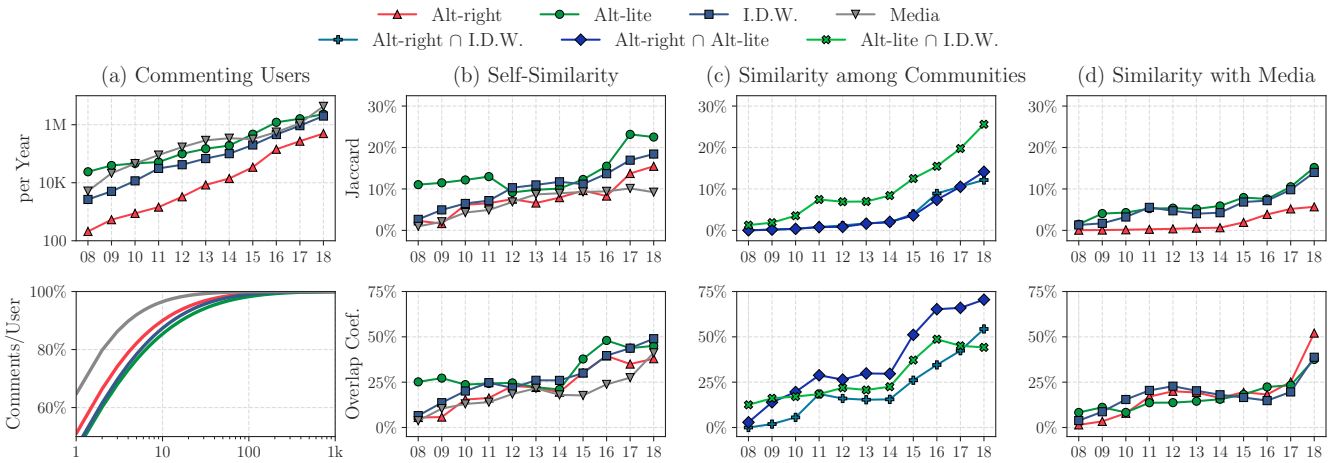

Figure 2: In *(a)*, the number of unique commenting users per year in the top plot and the CDF of comments per user for each one of the communities in the bottom plot. In *(b)*—*(d)* we show two similarity metrics (Jaccard and Overlap Coefficient) for different pairs of sets of commenting users across the years. In *(b)* these pairs are the sets of users of each community in subsequent years. In *(c)* these pairs are the sets of users of each one of the communities of interest. In *(d)* these pairs are the sets of users of the communities compared with the users who commented in media channels.

5 USER INTERSECTION

We begin our in-depth analysis of users who commented on the channels of interest by analysing the *intersection* between the users in different channels and communities. In that context, we use two set similarity metrics: the Jaccard Similarity $\frac{|A \cap B|}{|A \cup B|}$; and the Overlap Coefficient $\frac{|A \cap B|}{\min(|A|, |B|)}$. Notice that the overlap coefficient is particularly useful to compare communities of different sizes. For example, a small subset of a large set may yield low Jaccard Similarity, but will necessarily yield an Overlap Coefficient of 1.

Column (a) of Fig. 2 characterizes commenting users. The top plot shows the absolute number of commenting users per year, while the bottom one shows the CDF of the number of comments per user per community. It is interesting to compare these plots with that of Fig. 1(e), as we can see that the communities of interest have many more highly active commenters. This supports the hypothesis that users who consume content in the communities of interest are more "engaged" than those who consume the content from the media channels. Notice also that, although the Alt-right commenters have, on average, fewer comments than those in Alt-lite or the I.D.W., the community is much younger (as discussed in Sec. 4), and thus it is hard to tell whether their users are less engaged.

In columns (b)—(d) of Fig. 2 we consider the intersection between the commenting users of the I.D.W., the Alt-lite, the Alt-right and media channels. The top figure for each column shows the Jaccard Similarity and the bottom one shows the Overlap Coefficient.

Column (b) in Fig. 2 shows the similarity measures for a community with itself a year before (which here we name self-similarity). We find that the retention of users among the three communities is growing with time for both metrics. However, for media channels, we find that the Jaccard similarity is plateauing since 2014 and that the overlap coefficient only recently started to grow, perhaps due to the sharp increase in commenting users since 2015. Commenting users from the communities of interest seem to go back more often than those in media channels.

Column (c) in Fig. 2 shows the pairwise similarity between the three communities. Notably, in 2018, the Jaccard Similarity between the Alt-lite and the I.D.W. reached almost 30%, which is more than the self-similarity between the two communities. Moreover, the Overlap Coefficient of the Alt-right with the Alt-lite and the I.D.W is high: reaching around 50% in 2018. This means around half of the users who commented in Alt-right channels commented in the other communities.

Lastly, column (d) in Fig. 2 shows the similarity of the three communities with the media channels. We have that the Jaccard similarity between the I.D.W. and the Alt-lite and the media channels is not so different from the similarity between these communities and the Alt-right. This is a subtle finding. On one hand, it means that individuals in these communities make up a significant portion of the massive media channels we collected, which gather billions of views. These communities do not exist in a vacuum but are part of the existing online information environment. On the other, it shows that the Alt-right, a group of channels with order of magnitudes fewer views, subscribers and comments, are actually *on par* with these large channels. Inspecting the Overlap Coefficient, however, we get a different view: there we have that the communities overlap more with themselves than with the media channels, particularly since 2015. However, in 2018, there is a sharp growth in the similarity with media channels. A hypothesis for this is that, as these channels grew more popular (as previously discussed in Sec. 4, they became more mainstream).

These analyses take us one step further in understanding the communities being studied. We again see that their users are more engaged, and, notably, find that the I.D.W, the Alt-lite, and the Alt-right increasingly share the same commenting user base.

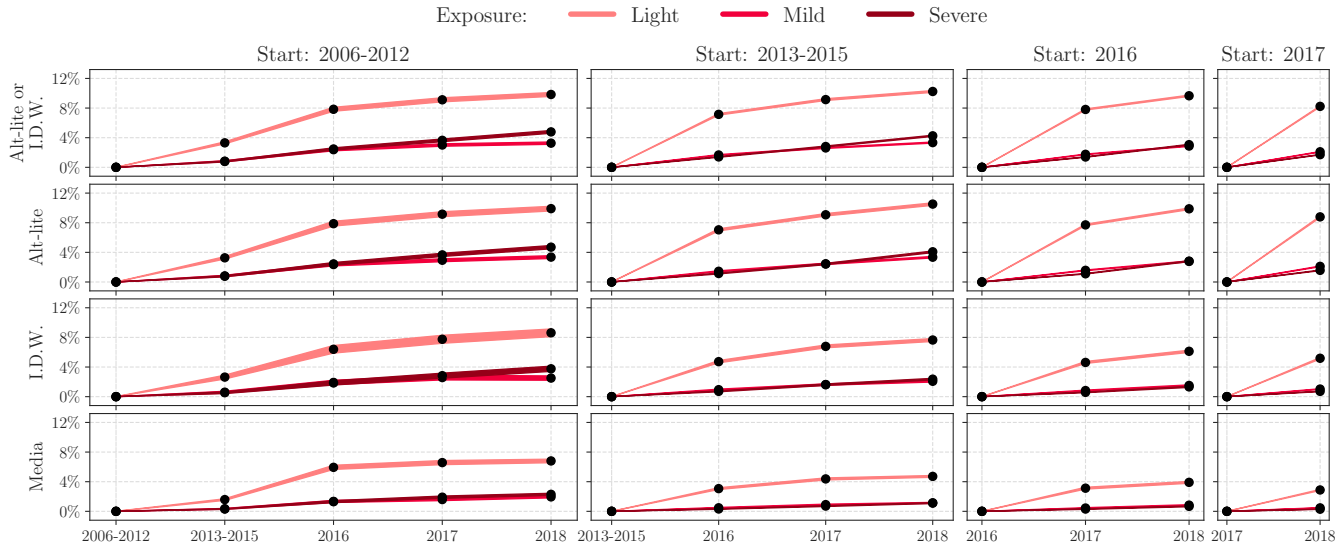

Figure 3: We show how users migrate towards Alt-right content. For users who consumed only videos in the communities indicated by the labels in the rows (Alt-lite or I.D.W., Alt-lite, I.D.W., and Media), we show the chance that they go on to consume Alt-right content. We consider three levels of exposure: light (commented in 1 to 2 Alt-right videos), mild (3 to 5) and severe (6+). Each column tracks users on a different starting date. Initially, their exposure rates are 0 (as they did not consume any Alt-right content). As time passes, we show the exposure rates in the y-axis, for each of the years, in the x-axis. Line widths represent 95% confidence intervals.

6 USER MIGRATION

In the previous section, we showed that the commenting user bases among the I.D.W., the Alt-lite, and the Alt-right are increasingly similar —and the effect is stronger than for media channels. This indicates that there is a growing percentage of users consuming extreme (Alt-right) content on YouTube *while also* consuming content from other milder communities (Alt-lite/I.D.W.). Yet, it does not, *per se*, indicate that there is a radicalization pipeline on the website. It could be, for example, that new users who join the website go on to consume content from all three communities. To better address this question, we find users who *did not comment* in Alt-right content in a given year and track their subsequent activity. Notice that we do not have the user's entire activity history, and thus, we track their activity only in the channels whose videos we collected.

For four time brackets [(2006–2012), (2013–2015), (2016), (2017)] we track four sets of users: those who only commented on videos of the Alt-lite or the I.D.W, those who did so only for videos on the Alt-lite, those who did so only for videos on the I.D.W., and those who commented only on videos of the media channels. Then, for subsequent years, we track the same users. Notice that when users are tracked for one year they are not eligible for selection in upcoming years. We consider these users to be exposed if they commented on 1-2 (light), 3-5 (mild) or 6+ (severe) Alt-right videos.

The results for this analysis are shown in Fig. 3. We show the percentage of users we managed to track that were exposed. The number of users tracked and exposed at each step may be found in Appendix C. Consider, for example, users who on 2006 – 2012 commented only on I.D.W. or Alt-lite content (227,945 users), as shown in the subplot in the first column and the first row. By 2018, around 10% were lightly exposed, and roughly 4% severely or mildly so — which amounts to approximately 9K users in total. From the ones who in 2017 commented only on Alt-lite or I.D.W. videos (1,251,674 users), as shown in the last column of the first row, approximately 12% of them were exposed — more than 60K users altogether.

We also find that media channels present lower exposure rates, as can be seen in the last row of the figure. The difference is particularly large for the last three time brackets. Less than 1% of users in media channels were mildly or severely exposed, against 3% to 4% for Alt-lite or I.D.W. users, and roughly 4% were lightly exposed versus approximately 8% for Alt-lite or I.D.W users.

When teasing apart users that commented only on Alt-lite or only on I.D.W. content, we find that, not only users who commented *only* on I.D.W. get less exposed, but increasingly less so. The same applies to the media channels. For example, the exposure rates of users who watched only Alt-lite (second row) or only I.D.W. (third row) content are much more similar for those tracked in 2006 – 2012 (first column) than for those tracked in 2017 (last column). For users who were tracked in 2006 – 2012, around 15% were exposed in both scenarios, while for those tracked in 2017, this difference grew farther apart (~ 12% Alt-lite vs. ~ 6% I.D.W.).

The previous study suggests that the pipeline effect does exist, and that indeed, users systematically go from milder communities to the Alt-right. However, it does not give insight into how expressive the effect is in terms of what part of the Alt-right user base has gone through it. We address this question by tracking users exactly as we did before, and then analyzing what percentage of exposed

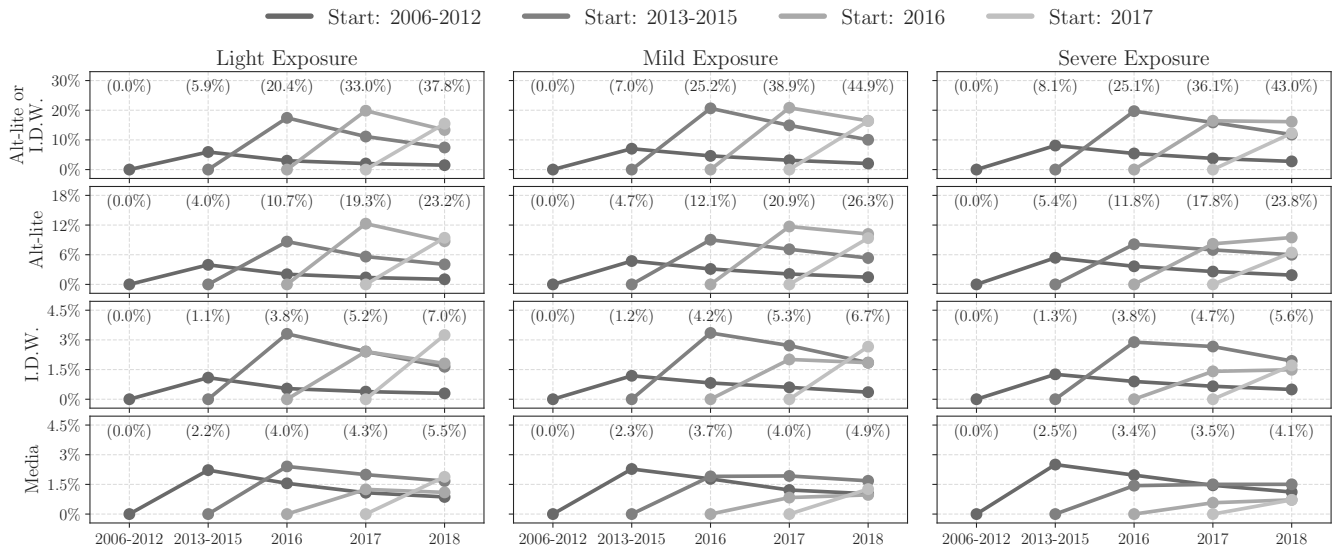

Figure 4: We show how expressive the tracked users are in terms of the Alt-right user base. Each row shows a different condition for tracking users and each column shows a different level of exposure. Each line corresponds to users tracked at a different starting date (in the x-axis), and the y-axis shows the percentage of the total Alt-right commenting users they went to become (notice that all lines begin at 0, because initially they did not consume any Alt-right content).

users at each year can be traced back to users who initially watched content from other communities. In other terms, for each year we calculate, of the users who are exposed (i.e. who watched Alt-right videos), which percentage belongs to each one of the sets of tracked users we just described.

The results for this analysis are shown in Fig. 4. We find that these users are a considerable fraction of the Alt-right commenting audience. In 2018, for all kinds of exposure, roughly 40% of commenting users can be traced back from cohorts of users that commented only on Alt-lite or I.D.W. videos in the past. This can be seen in the first row of the plot. Moreover, we can observe that, consistently, users who consumed Alt-lite or I.D.W. content in a given year, go on to become a significant fraction of the Alt-right user base in the following year. This number is much more expressive than the number of users which came from media channels — in the last row — which never surpasses 6% for any level of exposure.

Looking at the second and third row of Fig. 4, we find a substantial difference between the I.D.W. and the Alt-lite. Whereas in Sec. 5 we find that the intersection between them both and the Alt-right are similar, here we see that users who initially commented *only* on I.D.W. channels constitute a much less significant percentage of the Alt-right consumer base in upcoming years. For all levels of exposure, at all times, the number of exposed users that can be traced back to commenting exclusively on I.D.W. channels is around 3 times lower. So, while in 2018, 23.3% of users who were lightly exposed can be traced back to users who commented on Alt-lite channels in previous years, only 7.6% can be traced back to I.D.W. channels. Overall, in both analyses, users who consumed only I.D.W. channel seem to behave more similarly to the users in the media channels. Yet, as we see in Sec. 5, the intersection between the Alt-lite and the I.D.W. is increasing with time, which means this population is becoming less significant.

The experiments performed show that, not only the commenting user bases are becoming increasingly similar (as shown in Sec. 5), but that, systematically, users who commented only on I.D.W. or Alt-lite content go on to comment on Alt-right channels. This phenomenon is significant both in terms of the percentage of the users tracked — as in Fig. 3 — and in terms of the total Alt-right commenting user base — as in Fig. 4. We present the raw numbers associated with these figures in Appendix D.

7 THE RECOMMENDATION ALGORITHM

In this section, we inspect the impact of YouTube's recommendation algorithm. Unfortunately, we have only a snapshot of the recommender system which does not take into account personalization. Thus, it is hard to reach significant conclusions on what was the role of the recommender system in the radicalization process we depicted in Sec. 6. Yet, we argue that analyzing these data is relevant, for it is a blueprint of how the influence of the recommender

Table 3: Percentage of edges in-between communities in the recommendation graphs (normalized per weight). Video recommendations are in bold. Rows indicate the source of edges columns indicate their destination.

Src|Dst	I.D.W.	Alt-lite	Alt-right	Media	Other
I.D.W.	52.78/**19.03**	22.88/**1.57**	0/**0.03**	3.12/**3.03**	21.23/**76.35**
Alt-L	13.69/**2.46**	55.15/**12.70**	3.38/**0.13**	2.82/**3.24**	24.96/**81.47**
Alt-R	25.73/**1.89**	42.94/**1.15**	25.73/**8.55**	1.35/**3.38**	21.08/**85.03**
Media	4.94/**0.31**	4.36/**0.08**	0/**0**	28.78/**14.84**	61.92/**84.77**

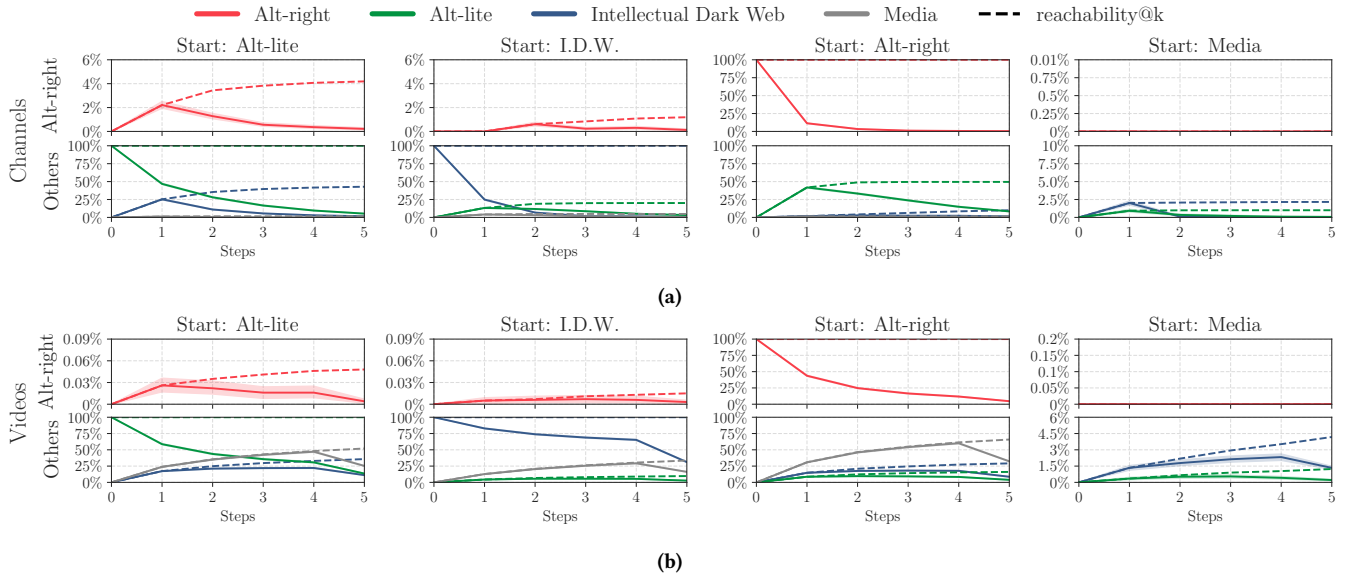

Figure 5: We show the results for the simulation of random walks for channels (a) and videos (b). We show two metrics, as described in the text, the probability of the walker being in a given community at each step (solid line) and the reachability at each step for a given community (dashed line). The different columns portray different starting rules for the initial node in the simulations. Error bands are 95% confidence intervals.

system may be measured, and because it allows us to understand how the recommender system is behaving for our scenario.

We perform our analysis in a recommendation graph, built using the data collected. The graph is built as follows: for each channel, we join together all recommendations obtained in all rounds of data collection. Each channel is a node, and edges between nodes indicate recommendations from a channel to another (for both video and channel recommendations). Notice that, in case there was a recommendation towards a channel or a video we are not aware of, we add an edge to a special sink node we name "Other". Each edge is weighted proportionally to the number of times that recommendation appeared in the data collection, and weights are normalized so that outgoing edges of each node sum up to 1.

The percentage of edges between communities (normalized by their weight) is shown in Tab. 3 for channel and video recommendations. For channel recommendations, we have that media channels are recommended scarcely by the communities of interest. In fact, there are more edges flowing *out* of media channels towards Alt-lite/I.D.W. channels than the other way around. Alt-lite and I.D.W. channels recommend channels from the same community around 50% of the time, and recommend each other around 14% (Alt-L to I.D.W.) and 23% (I.D.W. to Alt-L) of the time. Alt-right channels are only recommended by Alt-lite channels (3.08%). For video recommendations, there is a high prevalence of recommendation to videos we were not able to track (more than 75% of outgoing edges from all communities pointed towards the "Other" node). We also find that media channels are more often recommended in this setting (~ 3% for all communities), while the Alt-lite and the I.D.W. recommend each other roughly 2% of the time. Lastly, Alt-right videos are not significantly recommended here.

Given these graphs, we experiment with random walks. The random walker begins in a random node, chosen with chance proportional to the number of subscribers in each channel. Then, the random walker randomly navigates the graph for 5 steps, choosing edges at random with probabilities proportional to their weights. We store the random walks and calculate two metrics: *(i)* the probability of it being in a channel from each of the communities, that is, the probability that there is a channel of a given community in the k-th step. *(ii)* the reachability of each of the communities at step k. That is, at step k, the percentage of times that the random walker has found a node of a given community. We run the simulation 10K times for scenarios where the initial node is restricted to one of the three communities or the media channels.

Importantly, we consider a small difference in the experimental set-up for each of the graphs. In the channel recommendation graph, we allow the random walker to choose the "Other" node. When this happens the walk stops, thus at each step there is a probability this walk is interrupted by this — or by the fact that there are no recommended channels. In the channel recommendation graph, as the number of edges to the "Other" node is too high, we do not allow the random walker to go towards it. Notice that the scenario for the channels is more realistic, and we give more weight to the conclusions drawn there. The two aforementioned metrics, at each step, given different starting conditions, are shown in Fig. 5, for channel and video recommendations.

For channel recommendations, we have that the reachability@5 of Alt-right channels is of approximately 4% for the simulations starting from Alt-lite 1.5% for I.D.W. channels. Moreover, starting from an I.D.W. channel, users have approximately 10% of chance of being in an Alt-lite channel at the next step, and in 5 steps, there is 25% of chance that the user has found at least one Alt-lite channel.

Starting from the media channels, reachability@5 of I.D.W. channels is of 2.5%, and of slightly less than 1% for Alt-lite channels. These can be seen on the bottom row of Fig. 5 (a).

For video recommendations, reaching Alt-right channels from other communities is less likely. From the Alt-lite, reachability@5 is of around 0.05%. Going from the I.D.W. to the Alt-lite is more difficult: the reachability@5 is roughly 7%. More relevant, though, starting from media channels, the reachability@5 of I.D.W. and Alt-lite channels is of around 4.5% and 1.5% respectively. It is worth recalling that this experiment is less realistic than the former, as here we ignore the possibility of the random walker being in a video we are not aware of.

Overall, we find that, in the channel recommender system, it is easy to navigate from the I.D.W. to the Alt-lite (and vice-versa), and it is possible to find Alt-right channels. From the Alt-lite we follow the recommender system 5 times, approximately 1 out of each 25 times we will have spotted an Alt-right channel (as seen in Fig. 5 (a)). In the video recommender system, Alt-right channels are less recommended, but finding Alt-lite channels from the I.D.W. and I.D.W. channels from the large media channels in the media group is also feasible. Considering the sheer amount of views the channels in the Alt-lite, the I.D.W. and the Alt-lite, these percentages, although low, may result in a very significant number of views towards fringe content. This process may also be amplified when taking personalization into account. Notice that we depict the two graphs in which we performed our experiments in Appendix E.

8 DISCUSSION

We performed a through analysis of three YouTube communities — the I.D.W., the Alt-lite, and the Alt-right — inspecting a large dataset with millions of comments and recommendations from thousands of videos. In this section, we discuss how the insights of our analyses shed light into our research questions. We also talk about the limitations and potential implications of this work.

RQ1. How have these channels grown on YouTube in the last decade? The three communities studied sky-rocketed in terms of views, likes, videos published and comments, particularly, since 2015, coinciding with the presidential election of that year, as shown in Sec. 4. However, this seems to be the case not only for these communities, but also for the larger channels in the media group. A key difference between the communities and media channels lies in the engagement of their users. The number of comments per view seems to be particularly high for extreme content (Sec. 4), and users in all three communities are more assiduous commentators than in the media channels (Sec. 5).

RQ2. To which extent do users systematically gravitate towards more extreme content? We find that the commenting user bases for the three communities are increasingly similar (Sec. 5), and, considering Alt-right channels as a proxy for extreme content, that a significant amount of commenting users systematically migrates from commenting exclusively on milder content to commenting on more extreme content (Sec. 7). We argue that this finding provides significant evidence that there has been, and there continues to be, user radicalization on YouTube, and our analyses of the activity of these communities (Sec. 4) is consistent with the theory that more extreme content "piggybacked" on the surge in popularity of I.D.W. and Alt-lite content [30]. We show that this migration phenomenon is not only consistent throughout the years, but also that it is significant in its absolute quantity. Noticeably, the findings related to this research question make the implicit assumption that commenting users are a good enough proxy for radicalization, and that comments in YouTube channels are supportive of the videos they are associated with. We established the validity of these assumptions as follows. First, the sheer number of comments and high prevalence of comments per views in Alt-right videos suggest that commenting users are a population worth studying, especially when in Sec. 4 we found that Alt-right channels have a very high percentage of comments per view. Secondly, during the three week annotation period, it was noted that the number of opposing comments is rather small, as we found by manually checking 900 randomly selected comments (300 for each community of interest), finding that only 5 could be interpreted as criticisms to the videos they were associated with. Moreover, we note that the proportion of likes for the communities of interest is higher for the communities of interest (> 91% mean, > 96% median) than for the media channels (85% mean, 93% median), which suggests the people interacting with the three communities agree with their videos.

RQ3. Do algorithmic recommendations steer users towards more extreme content? Our simulations suggest that YouTube's recommendation algorithms frequently suggest Alt-lite and I.D.W. content. From these two communities, it is possible to find Alt-right content from recommended channels, but not from recommended videos. Noticeably, our analysis has several shortcomings which do not allow us to make bold claims about this research question. Firstly, we are able to look only at a tiny fraction of actual recommendations — it could very well be that Alt-right content was being more widely promoted in the past. Secondly, our analysis does not take into account personalization, which could reveal a completely different picture. Still, even without personalization, we were still able to find a path in which users could find extreme content from large media channels.

Limitations and future work. Our work resonates with the narrative that there is a radicalization pipeline [36, 41]. Indeed, we manage to measure traces of user radicalization using commenting users. Although we argue this is strong evidence for the existence of radicalization pathways on YouTube, our work provides little insight on *why* these radicalization pipelines exist. Elucidating the causes of radicalization is an important direction to better understand user radicalization and the influence of social media in our lives. Moreover, in this paper we focused exclusively on basic statistics (likes, views and comments) and on the trajectory of users, be they inferred through comments or simulated in the recommendation graphs. Another interesting direction would be to trace the evolution of the *speech* of content creators and commenting users throughout the years, to study what are the narratives that arose and how their tone has changed.

Acknowledgements. We gratefully acknowledge support from the Brazilian agencies CNPq, Capes and Fapemig, from the projects Atmosphere, INCT-Cyber and MASWEB, from a Google Research Award for Latin America (Manoel Horta Ribeiro). We thank Jeremy (Jimmy) Blackburn for helpful discussions.

REFERENCES

[1] [n. d.]. Media Bias/Fact Check - Search and Learn the Bias of News Media. https://mediabiasfactcheck.com/

[2] ADL. [n. d.]. Glossary Terms Alt-Right. https://www.adl.org/resources/glossary-terms/alt-right

[3] ADL. 2019. From Alt Right to Alt Lite: Naming the Hate. https://web.archive.org/web/20190422202936/https://www.adl.org/resources/backgrounders/from-alt-right-to-alt-lite-naming-the-hate

[4] Swati Agarwal and Ashish Sureka. 2014. A Focused Crawler for Mining Hate and Extremism Promoting Videos on YouTube.. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. ACM.

[5] Monica Anderson and Jingjing Jiang. 2018. *Teens, Social Media & Technology 2018*. Technical Report. Pew Research Center. https://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/

[6] Andywarski. 2018. Richard Spencer, Styx and Sargon Have a Chat - Andy and JF moderate. https://web.archive.org/web/20190616134137/https://www.youtube.com/watch?v=UiUH-tWHbr8

[7] Anonymous. [n. d.]. The Intellectual Dark Web. https://web.archive.org/web/20190407170300/http://intellectualdark.website/

[8] David C. Atkinson. 2018. Charlottesville and the alt-right: a turning point? *Politics, Groups, and Identities* (2018).

[9] Khaled A. Beydoun. 2018. US liberal Islamophobia is rising – and more insidious than rightwing bigotry | Khaled A Beydoun. *The Guardian* (2018). https://www.theguardian.com/commentisfree/2018/may/26/us-liberal-islamophobia-rising-more-insidious

[10] Tyler Bridges. 2018. "Alt-Lite" Bloggers and the Conservative Ecosystem. https://shorensteincenter.org/alt-lite-bloggers-conservative-ecosystem/

[11] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM.

[12] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. 2010. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM.

[13] Nicholas Diakopoulos. 2014. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. (2014).

[14] Benjamin G. Edelman and Michael Luca. 2014. *Digital Discrimination: The Case of Airbnb.com*. Technical Report. Social Science Research Network. https://papers.ssrn.com/abstract=2377353

[15] Lisa W. Foderaro. 2018. Alexandria Ocasio-Cortez Likens $10,000 Debate Offer by Conservative Columnist to Catcalling. *The New York Times* (2018). https://www.nytimes.com/2018/08/10/nyregion/alexandria-ocasio-cortez-debate-catcalling-ben-shapiro.html

[16] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis. 2010. A Multimodal Approach to Violence Detection in Video Sharing Sites. In *2010 20th International Conference on Pattern Recognition*.

[17] Rosie Gray. 2015. How 2015 Fueled The Rise Of The Freewheeling, White Nationalist Alt- Movement. https://www.buzzfeednews.com/article/rosiegray/how-2015-fueled-the-rise-of-the-freewheeling-white-nationali

[18] Keegan Hankes and Alex Amend. 2018. The Alt-Right is Killing People. https://www.splcenter.org/20180205/alt-right-killing-people

[19] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM.

[20] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14)*. ACM.

[21] Mathew Ingram. 2018. Most Americans say they have lost trust in the media. *Columbia Journalism Review* (2018). https://www.cjr.org/the_media_today/trust-in-media-down.php

[22] Brendan Joel Kelly. 2017. Lauren Southern: The alt-right's Canadian dog whistler. https://www.splcenter.org/hatewatch/2017/11/07/lauren-southern-alt-right%E2%80%99s-canadian-dog-whistler

[23] Max Kutner. 2016. Roosh V's journey from pickup artist to right-wing provocateur. *Newsweek* (2016). https://www.newsweek.com/2016/10/21/roosh-v-pickup-artist-right-wing-provocateur-509319.html

[24] Rebecca Lewis. 2018. *Alternative influence: Broadcasting the reactionary right on YouTube*. Technical Report. Data and Society.

[25] Cristina Lopez G. 2019. Stefan Molyneux is MAGA Twitter's favorite white nationalist. https://www.mediamatters.org/people/stefan-molyneux

[26] Tim Lott. 2017. Jordan Peterson and the transgender wars. https://life.spectator.co.uk/2017/09/jordan-peterson-and-the-transgender-wars/

[27] Alex Mann, Kevin Nguyen, and Katherine Gregory. 2019. 'Emperor Cottrell': Accused Christchurch shooter had celebrated rise of the Australian far-right.

[28] Andrew Marantz. 2017. The Alt-Right Branding War Has Torn the Movement in Two. (2017). https://www.newyorker.com/news/news-desk/the-alt-right-branding-war-has-torn-the-movement-in-two

[29] Clark McCauley and Sophia Moskalenko. 2008. Mechanisms of Political Radicalization: Pathways Toward Terrorism. *Terrorism and Political Violence* (2008).

[30] Angela Nagle. 2017. *Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right*.

[31] Newman Nic, Richard Fletcher, Antonis Kalogeropoulos, David AL Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News Report 2018*. Technical Report. Reuters Institute for the Study of Journalism.

[32] Raphael Ottoni, Evandro Cunha, Gabriel Magno, Pedro Bernardina, Wagner Meira Jr., and Virgílio Almeida. 2018. Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '18)*. ACM.

[33] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. 2015. Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems. *Social Science Computer Review* (2015).

[34] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. 2019. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *arXiv:1901.07046 [cs]* (2019). http://arxiv.org/abs/1901.07046

[35] PowerfulJRE. 2017. Sargon of Akkad - Joe Rogan Experience #979. https://web.archive.org/web/20190616134332/https://www.youtube.com/watch?v=xrBCsLsSD2E

[36] Kevin Roose. 2019. The Making of a YouTube Radical. *The New York Times* (2019). https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html,https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html

[37] Roosh V. 2016. I Do Not Disavow Richard Spencer. https://web.archive.org/web/20190331073037/https://www.rooshv.com/i-do-not-disavow-richard-spencer

[38] Andrew Sellars. 2016. *Defining Hate Speech*. SSRN Scholarly Paper. Social Science Research Network. https://papers.ssrn.com/abstract=2882244

[39] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. 2010. Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. In *Information Retrieval Technology (Lecture Notes in Computer Science)*. Springer Berlin Heidelberg.

[40] The Rubin Report. 2018. Eric Weinstein: The Future of The Intellectual Dark Web. https://www.youtube.com/watch?v=tUl7-SvntQ4

[41] Zeynep Tufekci. 2018. YouTube, the Great Radicalizer. *The New York Times* (2018). https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

[42] Bari Weiss and Damon Winter. 2018. Meet the Renegades of the Intellectual Dark Web. *The New York Times* (2018). https://www.nytimes.com/2018/05/08/opinion/intellectual-dark-web.html

[43] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM.

[44] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources. In *Proceedings of the 2017 Internet Measurement Conference*. ACM.