DESINFORMAÇÃO, RADICALIZAÇÃO E ÓDIO NA PERSPECTIVA DOS USUÁRIOS

MANOEL HORTA RIBEIRO

DESINFORMAÇÃO, RADICALIZAÇÃO E ÓDIO NA PERSPECTIVA DOS USUÁRIOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: WAGNER MEIRA JR.   COORIENTADOR: VIRGÍLIO AUGUSTO FERNANDES ALMEIDA

Belo Horizonte

Julho de 2019

MANOEL HORTA RIBEIRO

MISINFORMATION, RADICALIZATION AND HATE THROUGH THE LENS OF USERS

Dissertation presented to the Graduate Program in Computer Science of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

ADVISOR: WAGNER MEIRA JR.   COORIENTADOR: VIRGÍLIO AUGUSTO FERNANDES ALMEIDA

Belo Horizonte

July 2019

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Misinformation, Radicalization and Hate Through the Lens of Users

MANOEL HORTA RIBEIRO
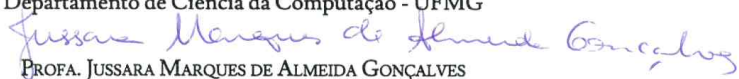
Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. WAGNER MEIRA JÚNIOR - Orientador
Departamento de Ciência da Computação - UFMG

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Coorientador
Departamento de Ciência da Computação - UFMG

PROF. LUIS DA CUNHA LAMB
Departamento de Informática Teórica - UFRGS

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 13 de Agosto de 2019.

para Jessica, com amor

Acknowledgments

Em 2013, quando ingressei na graduação, pedi uma oportunidade de IC para o Meira, agora meu orientador. Durante a graduação e o mestrado, ele me ajudou a descobrir minha paixão por pesquisa, me deu a liberdade para buscar tópicos que me motivavam, e me aconselhou de maneira que me fez crescer como pesquisador e como pessoa. Em 2017, tive o prazer de começar a trabalhar também com o Virgílio, que me co-orientou. Ele me ensinou a procurar (e responder) perguntas científicas que fazem diferença na sociedade, e a enxergar o mundo de maneira multidisciplinar e cosmopolita. O que aprendi com vocês vou levar para sempre comigo.

Meus dias no ICEx foram alegrados pela presença de amigos: Derick, Carlos, Walter, Rodrigo, Osvaldo, Ewerton, Karen, Roberto, Túlio, Emanuel, Evandro, Gabriel, Josemar. Obrigado pelas conversas e pela ajuda durante a temida monitoria de AEDS III. No DCC, também encontrei pessoas que viriam a ser meus co-autores: gostaria de agradecer ao Yuri, ao Pedro Calais e ao Rapha, que acrescentaram muito nos papers que originaram essa dissertação.

Os funcionários (docentes ou não) do DCC também não poderiam deixar de ser mencionados. Recebi todo o apoio que podia esperar da secretaria da pós do DCC. No sétimo andar, não só roubei café dos professores como fiz amizades e recebi conselhos. Queria agradecer em especial ao Fabrício e ao Renato Assunção pelos bate-papos.

Nos paragráfos abaixo, aproveito para agradecer pessoas com contribuições menos diretas para esse trabalho, porém não menos importantes. Meus amigos de Colégio Santo Antônio e de joguinho me acompanharam durante toda essa jornada, e sem Gansos nem Meusba Clan, ela seria bem menos divertida. Meus amigos de graduação também sempre estiveram presentes: Nini, Araroba, Nanato e Vivi.

Last (European) summer, I had the opportunity to do a research internship in Switzerland at EPFL. There, Gli, Tizi, Ramtin, Kiriusha, Luci and Bob taught me a lot about doing research (and about the internal workings of orange juice machines). Looking forward to join (most of) you guys soon. Also, I would like to thank the folks at iDrama, particularly "Jimmy" Blackburn, for all the feedback he gave me.

Também tive a oportunidade de trabalhar em um projeto interdisciplinar no HC durante boa parte do mestrado, o CODE. Em reuniões semanais, aprendi com Paulo, Gabriela, Jéssica, Derick (de novo), Milton, meu pai e meu irmão, um pouco mais sobre as dificuldades do trabalho interdisciplinar e de trabalhar com médicos em geral (hahaha).

E por falar em família, sem o apoio dela, nada disso seria possível, já que eles vêm me apoiando desde quando eu era uma criança que "perguntava demais" na escola. São minhas referências como pesquisadores e como pessoas; e compartilharam comigo as minhas aventuras nos confins da internet em jantares animados. Minha mãe esteve do meu lado nas horas mais complicadas da minha vida. Ela foi minha confidente, minha amiga, e me deu todo o carinho que um filho pode precisar (muitas vezes com um suquinho de laranja). Obrigado por me entender (mesmo eu sendo tão bobo). Meu pai colocou algum juízo em mim e me ajudou a sonhar meu futuro pessoal e profissional. Tenho muita sorte de ter um pai que sempre cativou o melhor em mim. Quero ser um cinquentão bacana igual você (apesar de que provavelmente serei mais careca). Meu irmão foi meu companheiro de cursos online, de projetos paralelos ligeiramente megalomaníacos, e de discussões sobre assunto dos quais ele claramente sabia mais do que eu. Minhas cadelinhas, Laika, Cacau e Jasmim me acompanharam durante incontáveis tardes de trabalho, e voltinhas no quarteirão para descansar um pouco. Amo vocês todos.

Nos últimos meses, Jessica foi minha companheira de *modo deadline*, e de discussões animadas sobre os tópicos que eu estudei. Durante a escrita desse texto, você foi minha namorada, minha noiva e minha esposa. Dedico essa dissertação pra você.

"The metaphysicians of Tlön do not seek for the truth
or even for verisimilitude, but rather for the astounding."
(Tlön, Uqbar, Orbis Tertius (1940))

Resumo

A popularização das redes sociais mudou a dinâmica de criação e consumo de conteúdo. Barreiras para disseminar textos, imagens e vídeos tornaram-se significativamente menores do que em épocas anteriores, capacitando os usuários a criar, com poucos recursos, conteúdo de impacto e de longo alcance. Neste cenário, a sociedade testemunhou a amplificação de fenômenos como a desinformação e discurso de ódio.

Pesquisas recentes tentam resolver esses problemas estudando conteúdo odioso ou falso. No entanto, abordar com robustez esses fenômenos desta maneira muitas vezes não é viável. Considere, por exemplo, a tarefa de detectar o discurso de ódio. Pode haver dois textos iguais que, em diferentes contextos (por exemplo, uma letra de rap e o discurso de um político) podem ser considerados odiosos ou não. Além disso, pode ser que, dado um texto dentro de um contexto, dois indivíduos discordem sobre se o conteúdo é odioso ou não.

Esta dissertação tenta abordar estes problemas tomando outra perspectiva: a do usuário. Através da perspectiva do usuário, é possível abordar conteúdos possivelmente falsos ou odiosos com o contexto circundante: orientação política, padrões de atividade, conexões, etc. Além disso, somos capazes de estudar fenômenos mais complexos, como a radicalização de usuários, onde devemos estudar as trajetórias dos indivíduos —ou, mais realisticamente, seus traços on-line.

Em três estudos de caso em redes sociais, nós: (i) fornecemos insights sobre como a percepção do que é desinformação é alterada pela opinião política; (ii) propomos uma metodologia para estudar o discurso de ódio no nível do usuário, mostrando que a estrutura de rede dos usuários pode melhorar muito a detecção do fenômeno; e (iii) caracterizamos a radicalização de usuários em canais no YouTube ao longo do tempo, mostrando uma migração crescente para canais mais extremos.

Cada estudo de caso contribui para seu assunto mais específico: desinformação, discurso de ódio e radicalização. Em conjunto, eles suportam um argumento central: o de que devemos estudar fenômenos mal definidos como desinformação e o discurso de ódio sob a perspectiva dos usuários.

Abstract

The popularization of Online Social Networks has changed the dynamics of content creation and consumption. Barriers to disseminate texts, images and videos became significantly lower than in earlier times, empowering users to create, with little resources, content with far-reaching impact. In this setting, society has witnessed an amplification in phenomena such as misinformation and hate speech.

Recent research attempts to address these issues by studying hateful or fake content. Yet, robustly addressing these phenomena in this fashion is often not feasible. Consider for example the task of detecting hate speech. There can be two exact texts that, in different contexts (e.g. a rap lyric and a politician's speech), may be considered to be hateful or not. Moreover, there may even be that, given a text within a context, two individuals disagree on whether the content is hateful or not.

This dissertation attempts to address these issues by taking another perspective: that of the user. Through the lens of users, it is possible to approach possibly fake or hateful content with surrounding context: political orientation, activity patterns, connections, etc. Furthermore, we can study more complex phenomenon, such as user radicalization, where one must study the trajectories of individuals – or, more realistically, their online traces.

In three case studies on social networks, we: (i) provide insight on how the perception of what is misinformation is altered by political opinion; (ii) propose a methodology to study hate speech on a user-level, showing that the network structure of users can improve the detection of the phenomenon; (iii) characterize user radicalization in far-right channels on YouTube through time, showing a growing migration towards the consumption of extreme content in the platform.

Each case study contributes to their more specific subject: misinformation, hate speech and user radicalization. Yet, altogether, they advance a central argument: that studying users rather than the content itself is more productive to better understand (and eventually mitigate) ill-defined social phenomena such as hate speech and fake news.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In 1998, a publication by the U.S. Department of Defense (DoD) defined the term *information environment* as: *the aggregate of individuals, organizations, or systems that collect, process, or disseminate information, including the information itself* [Shelton, 1998]. At the time, institutions pointed out as key parts of the information environment included industry, academia, government, and commercial networks. Although ahead of its time, the report did not foresee the role of social media, which undoubtedly shapes today's information environment [Rainie et al., 2017].

The information environment in the times of social networks is very different from the picture painted by the DoD in 1998. The news ecosystem was deeply transformed: users increasingly consume more news-pieces and opinion content in social media [Gottfried and Shearer, 2016]; business models for traditional media organizations evolved [Newman, 2011], and their importance as gatekeepers diminished in favor of alternative sources [Lianne and Simmonds, 2013]. On darker corners of the internet, fringe websites, like *4chan*, and subreddits, like */r/TheDonald*, have great influence over which memes and news are shared in large social networks, such as Twitter [Zannettou et al., 2018b, Zannettou et al., 2018a], and often promote harassment campaigns and hateful narratives [Nagle, 2017]. On social media, in the feeds of websites like Facebook and Twitter, the information users are exposed to is selected through recommendation algorithms [Liao and Fu, 2013]. These black-boxes, tuned to optimize engagement, were accused of separating users from news and opinions they disagreed with [Pariser, 2011].

Overall, we can identify two troublesome phenomena that were amplified by this new information environment: the dissemination of hateful content and of misinformation (or *fake news*). As we discuss later, part of the challenges we propose to approach have to do with the subjectivity of such concepts. Yet, we broadly define them:

- **Fake news** is a recently popularized term which refers to fabricated or excessively biased news created with the intention to manipulate, deceive or (in case of satire) entertain users [Tandoc Jr et al., 2018].

- **Hate speech** is speech that targets a group, or individuals as members of a group, causing or intending harm, often as actions beyond the speech itself. It is often expressed publicly or directly at members of the group, in a context where violent response is possible [Sellars, 2016].

Although it is often hard to pinpoint exactly what constitutes either of these phenomena, their societal impact is significant. Fringe ideologies like White Supremacy got their voices amplified through online movements such as the Alt-right [ADL, 2019b], motivating terrorist attacks such as the one in Christchurch, NZ [Mann et al., 2019]. In the U.S., during the 2016 presidential election, researchers estimate that the average American "read and remember on the order of one, or perhaps several fake news articles" [Allcott and Gentzkow, 2017]. Worryingly, researchers and the media indicate that these false pieces of information were partially driven by an orchestrated effort to promote political turmoil [Ferrara et al., 2016, Zannettou et al., 2019].

In this scenario, ways of mitigating the diffusion of hate speech and fake news are clearly necessary. Yet, there are several challenges involved with that task. Moderating this content is hard due to the sheer amount of images and comments produced every day by users in OSNs [Schmidt and Wiegand, 2017], due to the inherent friction with values such as freedom of expression [Rainie et al., 2017], and due to the hardness to determine what exactly is fake or what is hateful [Davidson et al., 2017], which may differ, for example, in different cultures [NW et al., 2015]. These difficulties have been obstacles for both mass-hired human moderators [Julia Angwin, 2017], and for attempts to use automated techniques to characterize and detect these issues. For example, hate speech detection models fail to differentiate between harmful speech and offensive speech [Davidson et al., 2017], and several fake news detection models capture only stylistic cues, often not sufficient to tell whether something is fake [Shu et al., 2017]. Moreover, the dissemination of such information happens in an "adversarial" environment, in which agents may be spreading content to further a certain political agenda, which means that borderline cases will be exploited [Zannettou et al., 2019]. These challenges, which are largely shared in the task of tackling both hate speech and misinformation, make it logical to study these phenomena together. We argue that the development of methods to characterize and detect hate speech will further our understanding of fake-news (and vice-versa), as a big part of the challenge with dealing with both these phenomena is to deal with the elusiveness of their definition.

In this dissertation, we propose automated methodologies to characterize and detect hate speech and fake news by aggregating data at the *user level*. We argue that: *(i)* these two steps —characterization and detection— are essential parts in the larger task of mitigating these phenomena; and *(ii)* adopting a user-centric perspective —one which considers users as the central unit of study— allows to better address the aforementioned challenges. Throughout the chapters, we show examples of how focusing on users allowed us to better understand the nuances of these ill-defined but high-impact social phenomena; to develop detection methods better suited for the real world; and, lastly, to study more complicated processes, such as the alleged radicalization of users which is happening on YouTube [Lewis, 2018] —amidst plenty of hateful and fake content.

Before further describing the achievements of this work, we use hate speech detection as a motivating example for our user-centric approach. Consider the tweet: `Timesup, yall getting w should have happened long ago`. Which was in reply to another tweet that mentioned the holocaust. Although the tweet, whose author's profile contained white-supremacy imagery, incited violence, it is hard to conceive how this could be detected as hateful with only textual features. Furthermore, the lack of hate-related words makes it difficult for this kind of tweet to even be sampled in text-based approaches.

Fortunately, the data in posts, tweets or messages are not the only signals we may use to study hate speech in OSNs. Most often, these signals are linked to a profile representing a person or organization. Considering this profile, we could use plenty of other information to try to determine if the *user* is engaging in hateful behavior: other tweets, their network of friends and retweets, their activity patterns. The case can be made that this wider context is sometimes *needed* to define hate speech, such as in the example, where the abuse was made clear by the neo-nazi signs in the user's profile.

This motivates our user-centered approach, which is able to take into consideration all this extra information. Characterizing and detecting hateful *users* shares much of the benefits of detecting hateful content and presents plenty of opportunities to explore a richer feature space. Furthermore, on a practical hate speech guideline enforcement process, containing humans in the loop, its is natural that content needs to be surrounded with user context [1].

---

[1]This is present, for example, in YouTube's [Google, 2019] and Twitter's [Twitter, 2019] hateful conduct guidelines. To quote directly from Twitter Rules: *Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.*

The aforementioned example inspired one of the three case studies we present in the following chapters —one case study per chapter. They go as follows.

- In Chapter 2 we explore the connections between political polarization and the spread of fake news. By monitoring URLs and tweets associated with fake news related hashtags and keywords, we find the concept of "fake" to be highly associated with political polarization, and that users employ terms such as fake news to refer to content that they particularly disagree with. This touches on the already mentioned topic of the ill-definition of what is hate or fake.

- In Chapter 3 we explore —as mentioned in the motivating example— how it may be helpful to characterize and detect hateful *users*, rather than hateful *content*. Our characterization sheds light on how hateful users differ from normal ones with respect to their user activity patterns, network centrality measurements, and the content they produce. We show that these differences can be exploited to robustly detect such users.

- Lastly, in Chapter 4 we study the phenomena of user radicalization on YouTube. We find that users consistently migrate from milder to more extreme content; and that a large percentage of users who consume extreme content now, consumed milder content in the past. We also probe YouTube's recommendation algorithm, showing that more extreme content was not particularly favored by recommendation systems in the platform.

Each case study contributes to their more specific subject: misinformation, hate speech and user radicalization. Yet, altogether, they advance a central argument: that studying users rather than content itself is more productive to better understand (and eventually mitigate) ill-defined social phenomena such as hate speech and fake news.

Chapter 2

Polarized Users and Misinformation

Two phenomena have been increasingly receiving attention due to their potential impact on important societal processes [Allcott and Gentzkow, 2017, Bakshy et al., 2015]: the rapid spread of a growing number of unsubstantiated or false information online [Vicario et al., 2016], coined as *fake news*, and the increase of opinion polarization [Allcott and Gentzkow, 2017]. Previous studies suggest a dual interaction between the two: polarized communities are more susceptible to the dissemination of misinformation [Vicario et al., 2016], and, conversely, misinformation plays a key role in creating polarized groups [Zollo et al., 2015]. Another way these phenomena may interact is when users incorrectly classify news as being fake because they disagree with the narrative the news-piece pushes, and not because it reports actual false or imprecise facts [Lewandowsky et al., 2012]. This behavior creates alternative narratives of what is actually fake, which depends on one's political ideology, and that, ultimately, make the line between biased and fake information blurrier.

In this chapter, we conduct an initial analysis of the relationships and interactions between polarized debate and the perception of misinformation on Twitter data. We examine the following research questions:

RQ1 *How is polarization related to information perceived as or related to fake news?*

RQ2 *Are users designating content that they disagree with as misinformation?*

We analyze a dataset composed of tweets of content associated with fake news and general tweets about U.S. politics. Our methodology employs a community detection method designed to estimate the degree of polarization of each user leaning towards the Democratic or Republican parties, as depicted in Figure 2.1. Based on these estimates, we correlate user polarization levels to their interactions with fake news related tweets
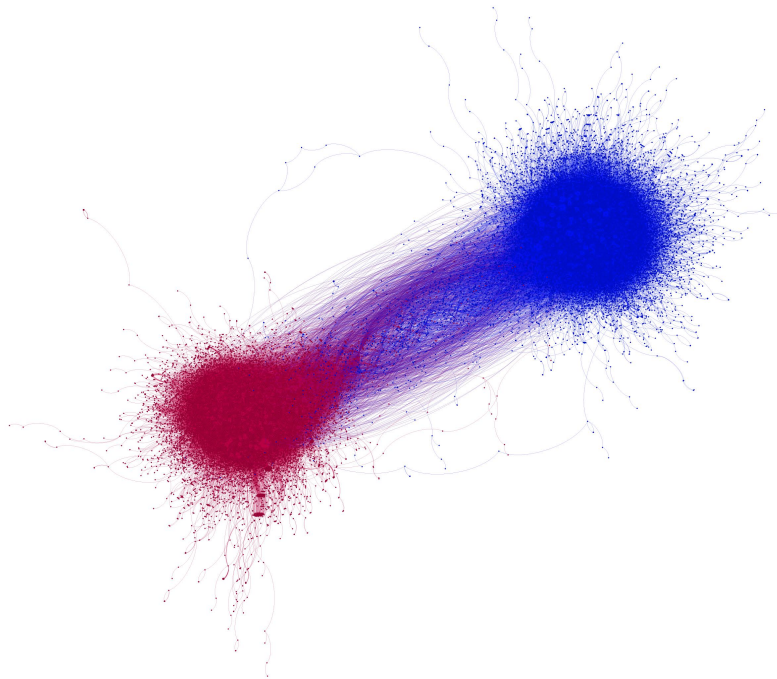
Figure 2.1. Network of retweets showing Democrats (blue) and Republicans (red) divided into two distinct communities. What is the impact of such polarization in what is perceived as fake news?

and external URLs (links to other websites). We analyze how the polarization of such tweets and URLs are related to their popularity and to the frequency they are associated with the theme of misinformation by users. We also analyze the polarization differences between users who merely discuss politics versus those who engage with tweets and URLs labeled by other users as fake news.

Our analyses show three main findings: *(i)* Fake news is associated with polarization. Users that associate content with fake news are more polarized, and content that is associated with fake news receives engagement from users of mostly one side of the political spectrum. *(ii)* Polarized groups cite sources on their side of the political spectrum to tag or condemn news and statements given by the other opposite group as fake; *(iii)* Polarized users employ terms such as fake news to refer to content that they particularly disagree with;

We discuss the impact of these findings in the ongoing battle against the spread of fake news, particularly to detect them. We suggest, for example, that approaches based on crowd-sourcing [Ratkiewicz et al., 2011] may become biased towards political ideologies —once the narratives on what is fake seems to differ among groups with different ideologies— and how machine learning models trained with naive data can do harm, as the umbrella term designates content with different characteristics.

2.1   Background

Online social networks have deeply transformed the news ecosystem [Kumar and Shah, 2018].   In 2016, 62% of adults in the U.S. allegedly obtained news from social media, whereas in 2012, only 49% reported *seeing* news on social media at all [Gottfried and Shearer, 2016].   This popularization process has drastically changed the way newsrooms function and how news outlets profit.   In the last years, traditional news organizations have struggled to maintain a sustainable business model [Newman, 2011, Pew Research, 2018], which has increased the sensationalism of headlines in order to attract clicks from social media [Chakraborty et al., 2016], and decreased resources and time available for newsrooms to conduct research [Mitchell and Page, 2015].   The way individuals consume news also changed.   The information to which users are exposed is selected through recommendation algorithms [Liao and Fu, 2013], which may separate users from information (and news) that disagree with their viewpoints [Pariser, 2011].   Moreover, users and alternative sources began to significantly influence the flow of information, diminishing mainstream media importance as gatekeepers [Lianne and Simmonds, 2013].

In this dynamic setting, social media platforms have become a fertile terrain for the phenomenon broadly referred to as fake news [Zannettou et al., 2018b].   Although trendy, the term fake news is highly ambiguous, as it may be used to refer to distinct phenomena [Lazer et al., 2018, Zannettou et al., 2018b] The term may be used to refer to: *(i)* information that is plainly false; *(ii)* information that intends to mislead or influence readers; *(iii)* information published without proper research or editorial process; and *(iv)* information that is biased.

Different news sources are often accused of spreading fake news for different reasons.   Traditional media outlets like CNN, for example, have been accused by right-wing politicians and pundits as being fake news due to bias [Trump, 2017, Fox News, 2019]; Whereas alternative news sources have been called fake news for, among other reasons, lack of accountability and due journalistic process [Shafer, 2017].   We avoid adopting an explicit definition for the phenomena, as part of the reason of this work is to explore the different perceptions of fake news by users in social media.   Some surveys attempt to create different categories, which leaves less room for ambiguity [Zannettou et al., 2018b, Kumar and Shah, 2018].   Zannettou et al., for example, create a typology which differentiates between Fabricated News, Propaganda, Conspiracy Theories, Biased News, Clickbaits, and others [Zannettou et al., 2018b].

Regardless of the particular definition of fake news, news-pieces with the aforementioned characteristics (false, biased, misleading, lacking editorial process, etc) have

been playing an important role in shaping the public discourse and global politics. In Brazil, *WhatsApp* was in the center of political debate during the 2018 presidential election, as it was widely used to disseminate hoaxes [Tardáguila et al., 2018]. In the U.S., during the 2016 presidential election, researchers estimate that the average American "read and remembered on the order of one or perhaps several fake news articles" [Allcott and Gentzkow, 2017]. In India, rumours have motivated a series of mob-lynchings, which, in 2018 alone, were responsible for more than 20 deaths [Bengaluru et al., 2018]. Overall, fake news has become an effective mechanism to influence public opinion and to create discord, possibly to the benefit of a certain agenda [Zannettou et al., 2019].

Recent research has found clever workarounds against the fuzzy definition of the phenomena. Grinberg et. al [Grinberg et al., 2019], for example, divided fake news sources into three categories which ranged from publishing almost exclusively fabricated stories, to websites where the creation of falsehoods was not systematic. Although careful consideration of sub-categories of fake news may be a successful approach for researchers, what is less clear is whether the general public is correctly parsing through this multi-faceted concept. In this setting, a major concern is that the term fake news is increasingly used to refer to news perceived as biased. This misuse may even be a strategy to dismiss opposing narratives. In the Philippines, for instance, President Duterte dismissed serious criticism by the news outlet Rappler as fake news [Yap, 2018]. In a similar fashion, in Burma, the army has dismissed accusations of Rohingya ethnic cleansing as fake [Lloyd Parry, 2017]. The hijacking of the term fake news, if widespread, would largely contribute to post-factual or post-truth politics, where emotion and ideological biases triumph over facts and empirical data [Corner, 2017]. It would potentially make fake news even harder to moderate, as the disagreement over *what is fake news* would make the line between content moderation and censorship blurrier.

The usage of the term fake news to refer to a narrative one disagrees with does not happen in a vacuum, but in a world stage increasingly marked by political polarization [Cunha et al., 2018]. In the US, between 2004 and 2014, the number of republicans that viewed the democratic party very unfavourably doubled (from 21% to 43%) [Mitchel et al., 2014]. In Europe, national and European parliament seats have become more polarized than ever, with a rising number of seats being held by far-right parties and a decreasing number seats being held by center parties [Groskopf, 2016]. In Brazil, the 2018 election was marked by extreme political polarization, raising concerns about the country's future [Brooks and Boadle, 2018]. In such scenarios, a troubling possibility is that individuals start perceiving any opposing narrative as extremely biased, and thus, fake [Habgood-Coote, 2017, Funke, 2018]. This has the po-
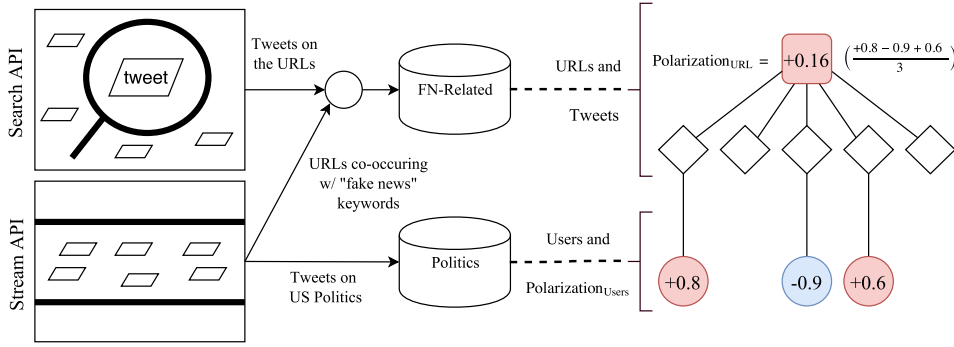
Figure 2.2. Methodology to collect the: *(i)* URLs/tweets flagged as fake news; *(ii)* general tweets that tweeted this URL/tweets; and *(iii)* general tweets on politics, and then build a dataset that encompasses the polarized reactions of users to a URL. We also exemplify how the calculation of the polarization of the URL is performed on the right-hand side. This is discussed in detail in Section 2.2.3

tential to poison public debate and lead to the adoption of a decision-making process grounded only in ideology and not in reality. Interestingly, it is important to notice that this increase in polarization has often been associated with Online Social Networks [Garimella et al., 2017]. The possibilities given by these platforms for users to filter only content they agree with, as well as algorithms which more often recommend content which gets engagement from users, have been accused of creating "filter bubbles": tailored information environments where users only find information they agree with [Pariser, 2011].

2.2   Methods

We describe our methodology to collect the data; estimate the political polarization of users, URLs, and Tweets; and how we analyse domains and impactful URLs.

2.2.1   Data Collection

Our data collection strategy is shown in Figure 2.2. We study two datasets in conjunction, both obtained from *Twitter*. The first dataset, `FN-Related`, was built to monitor narratives and discussions surrounding fake news. While the second, `Politics`, is used to infer the political orientation (Republican vs. Democrats) of the users in the first dataset. Notice that we collect tweets from all over the world in all cases, as filtering only geo-tagged tweets would greatly diminish the amount of data available.

To collect the `FN-Related` dataset, we performed, from *May 07 2017* to *May 25 2017*, two simultaneous data collection efforts, using the Stream API (which allows you to gather large amounts of data currently being tweeted) and the Search API (which allows you to search for tweets mentioning specific keywords, among other parameters), according to the following steps:

(Step 1) We collect tweets with the following keywords/hashtags from the Stream API:

> *{fakenews, #fakenews, fake-news, #fake-news, posttruth, #posttruth, post-truth, #post-truth, alternativefact, #alternativefact, alternative-fact, #alternative-fact}*

We then proceed to store the URLs being mentioned, whether it is an external URL or a URL to another tweet. In the examples below, we store external URL in the first case and the tweet being commented in the second case:

URL `Trump Schools CNN Reporter in 1990 {URL} #fakenews`

Tweet `(RT) This is an abuse of his office #posttruth {Tweet}`

(Step 2) In the second step, we use the stored external URLs and tweets. We use Twitter's Search API to extract tweets that include relevant URLs and tweets stored, and metadata about the users who tweeted about them. Unlike the Stream API, these tweets may have been tweeted a couple of hours or even days ago. For instance, in the aforementioned examples, we could find other users who tweeted the same link of Trump talking to the CNN reporter, or other people who had made a comment on the tweet quoted in the second example. Notice that this allows us to capture the URLs being used in different contexts, as exemplified by the two tweets below, which had the same URL. One of them simply shares a poll from The Huffington Post, whereas the other one suggests that the poll is untrustworthy:

Fake `HuffPost sucks, no one believes their polls #fakenews #MAGA {URL}`

Other `Canadian views of US hit an all-time low, poll shows, {URL}`

This data collection process allows us to build a more complete view of the fake news debate on Twitter: we can see both users who are referring to a piece of content (i.e. a URL or another tweet) as a potential source of fake news, and users who are citing, propagating or interacting with the same content *without* attaching to it the fake news label.

We obtained the second dataset, `Politics`, by collecting tweets about US Politics in general from the Twitter Stream API, from *August 2016* to *May 2017*. This is what is referred to as the third step in the Figure 2.2. We use keywords and hashtags such as {`Hillary Clinton, #potus, Donald Trump, White House, Democrats, Republicans...`}. The reason for collecting this dataset was to have sufficient data to accurately compute the degree of polarization of users from the `FN-related` dataset with respect to their leanings towards Republicans and Democrats. This is explained in detail in Section 2.2.2.

Using these data sources, we are able to analyze URLs that co-occurred with #FakeNews tags, the associated reactions to this URL in the form of tweets, and the polarization of some of the users who tweeted it. This is depicted on the right-hand side of Figure 2.2.

Some remarks on the methodology are: *(i)* Retweets and quoted tweets (a retweet with some additional text) are considered to be URLs (to other tweets). *(ii)* Every 15 minutes we update the URLs being searched in *Step 2* with the search according to the more recent results of the Stream API in *Step 1*. This is done due to limitations in using Twitter Stream and Search APIs.

2.2.2   Estimating User's Political Polarization

We want to correlate the degree of polarization of Twitter users to each main side in US Politics —Republicans and Democrats with fake news-related tweets. There are a plethora of methods designed to classify the political leaning of social media users, which typically group themselves in well-separated communities [Conover et al., 2011, Wong et al., 2016]. Finding communities on polarized topics is eased by the fact that it is usually simple to find seeds —users that are known to belong to a specific community. In the case of Twitter data, we consider that official profiles of politicians and political parties are natural seeds that can be fed to a semi-supervised clustering algorithm that expands the seeds to the communities formed around them [Calais Guerra et al., 2011, Kloumann and Kleinberg, 2014].

We assume that the number of communities K formed around a topic T is known in advance and it is a parameter of our method. To estimate user leanings toward each of the K groups (K=2 for Democrats, Republicans), we employ a label propagation-like strategy based on random walk with restarts [Tong et al., 2008]: a random walker departs from each seed and travels in the user-message retweet bipartite graph by randomly choosing an edge to decide which node it should go next. With a probability $(1 - \alpha) = 0.85$, the random walker restarts the random walking process from its original

seed. As a consequence, the random walker tends to spend more time inside the cluster its seed belongs to [Calais Guerra et al., 2011]. Each node is then assigned to its closest seed (i.e., community), as shown in the node colors in the sample of the graph displayed in Figure 2.1.

The relative proximity of each node to the two sets of seeds yields a probability that this node belongs to each of two communities, and can be interpreted as an estimate of his or her political leaning. For instance, if the proximity of node X to Republican seeds is 0.01 and its proximity to Democrat seeds is 0.04, the random-walk based community detection algorithm outputs that node belongs to the Democrat community with 80% probability. Note that this model captures that some nodes may be more neutral than others. For more details on the random walk-based community detection algorithm, please refer to [Calais Guerra et al., 2011].

In our specific case study there are only two communities, thus we can define the polarization of a user u with an assigned polarization value $v_u \in [0.5, 1.0] \cup [-1.0, -0.5]$:

$$P_u = \begin{cases} 2(-v_u + 0.5) \text{ if } u \in D \\ 2(v_u - 0.5) \text{ if } u \in R \end{cases} \tag{2.1}$$

Where R and D are the polarized groups of Republicans and Democrats. Notice that we are simply changing the domain of the value assigned by the polarization algorithm to a more intuitive one ($[-1, 1]$). We can further define the absolute user polarization:

$$A_u = |P_u| \tag{2.2}$$

2.2.3 Estimating URL's Political Polarization

Another aspect of the data that needs to be modeled is the polarization of a URL. Intuitively, polarized URLs are those which receive engagement from users of only one side of the political spectrum. Thus, we define the degree of polarization of a URL based on the polarization of users that reacted to it. The polarization of a URL (k), given two polarized groups of users (R, D), is defined as:

$$P_k = \frac{1}{n} \sum_{u \in \mathbf{U}(k)}^{n} P_u \tag{2.3}$$

Source	General Statistics				Shared Users		Shared Act. Users	
	#users	#active users	#tweets	#urls	FN-R	Politics	FN-R	Politics
FN-R	374,191	101,031	833,962	109,397	-	29.22%	-	37,61%
Politics	4,164,604	247,435	246,103,385	-	2.62%	-	15.72%	-

Table 2.1. General characterization of the data sources. The intersection between the `Politics` dataset and `FN-Related` is important as we use it to characterize the polarization of the users, and consequently of the URLs in the `FN-Related` datasets.

This is depicted in the right-hand side of Figure 2.2. We also define the absolute URL polarization:

$$A_k = |P_k| \tag{2.4}$$

Remember that *(i)* we consider as URLs any links external to Twitter or links to other tweets; and *(ii)* reactions are tweets of all kinds that interact with the URL.

2.2.4 Domains and Impactful URLs

An important part of our analysis is trying to find evidence that users are employing fake news related terms to express disagreement, rather than a factual lack of veracity in content they tweet about. To do so, we analyze the URL domains mentioned by each polarized side in tweets associated with misinformation. We also qualitatively analyze the content of some of the URLs that generated the most significant reactions.

To generate the domains we extract the external URLs mentioned in the tweets and then calculate the political polarization of each domain exactly like we do for full URLs. The wordclouds are generated for all the external URLs with absolute polarization A_k bigger than 0.5, one for each respective polarized group. For the analysis of the content of the URLs which garnered the most reactions, we randomly select 75 of the top 150 URLs, including both URLs referring to other tweets as well as URLs to external websites. Of those, we have equally sized stratum where A_k belongs to intervals $[0, 0.32]$, $[0.33, 0.66]$ or $[0.67, 1]$. We then analyze the content of these top 75 URLs for insights on different ways the fake news themes may emerge.

2.3 Results

We begin by characterizing the two datasets in terms of tweets, URLs, and users – as depicted in Table 2.1. Recall that the analysis concerning URLs are all performed using the tweets of the dataset we call `FN-Related`; the analysis concerning the polarization
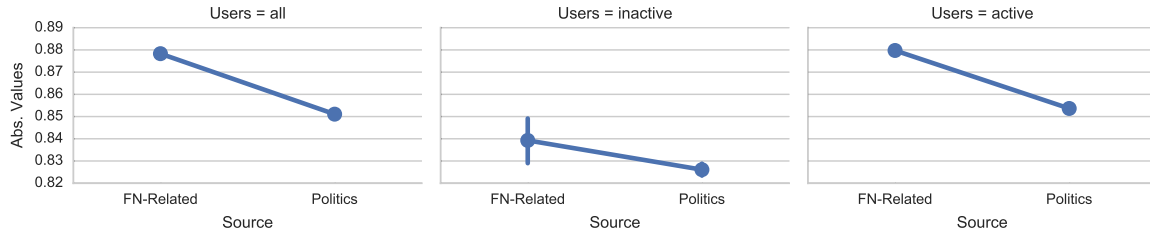
Figure 2.3. Average absolute user polarization for the users in the `FN-Related` and the `Politics` dataset. The error bars are the 95% confidence intervals calculated using bootstrap. The increase in the polarization in the `FN-Related` dataset suggests that the theme of misinformation increases polarization in an already polarized topic (politics).

of users is performed using the dataset named `Politics`. The dataset sizes differ significantly, but the overlap amongst them grants us a significant number of users with which to perform the analysis (29.22% of the 374, 191 users in the `FN-Related` dataset). If we define the active users in the `Politics` dataset as the smallest set of users responsible for 80% of the tweets collected, we have that the intersection with the users from the `FN-Related` dataset grows significantly, increasing to 15.72% from the original 2.62%. Notice that this overlap is important because these users present in both datasets are the ones used to calculate the polarization of URLs. These are the users we know the political orientation of.

2.3.1   Polarization and Tweets

We analyze the difference in polarization of the users in both datasets. Notice that in the `Politics` dataset we inspect the polarization of all users, whereas in the `FN-Related` dataset we only know the polarization of the users who are also present in the `Politics` dataset. Figure 2.3 shows the average polarization in such datasets, considering all users, active users, and inactive users. A significant increase in polarization is evident among the users who were associated with URLs that co-occurred with fake news related terms.

We were also interested in investigating the relationship between the polarization of URLs and the characteristics of the reactions associated with it. This analysis is performed only in the `FN-Related` dataset. We analyze two aspects: *(i)* the impact of the number of reactions surrounding a URL on its level of polarization, and *(ii)* the impact of the percentage of reactions which use fake news-related keywords and hashtags to the polarization of the URL. Ordering the URLs according to these metrics, we plot the average polarization of each one of its quartiles in Figure 2.4.

(i)   **(ii)**
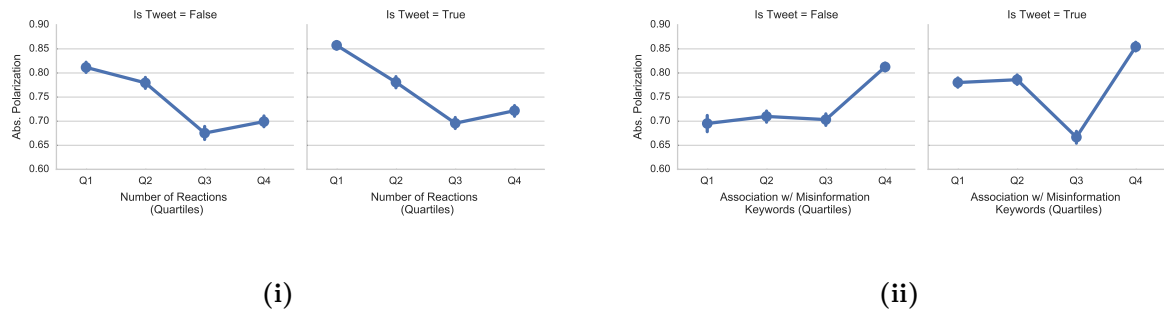
Figure 2.4. **(i)** Average polarization per number of reactions to a URL (quartiles). Error bars represent the 95% confidence interval. **(ii)** Average polarization per ratio of tweets with the URL containing the misinformation related keywords (quartiles).

(a) Democrat-leaning users.   *(b)* Republican-leaning users.
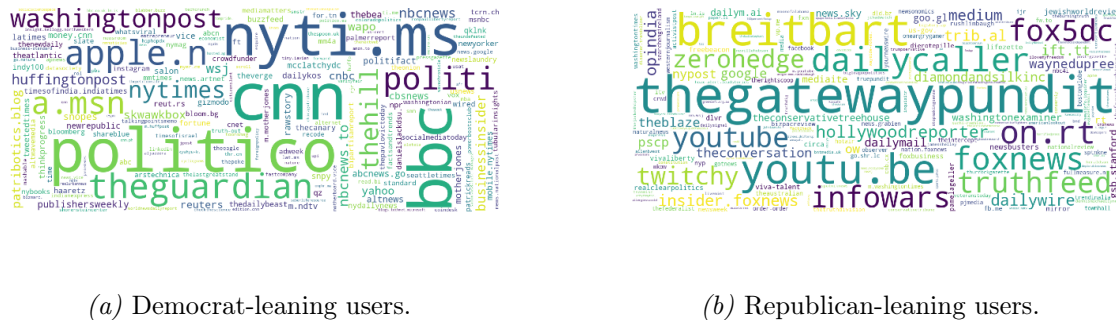
Figure 2.5. Domain clouds of #FakeNews-related tweets. Notice the presence of websites with the same ideology as the users in the polarized groups. This indicates that users are reacting to sources they agree with on fake-news related narratives.

We analyze other tweets and external URLs separately. Figure 2.4 (i) shows that the increase in the number of reactions has a negative impact on the average polarization of a URL, which suggests that more reactions mean less polarization. This is a nuanced finding, due to how we defined polarization. Notice that a URL with heated comments from both Republican and Democrat leaning users would be considered not very polarized here. A polarized URL, on the other hand, would be one where comments were predominantly from users of just one side of the political spectrum. In practice, this means that for popular URLs, you will have engagement from users of both sides of the political spectrum, while for less popular ones, mostly from one side. Figure 2.4 (ii) shows that polarization increases when URLs are associated with fake news-related keywords. This, along with Figure 2.3, contributes to the hypothesis that the fake news thematic is a polarizing one.

2.3.2   Polarization and URL Domains

We generate a wordcloud as described in Section 2.2.4, and the results can be seen in Figure 2.5*(a)* for Democrat-leaning users and in Figure 2.5*(b)* for Republican-leaning users. Recall that the users in the `FN-Related` dataset were put in the Democrat-Republican spectrum using the `Politics` dataset. Using different media sources determined to be trustworthy on both sides of the political spectrum according to the Pew Research Center [Mitchel et al., 2014], we can see that Democrat-leaning wordcloud contains domains related to news sources such as *The Washington Post* and *The New York Times*, which are trusted by liberals and distrusted by conservatives. Similarly, the republican-leaning wordcloud contains domains related to news sources such as *Breitbart* and *Fox News*, which are trusted by conservatives and distrusted by liberals.

The wordcloud analysis implies that polarized groups does not directly mention some reports or news stories as fake, but do react to links from sources that they agree with on the fake news theme. It also indicates that sources that members of a certain political ideology trust can significantly impact their view of what is fake, as they rely on their trusted sources, rather than pieces of information from outside sources that they believe to be misinformation or denounced as fake news.

2.3.3   Analyzing Top Reacted URLs

Analyzing the tweets which received the most reactions, and that co-occurred with fake news related keywords, allows us to better understand how such keywords are being used. We perform our qualitative analysis by giving and discussing examples of the different stratum we defined and inspected.

2.3.3.1   Opposing Narratives as #FakeNews

Among the randomly selected URLs, the top reacted external URL in the highly polarized stratum is a news-piece on Michael Flynn being cleared by the FBI as innocent from his relationship with Russia [Tacopino, 2017]:

> *New York Post:* FBI clears Michael Flynn in probe linking him to Russia

It is important to notice that Flynn's involvement with Donald Trump's campaign makes this piece of information more favorable to Republican-leaning users. Confirming the result obtained with the analysis of the wordclouds in Section 2.3.2, however, the result is polarized towards individuals supporting the Republican party. This suggests

that users are mainly dismissing a narrative of other media sources that suggested the link of Flynn with Russia. The terms associated with fake news are thus not being employed to denote that a piece of content itself is fake, but to denote other pieces of information as false.

2.3.3.2 Humor or absurdities as #FakeNews

Another usage of the term we can find analyzing the most reacted URLs is to refer to news which can be seen as ridicule. One of the most reacted URLs in the less polarized stratum $A_k \in [0, 0.32]$, is about a prisoner who attempted to escape jail dressed as a woman in Honduras [Boult, 2017]:

> *Telegraph:* Prisoner dressed as woman in failed escape bid

This usage, although not necessarily harmful for the political debate, may present a challenge for automated techniques to detect misinformation. This could be particularly troublesome if what users tag as fake is used directly as a feature by a model.

2.3.3.3 Misinformation as #FakeNews

We can also find instances of users tagging actual facts or stories as fake. An example of such a case is a highly polarized Democrat-leaning news-piece $A_k \in [0.67, 1]$ pointing the rebuttal of a supposedly fake news story on the murder of DNC staff:

> *Raw Story:* Family blasts right-wing media for spreading fake news story about slain DNC staffer as Russia scandal deepens

This usage is what is often imagined when we think about the conversational usage of the term fake news. Yet, it is important to understand that this is only one among many of the ways it is employed.

2.4 Discussion

This work is a first attempt to observe correlations between political polarization and the spread of misinformation, in particular fake news. To tackle the practical challenge of having access to a preclassified set of fake news articles or tweets, we monitored the external URLs and tweets associated with fake news related hashtags and keywords. We searched for tweets reacting to these URLs and calculated the polarization of the users who reacted to them using an auxiliary and more general dataset on politics. We

examined the association between polarization and fake news by analyzing the impact of various factors in fake news related URLs, and users we knew the polarization of. We also analyzed the different sources that are mentioned as "fake" by users and qualitatively described different scenarios where the terminology is applied.

Our analyses show three main findings:

- Fake news is associated with polarization. Users that engage with content labeled as fake news are more polarized and that content that is labeled more often as fake news by users receives engagement from users of mostly one side of the political spectrum.

- Users employ terms such as fake news to refer to content that they particularly disagree with. We find examples of such cases in our qualitative analysis of the tweets and the URLs, and observe that content labelled as fake news is very polarized.

- Polarized groups cite sources on their side of the political spectrum to tag or condemn news and statements given by the other opposite group as fake. This was shown by the analysis of the wordclouds of the Democrat and Republican-leaning users and the qualitative examination of the contexts where misinformation-related keywords and hashtags were employed.

These findings present challenges for automated ways of detecting fake news. Two popular ideas for doing so are crowd-sourced systems (e.g. [Ratkiewicz et al., 2011]) and machine learning models (e.g. [Conroy et al., 2015]). Both cases involve humans labelling news-pieces as either true or false: In the first case, to take a crowd-sourced decision about the news-piece, and, in the second, to label a dataset with which a machine learning model will be trained. Our findings suggest that the political orientation of these humans-in-the-loop may have a significant impact on what they are labelling as fake, which should be taken into consideration. One way this could be addressed in crowd-sourcing setups would be to ensure that a news-piece was annotated by users from different political orientations. In that sense, actual misinformation could be found in the intersection of what users from groups with conflicting interests consider to be fake. In the machine learning scenario, a possible workaround would be to consider this factor in the labelling process, ensuring that an ideologically diverse group of users looked at each news-piece in the training data.

More generally, this work also corroborates with the hypothesis that vague terminologies to denote misinformation are harmful. To avoid blurring the line between

what is perceived as fake and what is perceived as biased, governments and the media should try to adopt specific terminology to differentiate between different types of content, such as click-baits, biased news, hoaxes, and etc. Several of these typologies have been proposed by researchers [Tandoc Jr et al., 2018, Zannettou et al., 2018b, Kumar and Shah, 2018], and they could address the ambiguity of the term.

The major possibilities to expand this work are tied to its limitations. Firstly, although we obtain evidence that users employ terms such as fake news to refer to content that they disagree with, we do not calculate to which extent is that so. In that context, a key direction to expand our understanding of the interactions between these two phenomena is to measure to which extent people actually believe that facts they disagree with are fake. A way to do so would be finding statements that are proven false by fact-checkers and modeling the complex interactions (such as quotes and replies) between users in a social network such as *Twitter*.

Another interesting direction has to do with the causal relationship between fake news and political polarization. Although this work shows a correlation between the two phenomena, the way they interact is not so clear. If there is a decrease in political polarization, is there a decrease in fake news? How about the opposite scenario? Understanding these effects would enable new ways of fighting misinformation and political polarization.

Chapter 3

Hateful Users on Twitter

The importance of understanding hate speech in Online Social Networks (OSNs) is manifold. Countries such as Germany have strict legislation against the practice [Stein, 1986], the presence of such content may pose problems for advertisers [Solon, 2017] and users [Sabatini and Sarracino, 2017], and manually inspecting all possibly hateful content in OSNs is unfeasible [Schmidt and Wiegand, 2017]. Furthermore, the trade-off between banning such behavior from platforms and not censoring dissenting opinions is a major societal issue [Rainie et al., 2017].

This scenario has motivated work that aims to understand and detect hateful content [Greevy and Smeaton, 2004, Warner and Hirschberg, 2012, Burnap and Williams, 2016], by creating representations for text in OSNs, *e.g.* word2vec [Mikolov et al., 2013], and then classifying them as hateful or not, often drawing insights on the nature of hateful speech. However, in OSNs, the meaning of such content is often not self-contained, referring, for instance, to some event which just happened, and the texts are packed with informal language, spelling errors, special characters and sarcasm [Dhingra et al., 2016, Riloff et al., 2013]. Besides that, hate speech itself is highly subjective, reliant on temporal, social and historical context, and occurs sparsely [Schmidt and Wiegand, 2017]. These problems, although observed, remain unaddressed [Davidson et al., 2017, Magu et al., 2017]. Consider the tweet:

```
Timesup, yall getting w should have happened long ago
```

Which was in reply to another tweet that mentioned the holocaust. Although the tweet, whose author's profile contained white-supremacy imagery, incited violence, it is hard to conceive how this could be detected as hateful with only textual features. Furthermore, the lack of hate-related words makes it difficult for this kind of tweet to even be sampled in text-based approaches.
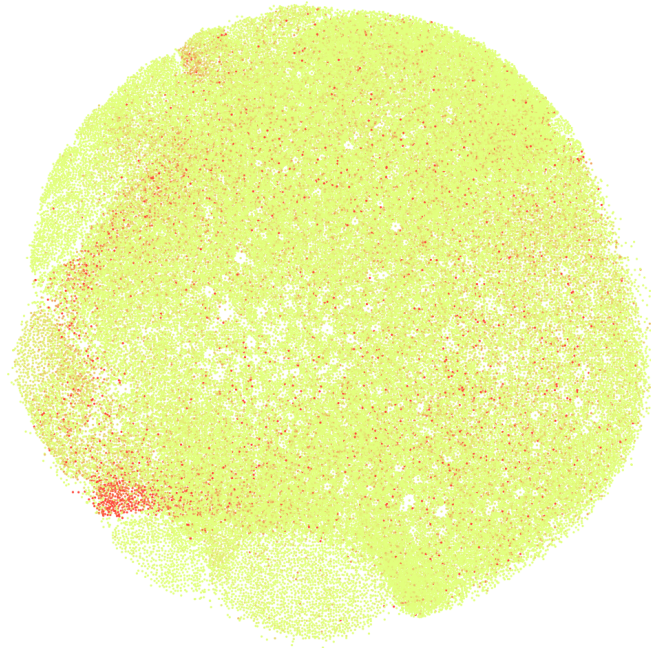
Figure 3.1. Network of $100,386$ users sampled from Twitter after our diffusion process. Red nodes indicate the proximity of users to those who employed words in our lexicon.

Fortunately, as we just hinted, the data in posts, tweets or messages are not the only signals we may use to study hate speech in OSNs. Most often, these signals are linked to a profile representing a person or organization. Characterizing and detecting hateful *users* shares much of the benefits of detecting hateful content and presents plenty of opportunities to explore a richer feature space. Furthermore, on a practical hate speech guideline enforcement process, containing humans in the loop, it is natural that user profiles will be checked, rather than isolated tweets. The case can be made that this wider context is sometimes *needed* to define hate speech, such as in the example, where the abuse was made clear by the neo-nazi signs in the user's profile. Analyzing hate on a *user-level* rather than *content-level* enables our characterization to explore not only content, but also dimensions such as the user's activity and connections. Moreover, it allows us to use the very structure of Twitter's network in the task of detecting hateful users [Hamilton et al., 2017b].

In this chapter, we characterize and detect hateful *users* on Twitter, which we define according to Twitter's hateful conduct guidelines. We collect a dataset of $100,386$ users along with up to 200 tweets for each with a random-walk-based crawler on Twitter's retweet graph. We identify users that employed words from a set of hate speech related lexicon and generate a subsample, selecting users that are in different "distances"

to those. The latter are manually annotated as hateful or not through crowdsourcing. The aforementioned distances are real valued numbers obtained through a diffusion process in which the users who used the words in the lexicon are the seeds. With this, we create a dataset containing $4,972$ manually annotated users, of which 544 were labeled as hateful [1]. We also find the users that have been suspended after the data collection —before and after Twitter's 18/Dec/17 guideline changes.

Studying these users, we find differences between the activity patterns of hateful and normal users: hateful users tweet more frequently, follow more people each day and their accounts are more short-lived and recent While the media stereotypes hateful individuals as "lone wolves" [Burke, 2017], we find that hateful users are not in the periphery of the retweet network we sampled. Although they have fewer followers, the median for several network centrality measures in the retweet network is higher for those users. We also find that these users do not seem to behave like spammers.

A lexical analysis using *Empath* [Fast et al., 2016] shows that their choice of vocabulary is particular: words related to hate, anger and politics occur *less* often when compared to their normal counterparts, and words related to masculinity, love and curses occur more often. This is noteworthy, as much of the previous work directly employs hate-related words as a data-collection mechanism.

We compare the neighborhood of hateful with the neighborhood of normal users in the retweet graph, as well as accounts that have been suspended with those who were not. We argue that these suspended accounts and accounts that retweeted hateful users are also proxies for hateful speech online, and the similar results found in many of the analyses performed increase the robustness of our findings.

We also compare users who have been banned before and after Twitter's recent guideline change, finding an increase in the number of users banned per day, but little difference in terms of their vocabulary, activity and network structure.

Finally, we find that hateful users and suspended users are very densely connected in the retweet network we sampled. Hateful users are 71 times more likely to retweet other hateful users and suspended users are 11 times more likely to retweet other suspended users. This motivates us to pose the problem of detecting hate speech as a task of supervised learning over graphs. We employ a node embedding algorithm that creates a low-dimensional representation of nodes in a network to then classify it. We demonstrate robust performance to detect both hateful and suspended users in such fashion (95% AUC and 93% AUC) and show that this approach outperforms traditional state-of-the-art classifiers (88% AUC and 88% AUC, respectively).

---

[1] We used Crowdflower' crowdsourcing service to annotate users and Twitter's hateful conduct guidelines as a criteria for what we consider to be a hateful user.

3.1   Background

3.1.1   Hateful Users

We define "hateful user" and "hate speech" according to Twitter's guidelines. For the purposes of this chapter, "hate speech" is any type of content that 'promotes violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease" [Twitter, 2019]. On the other hand, "hateful user" is a user that, according to annotators, endorses such type of content.

3.1.2   Retweet Graph

The retweet graph G is a directed graph $G = (V, E)$ where each node $u \in V$ represents a user in Twitter, and each edge $(u_1, u_2) \in E$ implies that the user u_1 has retweeted u_2. Previous work suggests that retweets are better than followers to judge users' influence [Cha et al., 2010]. As influence flows in the opposite direction of retweets, we invert the graph's edges.

3.1.3   Offensive Language

We employ [Waseem et al., 2017] definition of explicit abusive language, which defines it as *language that is unambiguous in its potential to be abusive, for example, language that contains racial or homophobic slurs.* The use of this kind of language doesn't imply hate speech, although there is a clear correlation [Davidson et al., 2017].

3.1.4   Suspended Accounts

Most Twitter accounts are suspended due to spam, however, they are harder to reach in the retweet graph as they rarely get retweeted. We use Twitter's API to find the accounts that have been suspended among the $100,386$ collected users, and use these as another source for potentially hateful behavior. We collect accounts that have been suspended two months after the data collection, on 12/Dec/2017, and after Twitter's hateful conduct guideline changes, on 14/Jan/2018. The new guidelines are allegedly stricter, considering, for instance, off-the-platform behavior. Importantly, not all accounts are suspended due to hateful conduct. Yet, it is worth noticing that our analyses show that our accounts not present spammer-like behavior, a common reason for suspension.

3.2   Methods

3.2.1   Data Collection

Most previous work on detecting hate speech on Twitter employs a lexicon-based data collection, which involves sampling tweets that contain specific words [Davidson et al., 2017, Waseem and Hovy, 2016], such as `wetb*cks` or `fagg*t`. However, this methodology is biased towards a very direct, textual and offensive hate speech.   It presents difficulties with statements that subtly disseminate hate with no offensive words, as in `"Who convinced Muslim girls they were pretty?"` [Davidson et al., 2017]; And also with the usage of code words, as in the use of the word `"skypes"`, employed to reference jews [Magu et al., 2017, Know Your Meme, 2018]; In this scenario, we propose collecting users rather than tweets, relying on lexicon only *indirectly*, and collecting the structure of these users in the social network, which we will later use to characterize and detect hate.

We represent the connections among users in Twitter using the retweet network [Cha et al., 2010]. Sampling the retweet network is hard as we can only observe out-coming edges (due to API limitations), and as it is known that any unbiased in-degree estimation is impossible without sampling most of these "hidden" edges in the graph [Ribeiro et al., 2012]. Acknowledging this limitation, we employ Ribeiro et al. Direct Unbiased Random Walk algorithm, which estimates out-degrees distribution efficiently by performing random jumps in an undirected graph it constructs online [Ribeiro et al., 2010]. Fortunately, in the retweet graph the outcoming edges of each user represent the other users she —usually [Guerra et al., 2017]— endorses. With this strategy, we collect a sample of Twitter retweet graph with $100,386$ users and $2,286,592$ retweet edges along with the 200 most recent tweets for each users, as shown in Figure 3.1. This graph is unbiased w.r.t. the out degree distribution of nodes.

Notice that this graph is not biased in any way towards hateful users. It is just a sample of Twitter's retweet graph with a nice property (out-degree distribution is preserved).

3.2.2   Choosing the Subsample to Annotate

As the sampled graph is too large to be annotated entirely, we need to select a subsample to be annotated. If we choose tweets uniformly at random, we risk having a very insignificant percentage of hate speech in the subsample. On the other hand, if we choose only tweets that use obvious hate speech features, such as offensive racial
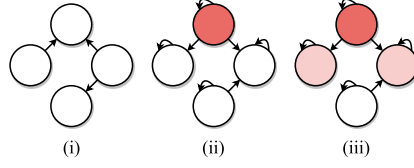
Figure 3.2. Toy examplé of the diffusion process. *(i)* We begin with the sampled retweet graph G textit(ii) We revert the direction of the edges (the way influence flows), add self loops to every node, and mark the users who employed words in our lexicon; *(iii)* We iteratively update the belief of other nodes.

slurs, we will stumble in the same problems pointed in previous work. We propose a method between these two extremes. We:

1. Create a lexicon of words that are mostly used in the context of hate speech. This is unlike other work [Davidson et al., 2017] as we do not consider words that are employed in a hateful context but often used in other contexts in a harmless way (*e.g.* `n*gger`); We use 23 words such as `holohoax`, `racial treason` and `white genocide`, handpicked from Hatebase.org [Hatebase, 2018], and ADL's hate symbol database [ADL, 2018].

2. Run a diffusion process on the graph based on DeGroot's Learning Model [Golub and Jackson, 2010], assigning an initial belief $p_i^0 = 1$ to each user u_i who employed the at least one of the words in the lexicon; This prevents our sample from being excessively small or biased towards some vocabulary.

3. Divide users into 4 strata according to their associated beliefs after the diffusion process, and perform a stratified sampling, obtaining up to 1500 user per strata.

We briefly present our diffusion model, as illustrated in Figure 3.2. Let A be the adjacency matrix of our retweeted graph $G = (V, E)$ where each node $u \in V$ represents a user and each edge $(u, v) \in E$ represents a retweet. We have that $A[u, v] = 1$ if u retweeted v. We create a transition matrix T by inverting the edges in A (as influence flows from the retweeted user to the user who retweeted him or her), adding a self loop to each of the nodes and then normalizing each row in A so it sums to 1. This means each user is equally influenced by every user he or she retweets.

We then associate a belief $p_i^0 = 1$ to every user who employed one of the words in our lexicon, and $p_i^0 = 0$ to all who did not. Lastly, we create new beliefs $\mathbf{p}^t$ using the update rule: $\mathbf{p}^t = T\mathbf{p}^{t-1}$. All the beliefs p_i^t converge to the same value as $t \to \infty$, thus we run the diffusion process with $t = 2$. With this real value ($p_i^2 \in [0, 1]$) associated with each user, we get 4 strata by randomly selecting up to 1500 users with p_i in the

intervals $[0, .25)$, $[.25, .50)$, $[.50, .75)$ and $[.75, 1]$. This ensures that we annotate users that did not employ any of the words in our lexicon, yet have a high potential to be hateful due to homophily.

3.2.3 Annotating the Users

We annotate $4,972$ users as hateful or not using *CrowdFlower*, a crowdsourcing service. The annotators were given the definition of hateful conduct according to Twitter's guidelines and asked, for each user:

> *Does this account endorse content that is humiliating, derogatory or insulting towards some group of individuals (gender, religion, race) or support narratives associated with hate groups (white genocide, holocaust denial, Jewish conspiracy, racial superiority)?*

Annotators were asked to consider the entire profile (limiting the tweets to the ones collected) rather than individual publications or isolate words and were given examples of terms and codewords in ADL's hate symbol database. Each user profile was independently annotated by 3 annotators, and, if there was disagreement, up to 5 annotators. In the end, 544 hateful users and $4,427$ normal ones were identified by them. The sample of the retweet network was collected between the 1st and 7th of Oct/17, and annotation began immediately after. We also obtained all users suspended up to 12/Dec/17 (387) and up to 14/Jan/18 (668).

3.3 Results

We analyze how hateful and normal users differ w.r.t. their activity, vocabulary and network centrality. We also compare the neighbors of hateful and of normal users, and suspended/active users to reinforce our findings, as homophily suggests that the neighbors will share a lot of characteristics with annotated users, and as suspended users may have been banned because of hateful conduct. We compare those in pairs as the sampling mechanism for each of the populations is different. We argue that each one of these pairs contains a proxy for hateful speech in Twitter, and thus inspecting

Table 3.1. Number of users in each group.

Hateful	Normal	Hateful Neigh.	Normal Neigh.	Banned	Active
544	4427	3471	33564	668	99718
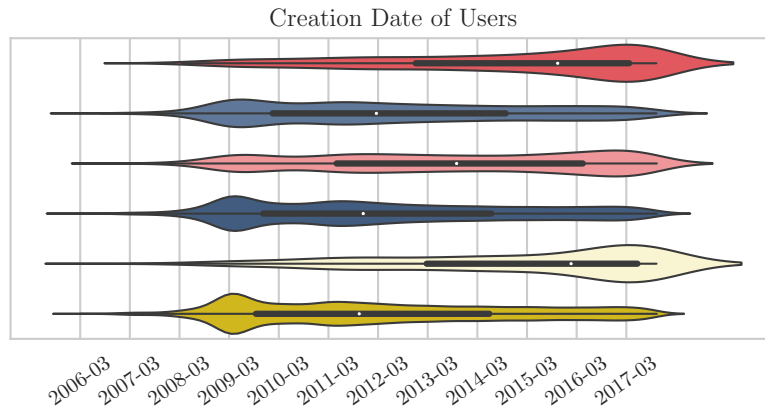
Figure 3.3. KDEs of the creation dates of user accounts. The white dot indicates the median and the thicker bar indicates the first and third quartiles.
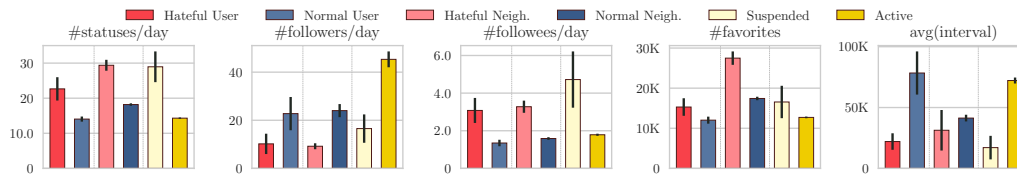
Figure 3.4. Average values for several activity-related statistics for hateful users, normal users, users in the neighborhood of those, and suspended/active users. The `avg(interval)` was calculated on the 200 tweets extracted for each user. Error bars represent 95% confidence intervals. The legend used in this graph is kept in the remainder of the chapter.

the three increases the robustness of our analysis. P-values given are from unequal variances t-tests to compare the averages across distinct populations. When we refer to "hateful users", we refer to the ones annotated as hateful. The number of users in each of these groups is given in the table bellow:

3.3.1 Activity

The account creation date of users is depicted in Figure 3.3. Hateful users were created later than normal ones (p-value < 0.001). A hypothesis for this difference is that hateful users are banned more often due to Twitter's guidelines infringement. This resonates with existing methods for detecting fake accounts in which using the account's creation date have been successful [Viswanath et al., 2015]. We obtain similar results w.r.t. the 1-neighborhood of such users, where the hateful neighbors were also created more recently (p-value < 0.001), and also when comparing suspended and active accounts

(p-value < 0.001).

Other interesting metrics for analysis are the number of tweets, followers, followees and favorite tweets a user has, and the interval in seconds between their tweets. We show these statistics in Figure 3.4. We normalize the number of tweets, followers and followees by the number of days the users have since their account creation date. Our results suggest that hateful users are "power users" in the sense that they tweet more, in shorter intervals, favorite more tweets by other people and follow other users more (p-values < 0.01). The analysis yields similar results when we compare the 1-neighborhood of hateful and normal users, and when comparing suspended and active accounts (p-values < 0.01, except for the number of favorites when comparing suspended/active users, and for the average interval, when comparing the neighborhood).

We investigate whether users that propagate hate speech are spammers. We analyze metrics that have been used by previous work to detect spammers, such as the numbers of URLs per tweet, of hashtags per tweet and of followers per followees [Benevenuto et al., 2010]. The boxplot of these distributions is shown on Figure 3.5. We find that hateful users use, in average, *less* hashtags (p-value < 0.001) and *less* URLs (p-value < 0.001) per tweet than normal users. The same analysis holds if we compare the 1-neighborhood of hateful and non-hateful, or suspended and active users (with p-values < 0.05, except for the number of followers per followees, where there is no statistical significance to the t-test). Additionally, we also find that, in average, normal users have more followers per followees than hateful ones (p-value < 0.05), which also happens for their neighborhood (p-value < 0.05). This suggests that the hateful and suspended users do not use systematic and programmatic methods to deliver their content. Notice that it is not possible to extrapolate this finding to Twitter in general, as there maybe be hateful users with other behaviors which our data collection methodology does not consider, as we do not specifically look for trending topics or popular hashtags.

3.3.2 Centrality

We analyze different measures of centrality for users, as depicted in Figure 3.6. The median hateful user is more central in all measures when compared to their normal counterparts. This is a counter-intuitive finding, as hateful crimes have long been associated with "lone wolves", and anti-social people [Burke, 2017]. We observe similar results when comparing the median eigenvector centrality of the neighbors of hateful and normal users, as well as suspended and active users. In the latter pair, suspended users also have higher median out degree. When analyzing the average for such statis-
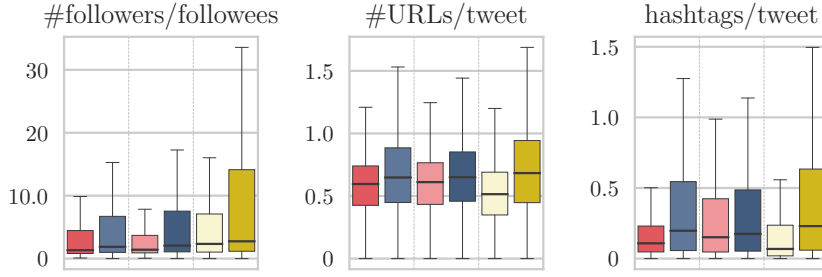
Figure 3.5. Boxplots for the distribution of metrics that indicate spammers. Hateful users have slightly *less* followers per followee, *less* URLs per tweet, and *less* hashtags per tweet.
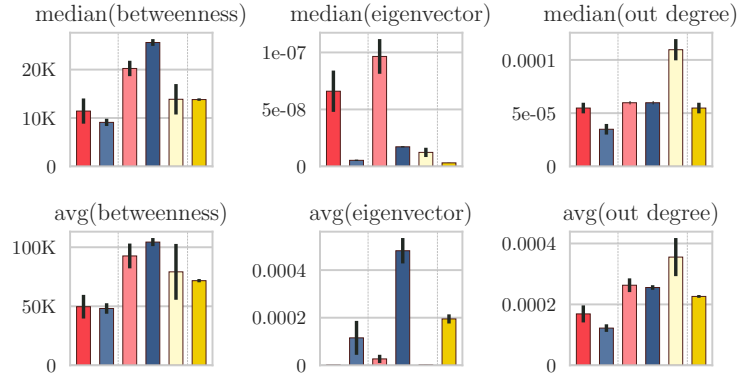
Figure 3.6. Network centrality metrics for hateful and normal users, their neighborhood, and suspended/non-suspended users calculated on the sampled graph.

tics, we observe the average eigenvector centrality is higher for the opposite sides of the previous comparisons. This happens because some very influential users distort the value: for example, the 970 most central users according to the metric are normal. Notice that despite of this, hateful and suspended users have higher average out degree than normal and active users respectively (p-value < 0.05).

3.3.3   Lexicon

We characterize users w.r.t. their content with *Empath* [Fast et al., 2016], as depicted in Figure 3.7. Hateful users use *less* words related to hate, anger, shame and terrorism, violence, and sadness when compared to normal users (with p-values < 0.001). A question this raises is how sampling tweets based exclusively in a hate-related lexicon biases the sample of content to be annotated to a very specific type of "hate-spreading" user, and reinforces the claims that sarcasm, code-words and very specific slang plays
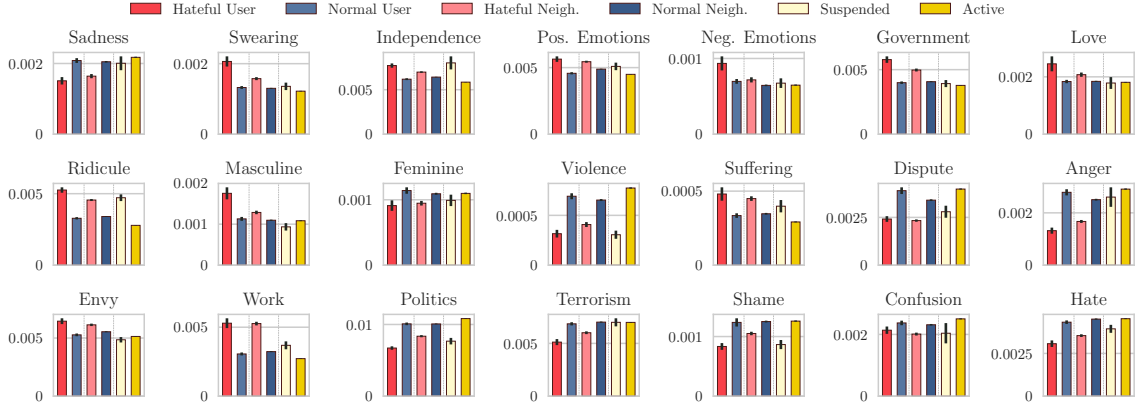
Figure 3.7. Average values for the relative occurrence of several categories in *Empath*. Notice that not all Empath categories were analyzed and that the to-be-analyzed categories were chosen before-hand to avoid spurious correlations. Error bars represent 95% confidence intervals.
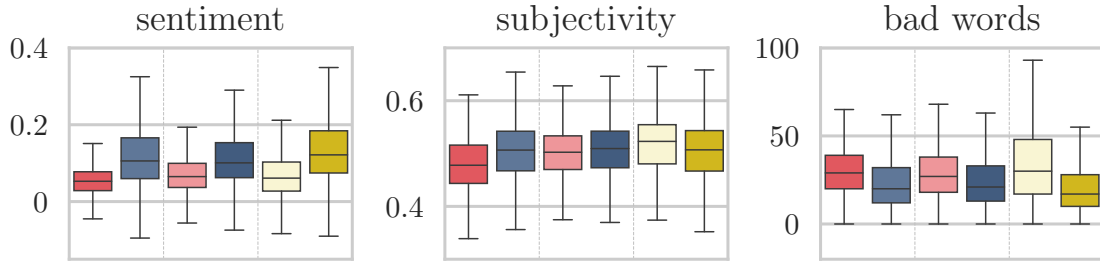
Figure 3.8. Boxplots for the distribution of sentiment and subjectivity and bad-words usage. Suspended users, hateful users and their neighborhood are more negative, and use more bad words than their counterparts.

a significant role in defining such users [Davidson et al., 2017, Magu et al., 2017].

Categories of words more used by hateful users include positive emotions, negative emotions, suffering, work, love and swearing (with p-values < 0.001), suggesting the use of emotional vocabulary. An interesting direction would be to analyze the sensationalism of their statements, as it has been done in the context of *click-baits* [Chen et al., 2015]. When we compare the neighborhood of hateful and normal users and suspended vs active users, we obtain very similar results (with p-values < 0.001 except for when comparing suspended vs. active users usage of anger, terrorism, sadness, swearing and love). Overall, the non-triviality of the vocabulary of these groups of users reinforces the difficulties found in the NLP approaches to sample, annotate and detect hate speech [Davidson et al., 2017, Magu et al., 2017].

We also explore the sentiment in the tweets users write using a corpus based

Table 3.2.   Occurrence of the edges between hateful (red) and normal (blue) users, and between suspended (lemon) and active (dark yellow)users. Results are normalized w.r.t. to the type of the source node, as in: P(source type→dest type|source type). Notice that the probabilities do not add to 1 in hateful and normal users as we don't present the statistics for non-annotated users.

Node Type	(%)	Node Type	(%)
🔴 → 🔴	41.50	🔴 → 🔵	13.10
🔵 → 🔵	15.90	🔵 → 🔴	2.86
⚪ → ⚪	7.50	⚪ → 🟡	92.50
🟡 → 🟡	99.35	🟡 → ⚪	0.65

approach, as depicted in Figure 3.8. We find that sentences written by hateful and suspended users are more negative, and are less subjective (p-value < 0.001). The neighbors of hateful users in the retweet graph are also more negative (p-value < 0.001), however not less subjective. We also analyze the distribution of profanity per tweet in hateful and non-hateful users. The latter is obtained by matching all words in Shutterstock's "List of Dirty, Naughty, Obscene, and Otherwise Bad Words" [2]. We find that suspended users, hateful users and their neighbors employ more profane words per tweet, also confirming the results from the analysis with *Empath* (p-value < 0.01).

3.3.4   Connections

Finally, we analyze the frequency at which hateful and normal users, as well as suspended and active users, interact within their own group and with each other. Table 3.2 depicts the probability of a node of a given type retweeting other types of node. We find that 41% of the retweets of hateful users are to other hateful users, which means that they are 71 times more likely to retweet another hateful user, considering the occurrence of hateful users in the graph. We observe a similar phenomenon with suspended users, which have 7% of their retweets redirected towards other suspended users. As suspended users correspond to only 0.68% of the users sampled, this means they are approximately 11 times more likely to retweet other suspended users. The high density of connections among hateful and suspended users suggest strong modularity. We exploit this, along with activity and network centrality attributes to robustly detect these users.

---

[2]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Table 3.3. Percentage/number of accounts that got suspended up before and after the guidelines changed. Notice that accounts may be suspended for reasons other than hateful conduct.

Susp. Accounts	Hateful	Normal	Others
2017-12-12	9.09%/55	0.32%/14	0.33%/318
2018-01-14	17.64%/96	0.90%/40	0.55%/532

End of Data Collection (07/Oct)   Twitter's Guideline Changes (18/Dec)

~6 bans/day   ~9 bans/day

387 banned (12/Dec)   677 banned (14/Jan)
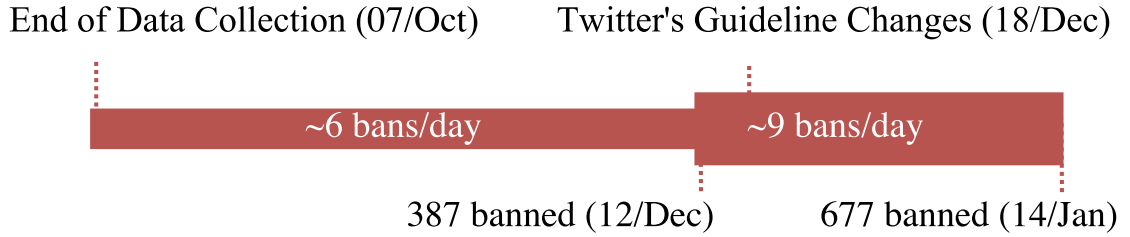
Figure 3.9. Corhort-like depiction of the banning of users. We find that in the period after Twitter's guideline change the number of bans a day increased 1.5 times, from 6 to 9.

3.3.5   Suspension of Users

Twitter has changed its enforcement of hateful conduct guidelines in 18/Dec/2017. We analyze the differences among accounts that have been suspended two months after the end of the annotation, in 12/Dec/2017 and in 14/Jan/2018.

The intersection between these groups and the ones we annotated as hateful or not is shown in Table 3.3. In the first period from the end of the data annotation to the 12/Dec, there were approximately 6.45 banned users a day whereas in the second period there were 9.05. This trend, illustrated in Figure 3.9, suggests an increased banning activity.

Performing the lexical analysis we previously applied to compare hateful and normal users we do not find statistically significant difference w.r.t. the averages for users banned before and after the guideline change (except for government-related words, where p-value < 0.05). We also analyze the number of tweets, followers/followees, and the previously mentioned centrality measures, and observe no statistical significance in the difference between the averages or the distributions (which were compared using KS-test). This suggests that Twitter has not changed the type of users banned.

Table 3.4. Prediction results and standard deviations for the two proposed settings: detecting hateful users and detecting suspended users. The semi-supervised node embedding approach performs better than state-of-the-art supervised learning algorithms in all the assessed criteria, suggesting the benefits of exploiting the network structure to detect hateful and suspended users.

Model	Features	Hateful/Normal			Suspended/Active		
		Accuracy	F1-Score	AUC	Accuracy	F1-Score	AUC
GradBoost	user+glove	84.6 ± 1.0	52.0 ± 2.2	88.4 ± 1.3	81.5 ± 0.6	48.4 ± 1.1	88.6 ± 0.1
	glove	84.4 ± 0.5	52.0 ± 1.3	88.4 ± 1.3	78.9 ± 0.7	44.8 ± 0.7	87.0 ± 0.5
AdaBoost	user+glove	69.1 ± 2.4	37.6 ± 2.4	85.5 ± 1.4	70.1 ± 0.1	38.3 ± 0.9	84.3 ± 0.5
	glove	69.1 ± 2.5	37.6 ± 2.4	85.5 ± 1.4	69.7 ± 1.0	37.5 ± 0.8	82.7 ± 0.1
GraphSage	user+glove	$\mathbf{90.9 \pm 1.1}$	$\mathbf{67.0 \pm 4.1}$	$\mathbf{95.4 \pm 0.2}$	$\mathbf{84.8 \pm 0.3}$	$\mathbf{55.8 \pm 4.0}$	$\mathbf{93.3 \pm 1.4}$
	glove	90.3 ± 1.9	65.9 ± 6.2	94.9 ± 2.6	84.5 ± 1.0	54.8 ± 1.6	93.3 ± 1.5

3.3.6   Prediction

As we consider users and their connections in the network, we can use information that is not available for models which operate on the granularity level of tweets or comments to detect hate speech.

- **Activity/Network:** Features such as number of statuses, followers, followees, favorites, and centrality measurements such as betweenness, eigenvector centrality and the in/out degree of each node. We refer to these as `user`.

- **GloVe:** We also use spaCy's off-the-shelf 300-dimensional GloVe's vector [Pennington et al., 2014] as features. We average the representation across all words in a given tweet, and subsequently, across all tweets a user has. We refer to these as `glove`.

Using these features, we compare experimentally two traditional machine learning models known to perform very well when the number of instances is not very large: Gradient Boosted Trees (`GradBoost`) and Adaptive Boosting (`AdaBoost`); and a model aimed specifically at learning in graphs [Hamilton et al., 2017a] (`GraphSage`). Interestingly, the latter approach is semi-supervised, and allows us to use the neighborhood of the users we are classifying even though they are not labeled, exploiting the modularity between hateful and suspended users we observed. The algorithm creates low-dimensional embeddings for nodes, given associated features (unlike other node embeddings, such as `node2vec` [Grover and Leskovec, 2016]). Moreover, it is inductive - which means we don't need the entire graph to run it. For additional information on node embeddings methods, refer to [Hamilton et al., 2017b].

The GraphSage algorithm creates embeddings for each node given that the nodes have associated features (in our case the *GloVe* embeddings and activity/network-centrality attributes associated with each user). Instead of generating embeddings for all nodes, it learns a function that generate embeddings by sampling and aggregating features from a node's local neighborhood. This strategy exploits the structure of the graph beyond merely using the features of the neighborhood of a given node.

We run the algorithms trying to detect both hateful and normal users, as annotated by the crowdsourcing service, as well as trying to detect which users got banned. We perform a 5-fold cross validation and report the F1-score, the accuracy and the area under the ROC curve (AUC) for all instances.

In all approaches we accounted for the class imbalance (of approximately 1 to 10) in the loss function. We keep the same ratio of positive/negative classes in both tasks, which, in practice, means we used the 4981 annotated users in the first setting (where approximately 11% were hateful) and, in the second setting, selected 6680 users from the graph, including the 668 suspended users, and other 5405 users randomly sampled from the graph.

Notice that, as we are dealing with a binary classification problem, we may control the trade-off between specificity and sensitivity by varying the positive-class threshold. In this work we simply pick the largest value, and report the resulting AUC score —which can be interpreted as the probability of a classifier correctly ranking a random positive case higher than a random negative case.

The results of our experiments are shown in Table 3.4. We find that the node embedding approach using the features related to both users and the *GloVe* embeddings yields the best results for all metrics in the two considered scenarios. The Adaptative Boosting approach yields good AUC scores, but incorrectly classifies many normal users as hateful, which results in a low accuracy and F1-score.

Using the features related to users makes little difference in many settings, yielding, for example, exactly the same AUC, and very similar accuracy/F1-score in the Gradient Boosting models trained with the two sets of parameters. However, the usage of the retweet network (in GraphSage) yields promising results, especially because we observe improvements in both the detection of hateful users and of suspended users, which shows the performance improvement occurs independently of our annotation process.

3.4   Discussion

We present an approach to characterize and detect hate speech on Twitter at a user-level granularity. Our methodology differs from previous efforts, which focused on isolated pieces of content, such as tweets and comments. [Greevy and Smeaton, 2004, Warner and Hirschberg, 2012, Burnap and Williams, 2016]. We developed a methodology to sample Twitter which consists of obtaining a generic subgraph, finding users who employed words in a lexicon of hate-related words and running a diffusion process based on DeGroot's learning model to sample for users in the neighborhood of these users. We then used *Crowdflower*, a crowdsourcing service to manually annotate 4,988 users, of which 544 (11%) were considered to be hateful. We argue that this methodology aids two existing shortcomings of existing work: it allows the researcher to balance between having a generic sample and a sample biased towards a set of words in a lexicon, and it provides annotators with realistic context, which is sometimes necessary to identify hateful speech.

Our findings shed light on how hateful users differ from normal ones w.r.t. their user activity patterns, network centrality measurements, and the content they produce. We discover that hateful users have created their accounts more recently and write more negative sentences. They use lexicon associated with categories such as hate, terrorism, violence and anger *less* than normal ones, and categories of words such as love, work and masculinity *more* frequently. We also find that the median hateful user is more central and that hateful users are densely connected in the retweet network. The latter finding motivates the use of an inductive graph embedding approach to detect hateful users, which outperforms widely used algorithms such as Gradient Boosted Trees. As moderation of Online Social Networks in many cases analyzes users, characterizing and detecting hate on a user-level granularity is an essential step for creating workflows where humans and machines can interact to ensure OSNs obey legislation, and to provide a better experience for the average user.

Nevertheless, our approach still has limitations that may lead to interesting future research directions. Firstly, our characterization only considered the behavior of users on Twitter, and the same scenario in other Online Social Networks such as Instagram or Facebook may present different challenges. Secondly, although classifying hateful users provides contextual clues that are not available when looking only at a piece of content, it is still a non-trivial task, as hateful speech is subjective, and people can disagree with what is hateful or not. In that sense, an interesting direction would be to try to create mechanisms of consensus, where online communities could help moderate their content in a more decentralized fashion (like Wikipedia [Shi et al., 2019]). Lastly,

a research question in the context of detecting hate speech on a user-level granularity that this work fails to address is *how much hateful content comes from how many users*. This is particularly important as, if we have a Pareto-like distribution where most of the hate is generated by very few users, then analyzing hateful users rather than content becomes even more attractive.

An interesting debate which may arise when shifting the focus on hate speech detection from content to users is how this can potentially blur the line between individuals and their speech. Twitter, for instance, implied it will consider conduct occurring "off the platform" in making suspension decisions. In this scenario, approaching the hate speech detection problem as we propose could allow users to be suspended to "contextual" factors —and not for a specific piece of content he or she wrote. However, as mentioned previously, such models can be used as a first step to detect these users, which then will be assessed by humans or other more specific methods.

The broader question this brings is to what extent a "black-box" model may be used to aid in tasks such as content moderation, where this model may contain accidental or intentional bias. These models can be used to moderate Online Social Networks, without the supervision of a human, in which case its bias could be very damaging towards certain groups, even leading to possible suppressions of individual's human rights, notably the right to free speech. Another option would be to make a clear distinction between using the model to detect possibly hateful or inadequate content and delegating the task of moderation exclusively to a human. Although there are many shades of gray between these two approaches, an important research direction is how to make the automated parts of the moderation process fair, accountable and transparent, which is hard to achieve even for content-based approaches.

Chapter 4

User Radicalization on YouTube

On YouTube, channels that discuss social, political and cultural subjects have flourished. Among these, one may find individuals such as Jordan Peterson and Joe Rogan, associated with the so-called Intellectual Dark Web (I.D.W.): iconoclastic thinkers, academics and media personalities [Weiss and Winter, 2018], but also openly declared white nationalists like Richard Spencer and Jared Taylor, which have been broadly referred to as Alt-right [ADL, 2019b].

These individuals do not only share the same platform, but often publicly engage in debates and conversations in the website [Lewis, 2018]. All the previously mentioned individuals, for example, are connected by joint video appearances: Jordan Peterson was interviewed by Joe Rogan [PowerfulJRE,], who interviewed YouTuber Carl Benjamin [PowerfulJRE,], who debated Richard Spencer [Andywarski,], who was in a panel with Jared Taylor in an Alt-right conference [RedIceTV,]. According to Lewis [Lewis, 2018], this proximity would create "radicalization pathways" for audience members and content creators. Anecdotal examples of these journeys are plenty, including *Roosh V*'s content creator trajectory, going from a Pick Up Artist to Alt-right supporter [Kutner, 2016, Roosh V, 2016], and Caleb Cain's testimony of his YouTube-driven radicalization [Faraday Speaks, 2019, Roose, 2019].

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 [Nic et al., 2018, Ingram, 2018]. In this setting, recent research has shown that fringe websites (like *4chan*) and subreddits (like */r/TheDonald*) have great influence over which memes [Zannettou et al., 2018a] and news [Zannettou et al., 2017] are shared in large social networks, such as Twitter. YouTube is extremely popular, espe-

cially among children and teenagers [Anderson and Jiang, 2018], and, if the streaming website is actually radicalizing individuals, this can push fringe ideologies like white supremacy further into the mainstream [Tufekci, 2018].

A key problem in dealing with topics like radicalization and hate speech is the lack of agreement over what is "hateful" or "extreme" [Sellars, 2016]. A work-around this issue is to perform community-based analyzes, rather than trying to label what is or is not hateful. For the purpose of this work, we consider three communities that have been associated with user radicalization [Lewis, 2018, Weiss and Winter, 2018, Roose, 2019], and that differ significantly in the extremity of their content: the Intellectual Dark Web (I.D.W.), the Alt-lite and the Alt-right. While the I.D.W. discuss controversial subjects like race and I.Q. [Weiss and Winter, 2018], the Alt-right sponsor fringe ideas like that of a white ethnostate [Hankes and Amend, 2018]. Somewhere in the middle, individuals of the Alt-lite deny embracing white supremacist ideology, although they constantly flirt with concepts associated with it (e.g. the great replacement, globalist conspiracies). This community-driven focus allows one to understand where individuals consuming extreme content are coming from and how does YouTube recommendation algorithms lump these communities together.

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

RQ1 What are the dynamics of the consumption and production of extreme content on YouTube?

RQ2 To which extent do users systematically steer towards more extreme content?

RQ3 Do algorithmic recommendations steer users towards more extreme content?

We develop a data collection process where we: *(i)* obtain a large pool of relevant channels from these communities; *(ii)* obtain 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 additionally collect traditional and alternative media channels to employ as control, These efforts resulted in a dataset with more than 79 million comments in $331,849$ videos of 350 channels, and with more than 2 million video and 10 thousand channel recommendations. We analyze this large dataset extensively:

We look at the growth of these communities throughout the last decade in terms of videos, likes, and views, finding a step rise in activity and engagement in the communities of interest when compared with the control channels (Sec. 4.3.1). We inspect the intersection of commenting users within the communities, finding they increasingly share the same commenting user base (Sec. 4.3.2). Moreover, we 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 consume only content from the I.D.W. or the Alt-lite throughout the years, consistently start to consume Alt-right content. These users are an expressive fraction of the Alt-right commenting user base. Interestingly, although control channels share, on a yearly basis, a significant number of users with Alt-right channels, we cannot observe significant user migration from them to Alt-right channels (Sec. 4.3.3). Lastly, we take a look at the impact of YouTube's recommendation algorithms, running simulations on recommendation graphs we construct with our data collection. Our analyzes, *given the recommender system in the time of the data collection, and without personalization,* show that the communities are indeed connected by YouTube's recommendation algorithms, but that it does not significantly steer users towards the Alt-right (Sec. 4.3.4).

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 algorithm does bind these communities together. However, its effect, given our experimental setup, is not as strong as suggested by anecdotal evidence. Yet, there are several limitations to our work, especially given that the recommendations obtained are not personalized. We discuss our findings and our limitations in light of the research questions further in Sec. 4.4. We argue that regardless of the influence of the recommender system in the process of radicalizing users, there is significant evidence that this process is happening, and that appropriate measurements should be taken.

4.1   Background

We discuss three of YouTube's prominent communities: the Alt-Right, the Alt-lite, and the Intellectual Dark Web. We argue that all of them are *contrarians*, in the sense that they strongly 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 [Nagle, 2017].

According to the Anti Defamation League [ADL, 2019a], 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.

The term Alt-lite was created to differentiate right-wing activists who deny to 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 [Atkinson, 2018]. Alt-right writer and white supremacist Greg Johnson [ADL, 2019b] describes the difference between Alt-right and Alt-lite by the origin of its nationalism: "The Alt-light 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 a The New Yorker article [Marantz, 2017]. Yet it is important to point out that the line between the Alt-right and the Alt-lite is blurry [ADL, 2019b], this is particularly tricky because many of the Alt-liters are accused of dog-whistling: attenuating their real beliefs to appeal to a more general public and to prevent getting banned [Lopez G., 2019, Joel Kelly, 2017]. To address this problem, in this paper we take a very conservative approach to our labeling, naming only the most extreme content creators as Alt-right. This is explained in further detail in Sec. 4.2.1.

The Intellectual Dark Web (I.D.W.), is a term coined by Eric Ross Weinstein to refer to a particular group of academics and podcast hosts. The neologism was later popularized in a New York Times opinion article [Weiss and Winter, 2018], where it is employed to describe: *"collection of 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 Sam Harris, Jordan Peterson, Ben Shapiro, Dave Rubin, and Joe Rogan, and also mentions a website with an unofficial list of members. Members of the so-called I.D.W have been accused of bigotry, including Islamophobia [Beydoun, 2018], transphobia [Lott, 2017] and sexism [Foderaro, 2018]. Moreover, a recent report by Data & Society research institute has claimed these channels are "pathways to radicalization" [Lewis, 2018]: they would act as an entry point to more radical channels, such as those in Alt-right. Broadly, members of this loosely defined movement see these critics as a consequence of discussing controversial subjects [Weiss and Winter, 2018] and largely ignored/dismissed

the report [The Rubin Report, 2018]. Similarly to what happens between Alt-right and Alt-lite, there is also blurry lines between the I.D.W. and the Alt-lite, especially for non-core members, like those listed in the website. Here, again, we take a conservative approach, considering borderline cases to belong to the Alt-lite.

4.1.1 Radicalization

We approach this central concept with the definition of McCauley and Moskalenko [McCauley and Moskalenko, 2008]: *Functionally, political radicalization is increased preparation for and commitment to intergroup conflict. Descriptively, radicalization means a change in beliefs, feelings, and behaviors in directions that increasingly justify intergroup violence and demand sacrifice in defense of the ingroup.* We use the consumption of Alt-right content as a proxy for radicalization. We argue this is reasonable because the rhetoric preached by the Alt-right has been associated with multiple recent terrorist attacks (e.g. the Christchurch mass shooting [Mann et al., 2019]), and because it champions ideas associated with intergroup conflict (e.g. a white ethnostate [Hankes and Amend, 2018]). Our conservative strategy when labeling channels is of particular importance here: Alt-right channels are closely related to these ideas, while the Alt-lite and the I.D.W. are given the benefit of the doubt.

4.1.2 Auditing recommendation systems

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 [Diakopoulos, 2014]. Previous large scale algorithmic auditing include measuring discrimination on AirBnB [Edelman and Luca, 2014], personalization on web search [Hannak et al., 2013] and price discrimination on e-commerce web sites [Hannak et al., 2014]. We argue this work is an audit in the sense that it sheds light into a troublesome phenomenon (user radicalization) in a content-sharing social environment heavily influenced by algorithms (YouTube). Unfortunately, it is impossible to obtain the entire history of YouTube recommendation, so we must limit algorithmic analyzes to a time slice of a constantly changing black-box. Although comments may give us insight into the past, it is impossible 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 analyzes provide answers to important question related with impactful societal processes that are allegedly happening on YouTube, and *(ii)* our framework for auditing user radicalization

can be replicated through time, and expanded to handle personalization. Regardless of the extent of the contribution of YouTube's algorithm towards the process of user radicalization, understanding this process and finding ways to fight it is still a timely question.

4.1.3 Previous research from/on YouTube

Previous work by Google sheds light into some of the high-level technicalities of YouTube's recommender system [Covington et al., 2016, Davidson et al., 2010]. Their latest paper indicates they use feed embeddings for video searches and video histories into a dense feed-forward neural network [Davidson et al., 2010]. There also exists a large body of work studying violent [Giannakopoulos et al., 2010], hateful or extremist [Sureka et al., 2010, Agarwal and Sureka, 2014] and disturbing content [Papadamou et al., 2019] 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 [Agarwal and Sureka, 2014, Giannakopoulos et al., 2010]. Notably, Sureka et al. [Sureka et al., 2010] 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 favourites. Ottoni et al. perform an in-depth textual analysis of 23 channels (13 broadly defined as Alt-right), finding significantly different lexicon and topics across the two groups [Ottoni et al., 2018].

4.2   Methods

4.2.1   Data Collection

We are interested three communities on YouTube: the Alt-lite, the I.D.W., 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 challenges multi-faceted: on one hand, we want to study user radicalization, and know, 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.

Table 4.1. Top 16 YouTube channels with the most views per each community and for controls.

	Alt-right	Views	Alt-lite	Views
1	James Allsup	62.20M	StevenCrowder	727.01M
2	Black Pigeon Speaks	49.97M	Rebel Media	405.12M
3	ThuleanPerspective	44.55M	Paul Joseph Watson	356.37M
4	Red Ice TV	41.98M	MarkDice	333.95M
5	The Golden One	12.06M	Stefan Molyneux	193.29M
6	AmRenVideos	9.08M	hOrnsticles3	144.98M
7	NeatoBurrito Productions	7.19M	MILO	133.03M
8	The Last Stand	6.52M	Styxhexenhammer666	132.17M
9	MillennialWoes	6.15M	OneTruth4Life	111.97M
10	Mark Collett	5.58M	No Bullshit	104.07M
11	AustralianRealist	5.29M	SJWCentral	89.99M
12	Jean-François Gariépy	4.80M	Computing Forever	86.69M
13	Prince of Zimbabwe	4.61M	The Thinkery	86.43M
14	The Alternative Hypothesis	4.60M	Bearing	81.16M
15	Matthew North	4.41M	RobinHoodUKIP	64.00M
16	Faith J Goldy	4.07M	patcondell	63.67M

	Intellectual Dark Web	Views	Control	Views
1	PowerfulJRE	1.07B	Vox	1.29B
2	JRE Clips	716.55M	Young Turks	1.12B
3	PragerUniversity	634.77M	GQ Magazine	1.09B
4	SargonofAkkad100	257.76M	Vice News	1.06B
5	The Daily Wire	246.64M	WIRED	1.05B
6	The Rubin Report	206.03M	ABC News	973.14M
7	ReasonTV	137.68M	MSNBC	824.96M
8	JordanPetersonVideos	90.49M	RT News	677.68M
9	Bite-sized Philosophy	62.39M	BBC	660.06M
10	Timcast	40.42M	Vanity Fair	639.13M
11	Owen Benjamin	34.80M	The Verge	636.06M
12	AgatanFoundation	32.92M	Glamour Magazine	619.75M
13	Essential Truth	32.44M	Fox News	584.50M
14	Ben Shapiro	29.99M	Business Insider	522.80M
15	YAFTV	29.61M	Next News Network	465.01M
16	joerogandotnet	24.66M	CBS News	452.79M

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)** label these channels according to the communities of interest using both manual input and the data that we collected for each channel.

(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. We do as follows:

1. We choose a set of *seed channels*. Seeds were extracted from the I.D.W. unofficial website [Anonymous,], ADL's report on the Alt-lite/the Alt-right [ADL, 2019b] and Data & Society's report on YouTube Radicalization [Lewis, 2018]. The idea is to 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.

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 questions.

3. We iteratively search the related and featured channels collected in steps 1 and 2, adding relevant channels (as defined in 2). We repeat the procedure for the recently collected channels. This, as well as Step (2), was done by an individual with more than 50 hours of watch-time of the communities of interest. Notice that here we are not labeling the channels, but creating a pool of channels to be further inspected and labeled in subsequent steps.

(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, 20 for channel recommendations and 10 for video recommendations. Each "round" consists of collecting all recommended channels (on the channel web page) and all recommended video (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 U.S.A, 2 in Canada, 1 in Switzerland and 1 in Brazil. Moreover, channels were always visited at random order, to prevent any biases from arising from session-based recommendations.

(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 researchers annotate them carefully. They both had significant experience with the communities being studied, and were given the following instructions:

> Given the set of channels in this table, you should carefully inspect each one
> of them: taking a look at the most popular videos and watching, altogether,

Table 4.2. Overview of our dataset.

Channels	350
Videos	331, 849
Comments	79,180,534
Video recommendation rounds	19
Video recommendations	2,474,044
Channel recommendation rounds	22
Channel recommendations	14,283

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 [link], and check out the website with some of the alleged members of the group [link]. 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 [link] and [link]. 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. [1]

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%. We ended up with 91 I.D.W., 114 Alt-lite and 88 Alt-right channels.

Controls Additionally we collect news-related channels as control channels. These were obtained from the *mediabiasfactcheck.com* [Check, 2019]. 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 [Check, 2019]. Some of the channels were not considered because they had too many videos (15, 000+) and we weren't able to retrieve all their videos (which is important, because our analysis are temporal). In total, we collect 68 channels in that way.

We summarize the dataset collected in the Tab. 4.2. Data collection was performed from the 19th to the 30th of May 2019, and the collection of the recommendations between May and July 2019.

---

[1] Links were ADL's report [ADL, 2019b] and the New Yorker's article on Alt-right/lite [Marantz, 2017], the NYT a article on the I.D.W. [Weiss and Winter, 2018], and the I.D.W. website [Anonymous,]

4.3 Results

4.3.1 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. Tab. 4.1 shows the 16 most viewed YouTubers for each of the communities and for the controls, and Figure 4.1 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. 4.1(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 control channels. Notice-ably, the rise in the number of active channels is much more recent for the communities of interest than for control channels, as shown in Fig. 4.1(a). In mid 2015, for example, while more than 75% of the channels in the control group were already created, only slightly more than 50% of Alt-lite and less than 50% of Alt-right and I.D.W. channels had been created. This growth in the communities of interest during 2015 may also be noted in Fig. 4.1(i), which shows the CDF of number comments per videos, and can also be seen between early 2014 and late 2016 in Figs. 4.1(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
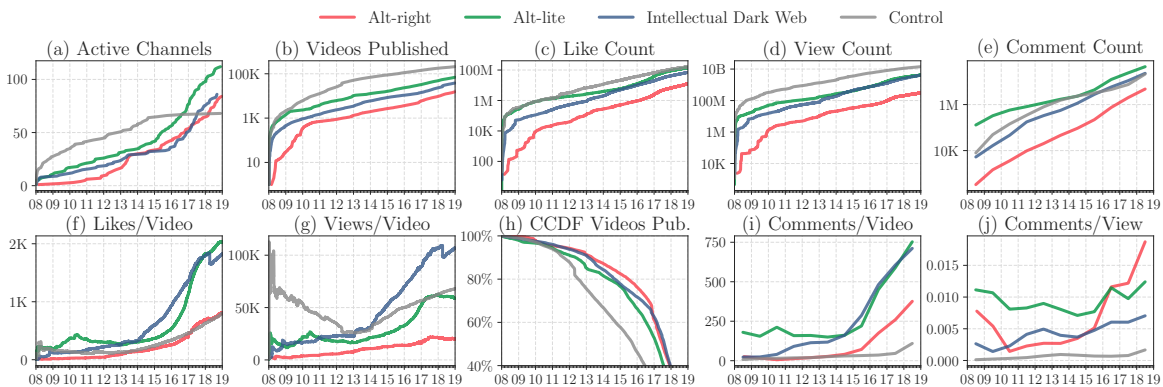
Figure 4.1. In the top row (a)-(e), for each community and for the control channels, we have the cumulative number of active channels (that posted at least one video), of videos published, of likes, views and of comments. Recall that the number of likes and views is obtained at the moment of the data collection. In the bottom row, we have CDFs for engagement metrics, and the CCDF of videos published, zoomed in in the range [40%, 100%] on the y-axis. Notice that for comments, we know only the year when they were published, and thus the CDFs granularity is coarser (years rather than seconds).

from those channels became popular later. Altogether, our data corroborates with the narrative that these communities gained traction (and fortified) Donald Trump's campaign during the 2016 presidential elections [Campaigns et al., , Gray, 2015].

Engagement. A key difference between the communities of interest and the control 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, portrayed in Figs. 4.1(f), (i), and (j), respectively. For all these metrics, the communities of interest have more engagement than the control channels: Although control channels have more views per video, as shown in Figs. 4.1(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 50 views by 2018.

Dormant Alt-right Channels. Although by 2013, approximately the same number of channels of all three communities had been created ($\sim$ 30), as it can be seen in Fig. 4.1(a) the number of videos they published by the Alt-right was very low before 2016. This can be better seen in the CCDF in Fig. 4.1(h): while control 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 on Tab. 4.1. Most of the initial videos in the channel are about working out or video-games, with politics related videos becoming increasingly occurring. When taking into account that 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, this resonates with the narrative that the rise of Alt-Lite and I.D.W. channels created fertile grounds for individuals with fringe ideas to prosper [Nagle, 2017, Lewis, 2018].

Although our data-driven analysis sheds light into existing narratives on the communities of interest, it is still very difficult to see, 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.

4.3.2 User Intersection

We begin our in-depth analysis of users who commented on the channels of interest by analyzing the *intersection* between the users in different channels and communities.
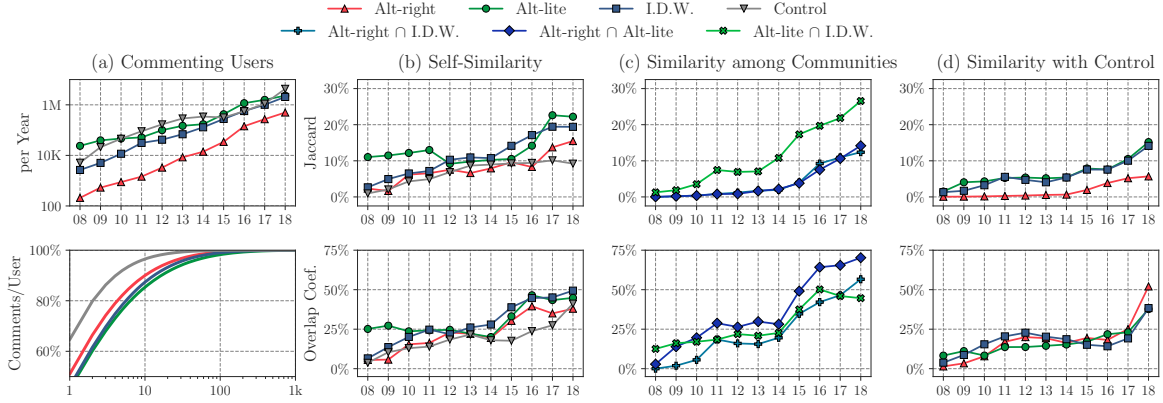
Figure 4.2. In *(a)*, the number of unique commenting users per year in the top figure and the CDF of comments per user for each one of the communities in the bottom figure. In *(b)—(d)* we show two similarity metrics (Jaccard and Overlap Coefficient) for different pairs of sets of commenting users across 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 on control channels. Notice that comments are clumped together per year, so here, unlike in Fig 4.1, 2017 means from 2017 to 2018, and so forth.

In that context, we use two set similarity metrics: the Jaccard Similarity $\frac{|A \cap B|}{|A \cup B|}$; and the Overlap Coefficient $\frac{|A \cap B|}{min(|A|,|B|)}$. Noticeably, previous studies have shown that commenting users are not a representative sample of users who engage with online content. In news, for example, they tend to be older and more often male [Ziegele et al., 2013]. Here, it is reasonable to assume they are more engaged with the content than simple viewers.

Column (a) of Fig. 4.2 characterizes commenting users. The top figure shows the absolute number of commenting users per year, while the bottom one shows the CDF of the number of comments per users per community. It is interesting to compare these plots with that of Fig. 4.1(e), as we can see that the communities of interest have way more assiduous commenters. This supports the hypothesis that users that consume content in the communities of interest are more "engaged" than those that consume the content from the control channels. Notice that although the Alt-right commenters have, in average, fewer comments than the Alt-lite or the I.D.W., the community is way younger (as discussed in Sec. 4.3.1), and thus it is hard to tell whether their users are less engaged.

In columns (b)—(d) of Fig.4.2 we consider the intersection between the commenting users of the *communities* of interest and the control channels. The top figure for each column shows the Jaccard similarity and the bottom one the Overlap Coefficient.

Column (b) in Fig. 4.2 shows the self-similarity for a community with itself a year before. We find that the retention of users among the three communities is growing with time for both metrics. However, for control channels, we find that the Jaccard similarity is actually decreasing since 2015, 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 the control channels.

Column (c) in Fig. 4.2 shows the similarity within the three communities. Notably, the Jaccard similarity between the Alt-lite and the I.D.W. is higher than the self-similarity of these both communities, reaching almost 30%. Moreover, the Overlap Coefficient of the Alt-right with the Alt-lite and the Alt-right is high: approximately 55% for the I.D.W. and 70% for the Alt-lite. This means more than half of the users who commented on Alt-right channels commented on both the other two communities.

Lastly, column (d) in Fig. 4.2 shows the similarity of the three communities with the control channels. We have that the Jaccard similarity between the I.D.W. and the Alt-lite and the control 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 on this communities actually make up a significant portion of the massive media channels we use for control, which gather billions of views. On the other, it shows that the Alt-right, a group of channels with an order of magnitudes less views, subscribers and comments, are actually *on par* with these large channels. Inspecting the Overlap Coefficient, however, we get a different panorama: there we have that the communities overlap more with themselves than with the control channels. Interestingly, for both similarity measurements, we can see a sharp growth in the similarity with control after 2016. This may be explained due to the sharp increase in popularity of these communities since 2015, as discussed in Sec. 4.3.1, as these communities become more "mainstream", it may be that users who comment watch the control channels eventually stumble upon them.

4.3.3   User Migration

In the previous section we portrayed a scenario where the commenting user bases among the communities is increasingly similar. Although that may indicate a growing percentage of users consuming extreme (here the Alt-right) content on YouTube *while also* consuming content from other milder communities (here the Alt-lite and the I.D.W.), it does not, *per se*, indicate that there is a radicalization pipeline in the website. To better address this question, we find users who *did not comment* in Alt-right content
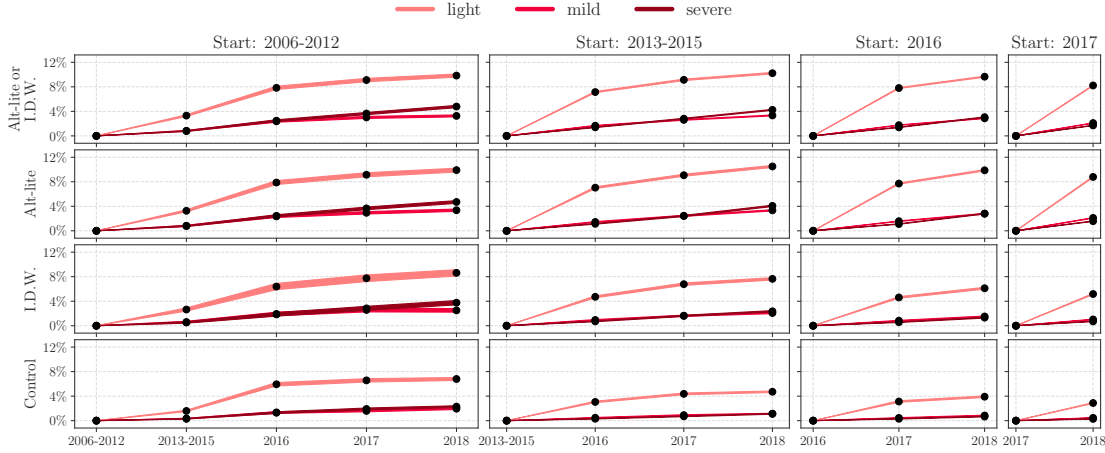
Figure 4.3. We show how users "migrate" towards Alt-right content. For users who consumed only videos in the communities indicated by the labels in the rows (Alt-lite or I.D.W., only Alt-lite, only I.D.W. or Control), we show the probability of them becoming consuming Alt-right content. We consider three levels of "infection": light (commented on 1 to 2 Alt-right videos), mild (3, 5) and severe (6+). Each column tracks users in a different starting date. Initially, their infection rates are 0 (as they did not consume any Alt-right content). As time passes, we show the infection rates in the y-axis, for each of the years, in the x-axis.

in a given year and look into their activity into subsequent years.

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

The results for this analysis are shown in Fig. 4.3. We show the percentage of users who become infected of the users we managed to track. We find that a high percentage of users became "infected", according to our criteria. Consider, for example, users who in $2006-2012$ commented only on I.D.W. or Alt-lite content $(227,945$ users$)$. By 2018, 21.83% were still active, and from those, 17.9% (around $9,000$ users) were "infected" in one of the three levels, around 4.8% of them severely so. From the ones who in 2017 commented only in Alt-lite or I.D.W. videos $(1,253,751$ users$)$ 50% were active the following year, and approximately 12% of them became infected, around 3.6% mildly or severely so —more than $26,000$ users. Interestingly, when considering users who commented both in the Alt-lite and the I.D.W., or only in the Alt-lite, the
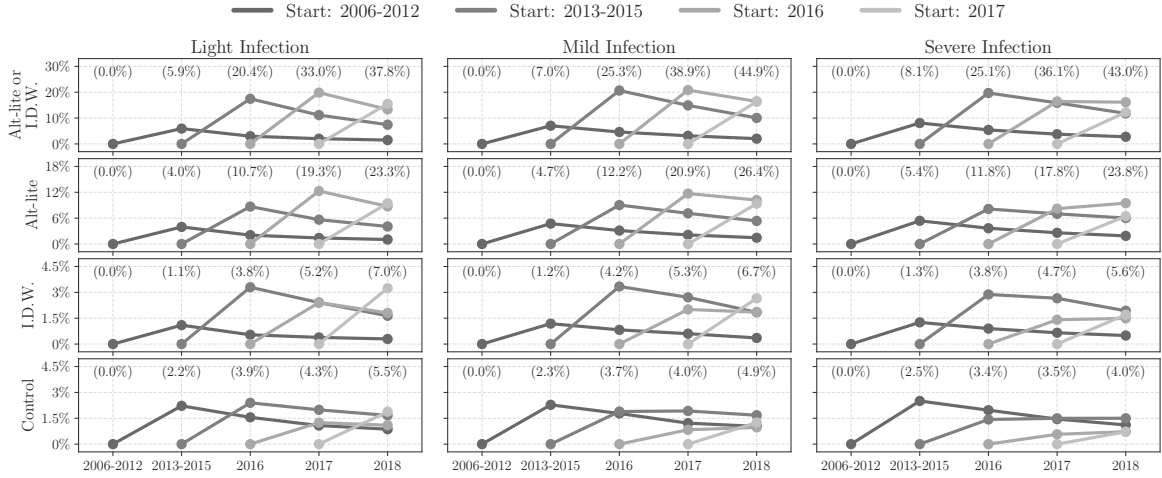
Figure 4.4. We show the percentage of users that can be traced back as not-infected users who commented on other communities. Each line represents users who, in a given start date, commented only Alt-lite or I.D.W. content, the y-axis shows the percentage of the total Alt-right commenting users they went to become (notice that all lines begin at 0 as users initially did not consume any Alt-right content).

speed of infection seems to be increasing year by year. When comparing the infection rates of the communities of interest with the control channels, we find that they all present higher infection rates, particularly for mild and severe infections. Moreover, we also find that, for light infections, the rate is significantly smaller for users who were tracked after 2013. When teasing apart users that commented only on Alt-lite or only on I.D.W. content, we find that, not only users that commented only on content from the I.D.W. get less infected, but increasingly less so, as with the control channels. For example, the radicalization rates of users who watched only Alt-lite or only I.D.W. content are much more similar for those tracked in $2006 - 2012$ than for those tracked in 2017 (infection is less prevalent among I.D.W. commenters in recent years).

The previous experiment 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 "infected" 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 infected (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. 4.4. We find that these users are

a considerable fraction of the Alt-right commenting audience. In 2018, for example, for all kinds of infections, roughly 40% of commenting users can be traced back from cohorts of users that commented only in Alt-lite or I.D.W. videos in the past. 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 generates the peaks in the first row of Fig. 4.4. Moreover, looking at the second and third row of Fig. 4.4, we find a substantial difference between the I.D.W. and the Alt-lite. Whereas in Sec. 4.3.2 we find that the intersection between them both and the Alt-right are very similar, here we see that users who commented *only* on I.D.W. channels constitute a way less significant percentage of the Alt-right consumer base in upcoming years when compared to the Alt-lite. For all levels of infection, at all times, there are more than roughly 3 times more users in the Alt-right commenting base that commented exclusively in the I.D.W. than in the Alt-lite. So, while in 2018, 23.3% of users who were lightly infected 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. As a matter of fact, the percentages of the I.D.W. are not so different from those of the control channels, shown in the last row of Fig. 4.4.

The experiments performed in this section are particularly relevant given our analysis (in Sec. 4.3.2) showing the growing intersection of the commenting user bases. There we find that *(i)* the differences in user intersection on a yearly basis are not so dire between the milder communities and the Alt-right, and control channels and the Alt-right; and *(ii)* that the intersection between the I.D.W. and the Alt-right is similar to that of the Alt-lite of the Alt-right. Here we see that, despite that, the systematic migration of users who consume Alt-lite content (along with I.D.W. or not) to Alt-right content is significantly more expressive than that of users who consume only I.D.W. content initially, which in itself is only marginally more expressive than that of users who consumed only control channels initially. Importantly, we mean "expressive" both in terms of the percentage of the users we manage to track through the comments (as in Fig. 4.3), but also in terms of the percentage of users comment on the Alt-right videos (as in Fig. 4.4).

4.3.4   Recommendation Algorithm

In this section, we inspect the impact of YouTube's recommendation algorithm, examining if the algorithm prioritizes more extreme content, as it has often been claimed [Tufekci, 2018]. We perform our analysis in a recommendation graph, built using the data collected. A noteworthy fact is that channels recommendations were

(a)

(b)
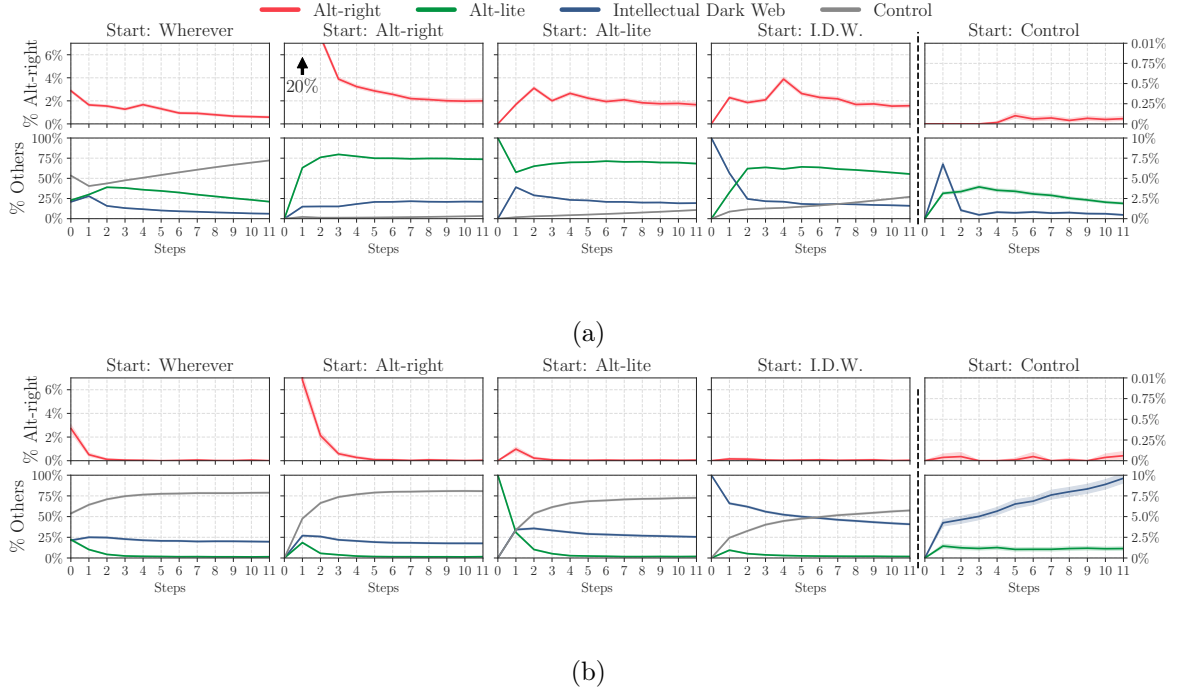
Figure 4.5. We show the results for the simulation of random walks for channels (a) and videos (b). The top row shows the chance of the random walker being in an Alt-right channel at each step, while the bottom row shows the chance of the random walker being in any of the other communities. The different columns portray different starting rules: in any channel, only in channels of the Alt-right, and so forth.

disabled for most of the channels we are studying following a dispute between Carlos Maza and Steven Crowder [Goggin, 2019], where the former asked for YouTube to act in response of several episodes of personal attacks, where Crowder made fun of Carlos' ethnicity and sexual orientation. 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). Each edge is weighted proportionally to the number of times that recommendation appeared in the data collection, and weights are normalized so that outcoming edges of each node sums to 1 (thus creating a stochastic graph).

The percentage of edges flowing from each community to another (normalized by their weight) is shown in Tab. 4.3 and Tab. 4.4 for channel and video recommendations, respectively. We have very different scenarios for each one of the recommender systems. For channel recommendations, we have that control channels are recommended scarcely by the communities of interest, although they do have some edges to Alt-lite and I.D.W. channels. Alt-lite and I.D.W. channels recommend another channel from the same

community around 68% of the time, and recommend each other around 25% of the time. Alt-right channels are rarely recommended by both the Alt-lite (2.72) and the Alt-right 4, although significantly more by the Alt-lite. Moreover, these channels recommend Alt-lite channels more than they recommend other Alt-right channels. They also very rarely recommend control channels. For video recommendations, control channels are recommended very often across the communities, more than communities themselves for the Alt-lite and the Alt-right. The only community that remains recommending itself often is the I.D.W (around 60% of the time). Whereas the Alt-lite recommends the I.D.W. about the same percentage of times as it recommends itself (roughly 30%), the I.D.W. recommends the Alt-lite less frequently.

Given these graphs, we perform experiments considering a random walker. We do as follows. 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 11 steps, choosing edges at random with probabilities proportional to their weights. We collect the communities of the channels visited by the random walker and calculate the probability of it visiting channels from each of the communities. We consider the case where the random-walk start in any node, but also the case where its starting point is confined within a single community. The probabilities of the random walker to be in each of the communities, at each step, given different starting conditions is shown in Fig. 4.5, for channel and video recommendations respectively.

For channel recommendations, we find that, from the three communities of interest, users eventually reach 2% of chance of being in an Alt-right channel. This is approximately 2/3 of the probability a user would land in such channel if they picked the channels at random given their number of subscribers ($\sim 3\%$). When starting in the Alt-right, the chance of visiting another Alt-right channel quickly drops, reaching 4% in the third hop. For the other communities, the situation changes according to the starting rule. If we start wherever or in control channels, we actually tend to increasingly consume control channels. On the other cases, we end up being steered towards Alt-lite content.

Video recommendations yield very low chances for users to be recommended Alt-right channels in all scenarios. The chance quickly converges to 0 in all cases. Here, control channels are also favoured much more than in the channel recommendation graph, and the chance of being in a control channel eventually increases to more than 50% in all cases. Moreover, in this scenario, the recommender system clearly steer users towards the I.D.W.. in detriment of the Alt-lite. For example, if you start in I.D.W. videos, there is less than 5% of chance that you will go towards Alt-lite channels, and

Table 4.3. Percentage of edges in-between communities in the channel recommendation graph (normalized per weight).

Src ↓ | Dest. →	Alt-lite	Alt-right	Control	I.D.W.
Alt-lite	68.02	2.72	4.27	25.00
Alt-right	46.38	35.64	1.75	16.23
Control	11.06	0.00	75.23	13.71
I.D.W.	26.29	0.43	5.52	67.76

Table 4.4. Percentage of edges in-between communities in the video recommendation graph (normalized per weight).

Src ↓ | Dest. →	Alt-lite	Alt-right	Control	I.D.W.
Alt-lite	28.67	2.59	38.86	29.88
Alt-right	17.74	16.39	38.00	27.88
Control	2.25	0.05	91.19	6.52
I.D.W.	12.70	0.52	26.95	59.82

if you start in Alt-lite videos, after 5 steps you are more likely to be in an I.D.W. video than in an Alt-lite video.

Overall, these findings are nuanced. Channel recommendations do seem to steer users towards the Alt-lite, which is definitely closer to some far-right talking points, while video recommendations do seem to steer users towards I.D.W. content. Yet, in both cases, the recommender systems seem particularly unfavorable to Alt-right content, which is less represented than it would be if we randomly picked the channel among our pool. It is worthwhile to mention that there are several limitations to our approach, and these are discussed further in Sec. 4.4.

4.4 Discussion

We performed a throughout analysis of three YouTube communities —the Alt-right, the Alt-lite and the Intellectual Dark Web— by inspecting a large dataset containing millions of comments and recommendations from thousands of videos. We find several data-driven insights associated with the questions proposed in the introduction.

The communities studied sky-rocketed in terms of views, likes, videos published and comments, particularly since 2015, coinciding with the turbulent presidential election of that year. The communities of interest also have a lot of engagement in the form of comments, gathering much more comments per video and comments per views than the Even if one considers only Alt-right content as "extreme", our results indicate

that content production and consumption are growing exponentially.

We find that the user bases for the three communities are increasingly similar, 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. We argue that this finding comprises significant evidence that there has been, and there continues to be, user radicalization on YouTube, and our analyzes of the activity of these communities are consistent with the theory that more extreme content "piggybacked" the surge in popularity of I.D.W. and Alt-lite content [Nagle, 2017]. The analyzes done in Sec. 4.3.3 show the phenomenon is not only consistent throughout the years, but also that it is very expressive in its absolute quantity. Even if one considers that only a very small quantity commenting users we deem as infected are getting radicalized, they are still in the magnitude of the thousands (not to mention those who only watch the videos and never comment on them). Our results also show that there is a significant difference for Alt-lite and for I.D.W. channels. Users who consume only the latter seems to migrate to more extreme content less, and are a less significant portion of the Alt-right commenting user base.

Our analysis also suggests that YouTube's recommendation algorithm, given our experimental setup, does not strongly favor the Alt-right, although it does bind the three communities of interest together. In our simulations, the representativity of Alt-right channels was smaller than what it should be considering the number of subscribers these channels have. Still the question of "is that enough?" does not seem to have a clear answer, particularly when Alt-lite content, seems to be in violation of YouTube's own guidelines against hate speech. 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 this content was being promoted in the past. Secondly, our analysis does not take into account personalization, which could reveal a completely different picture. Although it may be impossible to completely tease apart from historical data what role the algorithm had in influencing users to migrate from milder to more extreme content, techniques like the one employed here could be adapted to audit YouTube in years to come.

Future Work In this chapter, we focused almost exclusively at 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: what are the narratives that arose, how did their tone change. Looking at text could also improve the criteria

used for radicalization, we could rule out users that are having negative responses to extremist videos (although this is not substantial in the dataset). Moreover, we intend to extend the existing framework to audit radicalization to take into account user personalization —this is not trivial, as knowing the trajectories of radicalized users through YouTube content is unfeasible.

Chapter 5

Conclusion

In the introduction, we argued that a user-centric approach may aid towards better characterization and detection of ill-defined social phenomena. Here, we elaborate on this argument, and explain how this approach was beneficial in each of the three case studies presented. Moreover, we argue that each case study illustrate a way of thinking of users when modeling ill-defined phenomena such as hate speech and fake news.

In the first case study, we looked at the interaction between political polarization and misinformation, analyzing users and their social networks to understand *who* was engaging with content. Broadly, we want to associate users' characteristics (in this particular case, political leaning) to content. One could even try to infer one's political opinion from a piece of content, but this characteristic is not of the piece of content itself, but of the user who created or shared it. To put it more broadly, in Chapter 2, our user-centric focus allow us to relate a user-centered characteristic with content. Closer inspection shows that this is a surprisingly common modeling recipe. For example, [Shi et al., 2019] proposes correlating political polarization of users with the quality of Wikipedia articles. There, analogously to our paper, we are able to learn something about a piece of content (a Wiki article) by analyzing users associated with it.

In Chapter 3 we propose characterizing and detecting hateful users. This is different from what we did in Chapter 2, as here, we largely disassociate the content of the research question —it becomes just another dimension that we can analyze for users. Instead of trying to employ user context to detect whether a piece of content is hateful, we abandon the idea of trying to classify content altogether, and focus strictly on users. This abstraction is particularly useful for problems such as hate speech, as it is very hard to disassociate content from user —what is hateful depends on who is saying it— and as we are usually interested in taking action against the user who spread hateful content [Google, 2019, Twitter, 2019].

Lastly, in Chapter 4 we study a phenomenon that only makes sense in the granularity level of users —radicalization. If in Chapter 2 and 3, you could still argue that, at least ignoring trickier cases, there are things that are inherently hateful or inherently fake, regardless of who is sharing this content, here you cannot. In a sense, this is a third category of how we can better understand hate speech and fake news focusing on users: it allows us to study relevant phenomena which happens in a more complex level of abstraction.

In retrospect, we can pinpoint the different roles users had in each Chapter:

- In Chapter 2, we enriched our understanding of a kind of content by analyzing users.

- In Chapter 3, we studied a phenomenon associated with content at the user level.

- In Chapter 4, we studied a phenomenon associated with users.

After more explicitly defining what we mean by "user-perspective", understanding why it is beneficial to focus on users when studying ill-defined phenomena such as hate speech of fake news becomes clearer. As it is particularly hard to indicate what is hate or fake, focusing on users may help us: *(i)* to understand the nuances of these definitions (as done in Chapter 2); *(ii)* to simplify the problem, ignoring some of these nuances by adopting a more granular model of the world (as in Chapter 3); and, lastly, *(iii)* to explore phenomena that are associated with hateful and fake content, but which requires this more granular model to be studied.

5.1   Major Themes across the Chapters

After we have established in broad strokes the contribution of this dissertation, we take the opportunity to take a step back and intertwine the major themes from the different chapters. We identify 3 of such major themes:

Drawing Boundaries is Hard. In Chapters 3 and 4, we have posed questions that ultimately require that we subjectively categorize content or users —we had to make decisions about which user profiles were hateful, or which channels were radical. This kind of research is particularly challenging, as it requires that we approach the material with depth and domain experience. Yet, this close inspection is crucial, after all, the most important questions in the study of online hate or misinformation will require ground truth for *what is* hate speech and misinformation. In Chapter 3, drawing boundaries between hateful and non-hateful users allowed us to characterize the

behavior of such users and to develop a method to better detect them. In Chapter 4, classifying YouTube channels as belonging to one of the three categories of interest —I.D.W., Alt-lite, and Alt-right— allowed us to study the migration of such users across these communities. Importantly, drawing these boundaries is no easy task, and thus all efforts must be made that the methodology is throughout and transparent.

No Silver Bullet. Fake news, hate speech and radicalization are complex and ever-changing phenomena. As mentioned before, our work suggests that enriching the understanding of content related to these phenomena through user context and studying such phenomena at the user level presents many benefits. In that sense, our work proposes approaches that are distant from an out-of-the-box-solution framework. This avoidance of over-simplistic approaches towards these intricate issues is the driving force behind the methodologies of Chapters 2 and 3, and our findings suggest that over-simplification may indeed cause problems. In Chapter 2, we avoided simplistic definitions of what is fake or what is hateful and, as a consequence, realized that relying on what people name as fake in social media is not practical (as they disagree over what is fake news). In Chapter 3, we examined hate speech at a user level and developed a sampling methodology less dependant on a specific set of words. Not only this allowed us to paint a more realistic portrait of hate speech on Twitter, but it showed us that the choice of words of hateful users is highly counter-intuitive: they use fewer words related to anger and more words related to love, for example.

Research as an Auditing Tool. Lastly, a major theme across the dissertation is the role of research as a way to uncover social phenomena that: 1) interests society at large, 2) takes place in online social networks. In many cases, here particularly in Chapter 4, it is particularly hard to study such phenomena because social network platforms are not designed to be crawled or studied. Yet, this kind of research is important to understand the impact of different technologies in our society. In Chapter 3 we were able to characterize the users that Twitter banned as hateful during a months-long period. This is particularly important, as the discussion over what should be moderated is of tremendous interest to society. In Chapter 4 we were able to confirm that user radicalization indeed took place on YouTube in the last couple of years. Our work provides quantitative evidence previous anectotes and qualitative research that suggested so.

Bibliography

[ADL, 2018] ADL (2018). Hate on Display™ Hate Symbols Database.

[ADL, 2019a] ADL (2019a). Alt Right: A Primer about the New White Supremacy.

[ADL, 2019b] ADL (2019b). From Alt Right to Alt Lite: Naming the Hate.

[Agarwal and Sureka, 2014] Agarwal, S. and Sureka, A. (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*, HT '14, pages 294--296, New York, NY, USA. ACM. event-place: Santiago, Chile.

[Allcott and Gentzkow, 2017] Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211--236. ISSN 0895-3309.

[Anderson and Jiang, 2018] Anderson, M. and Jiang, J. (2018). Teens, Social Media & Technology 2018. Technical report, Pew Research Center.

[Andywarski,] Andywarski. Richard Spencer, Styx and Sargon Have a Chat - Andy and JF moderate.

[Anonymous,] Anonymous. The Intellectual Dark Web.

[Atkinson, 2018] Atkinson, D. C. (2018). Charlottesville and the alt-right: a turning point? *Politics, Groups, and Identities*, 6(2):309--315. ISSN 2156-5503.

[Bakshy et al., 2015] Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130--1132. ISSN 0036-8075, 1095-9203.

[Benevenuto et al., 2010] Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010). Detecting spammers on twitter. In *In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS*.

[Bengaluru et al., 2018] Bengaluru, C., Dhule, G., Hyderabad, K., and Raipur, R. (2018). Murderous mob — 9 states, 27 killings, one year: And a pattern to the lynchings. *The Indian Express.*

[Beydoun, 2018] Beydoun, K. A. (2018). US liberal Islamophobia is rising – and more insidious than rightwing bigotry | Khaled A Beydoun. *The Guardian.* ISSN 0261-3077.

[Boult, 2017] Boult, A. (2017). Prisoner dressed as woman in failed escape bid. *The Telegraph.* ISSN 0307-1235.

[Brooks and Boadle, 2018] Brooks, B. and Boadle, A. (2018). Divisive Brazil election careens into 'dangerous' polarization. *Reuters.*

[Burke, 2017] Burke, J. (2017). The myth of the 'lone wolf' terrorist. *The Guardian.* ISSN 0261-3077.

[Burnap and Williams, 2016] Burnap, P. and Williams, M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11. ISSN 2193-1127.

[Calais Guerra et al., 2011] Calais Guerra, P. H., Veloso, A., Meira, Jr., W., and Almeida, V. (2011). From Bias to Opinion: A Transfer-learning Approach to Real-time Sentiment Analysis. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 150--158, New York, NY, USA. ACM.

[Campaigns et al.,] Campaigns, i., Elections, Parties, Action, C., Groups, I., Papers, Politics, Government, Research, and Technology. "Alt-Lite" Bloggers and the Conservative Ecosystem.

[Cha et al., 2010] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *Fourth International AAAI Conference on Weblogs and Social Media.*

[Chakraborty et al., 2016] Chakraborty, A., Paranjape, B., Kakarla, S., and Ganguly, N. (2016). Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '16, pages 9--16, Piscataway, NJ, USA. IEEE Press.

[Check, 2019] Check, M. B. F. (2019). Media Bias Fact Check.

[Chen et al., 2015] Chen, Y., Conroy, N. J., and Rubin, V. L. (2015). Misleading Online Content: Recognizing Clickbait As "False News". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, WMDD '15, pages 15--19, New York, NY, USA. ACM. event-place: Seattle, Washington, USA.

[Conover et al., 2011] Conover, M. D., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., and Flammini, A. (2011). Political Polarization on Twitter. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM 2011*, Barcelona, Catalonia, Spain. AAAI Press.

[Conroy et al., 2015] Conroy, N. J., Rubin, V. L., and Chen, Y. (2015). Automatic Deception Detection: Methods for Finding Fake News. In *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ASIS&T '15, pages 82:1--82:4, Silver Springs, MD, USA. ASIS&T.

[Corner, 2017] Corner, J. (2017). Fake news, post-truth and media–political change. *Media, Culture & Society*, 39(7):1100--1107. ISSN 0163-4437.

[Covington et al., 2016] Covington, P., Adams, J., and Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191--198, New York, NY, USA. ACM. event-place: Boston, Massachusetts, USA.

[Cunha et al., 2018] Cunha, E., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (2018). Fake News as We Feel It: Perception and Conceptualization of the Term "Fake News" in the Media. In Staab, S., Koltsova, O., and Ignatov, D. I., editors, *Social Informatics*, Lecture Notes in Computer Science, pages 151--166. Springer International Publishing.

[Davidson et al., 2010] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., and Sampath, D. (2010). The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293--296, New York, NY, USA. ACM. event-place: Barcelona, Spain.

[Davidson et al., 2017] Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. In *Eleventh International AAAI Conference on Web and Social Media*.

[Dhingra et al., 2016] Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., and Cohen, W. (2016). Tweet2vec: Character-Based Distributed Representations for Social Media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 269--274, Berlin, Germany. Association for Computational Linguistics.

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

[Edelman and Luca, 2014] Edelman, B. G. and Luca, M. (2014). Digital Discrimination: The Case of Airbnb.com. SSRN Scholarly Paper ID 2377353, Social Science Research Network, Rochester, NY.

[Faraday Speaks, 2019] Faraday Speaks (2019). My Descent into the Alt-Right Pipeline.

[Fast et al., 2016] Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4647--4657, New York, NY, USA. ACM. event-place: San Jose, California, USA.

[Ferrara et al., 2016] Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016). The Rise of Social Bots. *Commun. ACM*, 59(7):96--104. ISSN 0001-0782.

[Foderaro, 2018] Foderaro, L. W. (2018). Alexandria Ocasio-Cortez Likens $10,000 Debate Offer by Conservative Columnist to Catcalling. *The New York Times*. ISSN 0362-4331.

[Fox News, 2019] Fox News (2019). Tucker: No American citizen has been charged with collusion. *Fox News*.

[Funke, 2018] Funke, D. (2018). Reporters: Stop calling everything 'fake news'. *Poynter*.

[Garimella et al., 2017] Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2017). Reducing Controversy by Connecting Opposing Views. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 81--90, New York, NY, USA. ACM. event-place: Cambridge, United Kingdom.

[Giannakopoulos et al., 2010] Giannakopoulos, T., Pikrakis, A., and Theodoridis, S. (2010). A Multimodal Approach to Violence Detection in Video Sharing Sites. In *2010 20th International Conference on Pattern Recognition*, pages 3244--3247.

[Goggin, 2019] Goggin, B. (2019). YouTube's week from hell: How the debate over free speech online exploded after a conservative star with millions of subscribers was accused of homophobic harassment. *Business Insider*.

[Golub and Jackson, 2010] Golub, B. and Jackson, M. O. (2010). Naïve Learning in Social Networks and the Wisdom of Crowds. *American Economic Journal: Microeconomics*, 2(1):112--149. ISSN 1945-7669.

[Google, 2019] Google (2019). Hate speech policy.

[Gottfried and Shearer, 2016] Gottfried, J. and Shearer, E. (2016). News Use Across Social Media Platforms 2016. Technical report, Pew Research Center.

[Gray, 2015] Gray, R. (2015). How 2015 Fueled The Rise Of The Freewheeling, White Nationalist Alt- Movement. *BuzzFeed News*.

[Greevy and Smeaton, 2004] Greevy, E. and Smeaton, A. F. (2004). Classifying Racist Texts Using a Support Vector Machine. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 468--469, New York, NY, USA. ACM. event-place: Sheffield, United Kingdom.

[Grinberg et al., 2019] Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425):374--378. ISSN 0036-8075, 1095-9203.

[Groskopf, 2016] Groskopf, C. (2016). European politics is more polarized than ever, and these numbers prove it. *Quartz*.

[Grover and Leskovec, 2016] Grover, A. and Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855--864, New York, NY, USA. ACM. event-place: San Francisco, California, USA.

[Guerra et al., 2017] Guerra, P. C., Nalon, R., Assunção, R., and Jr, W. M. (2017). Antagonism Also Flows Through Retweets: The Impact of Out-of-Context Quotes

in Opinion Polarization Analysis. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017*, pages 536--539, Montréal, Québec, Canada. AAAI Press.

[Habgood-Coote, 2017] Habgood-Coote, J. (2017). The term 'fake news' is doing great harm. *The Conversation*.

[Hamilton et al., 2017a] Hamilton, W., Ying, Z., and Leskovec, J. (2017a). Inductive representation learning on large graphs.

[Hamilton et al., 2017b] Hamilton, W. L., Ying, R., and Leskovec, J. (2017b). Representation Learning on Graphs: Methods and Applications. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

[Hankes and Amend, 2018] Hankes, K. and Amend, A. (2018). The Alt-Right is Killing People.

[Hannak et al., 2013] Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring Personalization of Web Search. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 527--538, New York, NY, USA. ACM. event-place: Rio de Janeiro, Brazil.

[Hannak et al., 2014] Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, pages 305--318, New York, NY, USA. ACM. event-place: Vancouver, BC, Canada.

[Hatebase, 2018] Hatebase (2018). Hatebase.

[Ingram, 2018] Ingram, M. (2018). Most Americans say they have lost trust in the media. *Columbia Journalism Review*.

[Joel Kelly, 2017] Joel Kelly, B. (2017). Lauren Southern: The alt-right's Canadian dog whistler.

[Julia Angwin, 2017] Julia Angwin, H. G. (2017). Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children.

[Kloumann and Kleinberg, 2014] Kloumann, I. M. and Kleinberg, J. M. (2014). Community Membership Identification from Small Seed Sets. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1366--1375, New York, NY, USA. ACM.

[Know Your Meme, 2018] Know Your Meme (2018). Operation Google.

[Kumar and Shah, 2018] Kumar, S. and Shah, N. (2018). False Information on Web and Social Media: A Survey. *arXiv:1804.08559 [cs]*. arXiv: 1804.08559.

[Kutner, 2016] Kutner, M. (2016). Roosh V's journey from pickup artist to right-wing provocateur. *Newsweek*.

[Lazer et al., 2018] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094--1096. ISSN 0036-8075, 1095-9203.

[Lewandowsky et al., 2012] Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, 13(3):106--131. ISSN 1529-1006.

[Lewis, 2018] Lewis, R. (2018). Alternative influence: Broadcasting the reactionary right on YouTube. Technical report, Data and Society.

[Lianne and Simmonds, 2013] Lianne, C.-F. and Simmonds, H. (2013). Redefining Gatekeeping Theory For A Digital Generation. *The McMaster Journal of Communication*, 8.

[Liao and Fu, 2013] Liao, Q. V. and Fu, W.-T. (2013). Beyond the Filter Bubble: Interactive Effects of Perceived Threat and Topic Involvement on Selective Exposure to Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2359--2368, New York, NY, USA. ACM.

[Lloyd Parry, 2017] Lloyd Parry, R. (2017). Rohingya ethnic cleansing is fake news, says Burma army. *The Times*. ISSN 0140-0460.

[Lopez G., 2019] Lopez G., C. (2019). Stefan Molyneux is MAGA Twitter's favorite white nationalist.

[Lott, 2017] Lott, T. (2017). Jordan Peterson and the transgender wars.

[Magu et al., 2017] Magu, R., Joshi, K., and Luo, J. (2017). Detecting the Hate Code on Social Media. In *Eleventh International AAAI Conference on Web and Social Media*.

[Mann et al., 2019] Mann, A., Nguyen, K., and Gregory, K. (2019). 'Emperor Cottrell': Accused Christchurch shooter had celebrated rise of the Australian far-right.

[Marantz, 2017] Marantz, A. (2017). The Alt-Right Branding War Has Torn the Movement in Two. ISSN 0028-792X.

[McCauley and Moskalenko, 2008] McCauley, C. and Moskalenko, S. (2008). Mechanisms of Political Radicalization: Pathways Toward Terrorism. *Terrorism and Political Violence*, 20(3):415--433. ISSN 0954-6553.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. arXiv: 1301.3781.

[Mitchel et al., 2014] Mitchel, A., Eva Matsa, K., Gottfried, J., and Kiley, J. (2014). Political Polarization & Media Habits. Technical report, Pew Research Center.

[Mitchell and Page, 2015] Mitchell, A. and Page, D. (2015). State of News Media 2015. Technical report, Pew Research Center.

[Nagle, 2017] Nagle, A. (2017). *Kill All Normies: Online Culture Wars From 4Chan And Tumblr To Trump And The Alt-Right.* John Hunt Publishing. ISBN 978-1-78535-544-8.

[Newman, 2011] Newman, N. (2011). Mainstream media and the distribution of news in the age of social media. Technical report.

[Nic et al., 2018] Nic, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., and Nielsen, R. K. (2018). Reuters Institute Digital News Report 2018. Technical report, Reuters Institute for the Study of Journalism.

[NW et al., 2015] NW, . L. S., 800Washington, S., and Inquiries, D. U.-.-. |. M.-.-. |. F.-.-. |. M. (2015). Global Support for Principle of Free Expression, but Opposition to Some Forms of Speech.

[Ottoni et al., 2018] Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira Jr., W., and Almeida, V. (2018). Analyzing Right-wing YouTube Channels: Hate, Violence and Discrimination. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, pages 323--332, New York, NY, USA. ACM. event-place: Amsterdam, Netherlands.

[Papadamou et al., 2019] Papadamou, K., Papasavva, A., Zannettou, S., Blackburn, J., Kourtellis, N., Leontiadis, I., Stringhini, G., and Sirivianos, M. (2019). Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. *arXiv:1901.07046 [cs].* arXiv: 1901.07046.

[Pariser, 2011] Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You.* Penguin UK. ISBN 978-0-14-196992-3.

[Pennington et al., 2014] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532--1543, Doha, Qatar. Association for Computational Linguistics.

[Pew Research, 2018] Pew Research (2018). Trends and Facts on Newspapers. Technical report, Pew Research Center.

[PowerfulJRE,] PowerfulJRE. Jordan Peterson - Joe Rogan Experience #1208.

[PowerfulJRE,] PowerfulJRE. Sargon of Akkad - Joe Rogan Experience #979.

[Rainie et al., 2017] Rainie, H., Anderson, J. Q., and Albright, J. (2017). The future of free speech, trolls, anonymity and fake news online. Technical report, Pew Research Center Washington, DC.

[Ratkiewicz et al., 2011] Ratkiewicz, J., Conover, M. D., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and Tracking Political Abuse in Social Media. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *Proceedings of the Fifth International Conference on Weblogs and Social Media, ICWSM 2011*, Barcelona, Catalonia, Spain. AAAI Press.

[RedIceTV,] RedIceTV. NPI 2016 Panel Discussion: Jared Taylor, Peter Brimelow, Kevin MacDonald & Millennial Woes.

[Ribeiro et al., 2012] Ribeiro, B., Murai, a. F., and Towsley, D. (2012). Sampling directed graphs with random walks. In *2012 Proceedings IEEE INFOCOM*, pages 1692--1700.

[Ribeiro et al., 2010] Ribeiro, B., Wang, P., and Towsley, D. (2010). On Estimating Degree Distributions of Directed Graphs through Sampling. Technical Report UM-CS-2010-046, University of Massachusetts.

[Riloff et al., 2013] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013). Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704--714, Seattle, Washington, USA. Association for Computational Linguistics.

[Roose, 2019] Roose, K. (2019). The Making of a YouTube Radical. *The New York Times*. ISSN 0362-4331.

[Roosh V, 2016] Roosh V (2016). I Do Not Disavow Richard Spencer.

[Sabatini and Sarracino, 2017] Sabatini, F. and Sarracino, F. (2017). Online Networks and Subjective Well-Being. *Kyklos*, 70(3):456--480. ISSN 1467-6435.

[Schmidt and Wiegand, 2017] Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1--10, Valencia, Spain. Association for Computational Linguistics.

[Sellars, 2016] Sellars, A. (2016). Defining Hate Speech. SSRN Scholarly Paper ID 2882244, Social Science Research Network, Rochester, NY.

[Shafer, 2017] Shafer, J. (2017). The Case for Accrediting Breitbart. *POLITICO Magazine*.

[Shelton, 1998] Shelton, H. (1998). Joint Doctrine for Information Operations. Technical report JOINT-PUB-3-13, JOINT CHIEFS OF STAFF WASHINGTON DC.

[Shi et al., 2019] Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. *Nature Human Behaviour*, 3(4):329. ISSN 2397-3374.

[Shu et al., 2017] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1):22--36. ISSN 1931-0145.

[Solon, 2017] Solon, O. (2017). Google's bad week: YouTube loses millions as advertising row reaches US. *The Observer*. ISSN 0029-7712.

[Stein, 1986] Stein, E. (1986). History against Free Speech: The New German Law against the "Auschwitz": And Other: "Lies". *Michigan Law Review*, 85(2):277--324. ISSN 0026-2234.

[Sureka et al., 2010] Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. (2010). Mining YouTube to Discover Extremist Videos, Users and Hidden Communities. In Cheng, P.-J., Kan, M.-Y., Lam, W., and Nakov, P., editors, *Information Retrieval Technology*, Lecture Notes in Computer Science, pages 13--24. Springer Berlin Heidelberg.

[Tacopino, 2017] Tacopino, J. (2017). FBI clears Michael Flynn in probe linking him to Russia. *New York Post.*

[Tandoc Jr et al., 2018] Tandoc Jr, E. C., Lim, Z. W., and Ling, R. (2018). Defining "Fake News". *Digital Journalism*, 6(2):137--153. ISSN 2167-0811.

[Tardáguila et al., 2018] Tardáguila, C., Benevenuto, F., and Ortellado, P. (2018). Opinion | Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It. *The New York Times.* ISSN 0362-4331.

[The Rubin Report, 2018] The Rubin Report (2018). Eric Weinstein: The Future of The Intellectual Dark Web.

[Tong et al., 2008] Tong, H., Faloutsos, C., and Pan, J.-Y. (2008). Random walk with restart: fast solutions and applications. *Knowledge and Information Systems*, 14(3):327--346. ISSN 0219-3116.

[Trump, 2017] Trump, D. J. (2017). The Fake News Media has never been so wrong or so dirty. Purposely incorrect stories and phony sources to meet their agenda of hate. Sad!

[Tufekci, 2018] Tufekci, Z. (2018). Opinion | YouTube, the Great Radicalizer. *The New York Times.* ISSN 0362-4331.

[Twitter, 2019] Twitter (2019). Hateful conduct policy.

[Vicario et al., 2016] Vicario, M. D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., and Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113(3):554--559. ISSN 0027-8424, 1091-6490.

[Viswanath et al., 2015] Viswanath, B., Bashir, M. A., Zafar, M. B., Bouget, S., Guha, S., Gummadi, K. P., Kate, A., and Mislove, A. (2015). Strength in Numbers: Robust Tamper Detection in Crowd Computations. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, COSN '15, pages 113--124, New York, NY, USA. ACM. event-place: Palo Alto, California, USA.

[Warner and Hirschberg, 2012] Warner, W. and Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19--26, Montréal, Canada. Association for Computational Linguistics.

[Waseem et al., 2017] Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (2017). Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78--84, Vancouver, BC, Canada. Association for Computational Linguistics.

[Waseem and Hovy, 2016] Waseem, Z. and Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88--93, San Diego, California. Association for Computational Linguistics.

[Weiss and Winter, 2018] Weiss, B. and Winter, D. (2018). Opinion | Meet the Renegades of the Intellectual Dark Web. *The New York Times*. ISSN 0362-4331.

[Wong et al., 2016] Wong, F. M. F., Tan, C. W., Sen, S., and Chiang, M. (2016). Quantifying Political Leaning from Tweets, Retweets, and Retweeters. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2158--2172. ISSN 1041-4347.

[Yap, 2018] Yap, C. (2018). Duterte Decries 'Fake News' as Critics Warn of Media Crackdown. *Bloomberg*.

[Zannettou et al., 2018a] Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Suarez-Tangil, G. (2018a). On the Origins of Memes by Means of Fringe Web Communities. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 188--202, New York, NY, USA. ACM. event-place: Boston, MA, USA.

[Zannettou et al., 2017] Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtelris, N., Leontiadis, I., Sirivianos, M., Stringhini, G., and Blackburn, J. (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*, IMC '17, pages 405--417, New York, NY, USA. ACM. event-place: London, United Kingdom.

[Zannettou et al., 2019] Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2019). Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion*

Proceedings of The 2019 World Wide Web Conference, WWW '19, pages 218--226, New York, NY, USA. ACM. event-place: San Francisco, USA.

[Zannettou et al., 2018b] Zannettou, S., Sirivianos, M., Blackburn, J., and Kourtellis, N. (2018b). The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *arXiv:1804.03461 [cs]*. arXiv: 1804.03461.

[Ziegele et al., 2013] Ziegele, M., Johnen, M., Bickler, A., Jakobs, I., Setzer, T., and Schnauber, A. (2013). Male, Hale, Comments? Factors Influencing the Activity of Commenting Users on Online News Websites. *SCM Studies in Communication and Media*, 2(1):67--114. ISSN 2192-4007.

[Zollo et al., 2015] Zollo, F., Novak, P. K., Vicario, M. D., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Emotional Dynamics in the Age of Misinformation. *PLOS ONE*, 10(9):e0138740. ISSN 1932-6203.