Online Intergroup Conflict: How the dynamics of online communication drive extremism and violence between groups

John David Gallacher
University College
University of Oxford

Supervisors:
Prof Jonathan Bright[a], Dr Joss Wright[a], Dr Marc Heerdink[b],

[a]Oxford Internet Institute, University of Oxford, 1 St Giles Oxford, Oxford OX1 3JS
[b]Department of Social Psychology, University of Amsterdam, Postbus 15900, 1001 NK Amsterdam, The Netherlands

Thesis Submitted for the degree of Doctor of Philosophy
Hilary Term 2021

Word Count: 88,212

Online Intergroup Conflict: How the dynamics of online communication drive extremism and violence between groups

John David Gallacher
University College
Thesis Submitted for the degree of Doctor of Philosophy, Hilary Term 2021

Abstract

In the early days of the Internet, the social web was suggested to hold the power to reduce conflict, prejudice, and discrimination, by promoting connection between those with diverse backgrounds and experiences. However, this optimistic view has not materialised. Instead, evidence shows that Western societies are increasingly polarised and divided, and social media has been particularly implicated with the promotion of violent extremism and the spread of hate speech. The exact role social media plays in driving these processes remains poorly understood. This thesis investigates the mechanisms through which social media may foster and exacerbate intergroup conflict both online and offline, and explores different intergroup and intragroup dynamics which may drive this trend. This work firstly investigates the prevalence and impact of online communication between opposing groups on offline intergroup relations, demonstrating how intergroup communication does frequently occur online, but can be confrontational, hostile, and fails to improve intergroup relations, instead predicting future offline violence. Following this, we then develop a novel approach to detect online hate speech by leveraging datasets from multiple social media platforms in order to train natural language machine learning classifiers. We use these classifiers to explore the influence of extreme outgroup denigration in ingroup discussions on fringe social media platforms popular with the far-right. We demonstrate how exposure to online hate plays a role in shaping individual radicalisation trajectories, increasing the likelihood of offline hate crimes, and driving mutual radicalisation with opposing groups. Finally, we explore the effect of hostile manipulation from state actors on the dynamics of online conversations, presenting novel techniques to reveal how manipulation of these discussions increases the polarisation of online conversations from genuine users. Together these findings shed light on the mechanisms by which the Internet promotes intergroup conflict, extremism, and violence, providing new insight into initial steps that could be used to counter these effects.

Acknowledgements

I am grateful to the Engineering and Physical Sciences Research Council (EPSRC), and the University College Oxford Radcliffe Scholarship, for providing the postgraduate studentship which made this work possible. I would not have been able to carry out my research without this support.

I would like then to express my sincere gratitude to my supervisors who have supported and helped me over the last four years of research. Firstly, thanks to Marc Heerdink for his invaluable advice and patience as I have consistently tried to take on too many research projects, and for his thoughtful feedback on my work throughout this thesis. I also give huge thanks to Jonathan Bright and Joss Wright for taking on supervision of my thesis part-way through, and for helping me develop my research focus and ambition for large-scale analysis, and for their encouragement in getting my thesis to this stage.

I would also like to thank Miles Hewstone and the whole team at the Oxford Centre for the Study of Intergroup Conflict for their support and community during the first half of my studies. Finally, thanks to Rolf Fredheim and the team at the NATO StratCom COE in Riga, Latvia, for hosting me for the Summer of 2018, an experience which gave rise to the final chapter in this thesis, and for making me feel so welcome during the time I spent with them. Thank you also to Robin Dunbar, Maura Conway, Balazs Vedres, and Matt Williams for their feedback on my work at different stages of my DPhil.

I also extend my thanks to both the Cyber Security Centre for Doctoral Training and the Oxford Internet Institute for providing an excellent environment for my studies over the last four years, in particular to David Hobbs for putting up with my never-ending administrative questions and to Laura Maynard for providing her support.

Finally, my appreciation goes out to my family and friends for their encouragement and support throughout my studies. In particular thanks go to Annette for convincing me to embark on a DPhil in the first place and for supporting me through all of the ups and downs that this has entailed, I really couldn't have done it without you.

"We connect people. Period. That's why all the work we do in growth is justified. All the questionable contact importing practices. All the subtle language that helps people stay searchable by friends. All of the work we do to bring more communication in. The work we will likely have to do in China someday. All of it. So, we connect more people. That can be bad if they make it negative. Maybe it costs someone a life by exposing someone to bullies […] Maybe someone dies in a terrorist attack coordinated on our tools, and still we connect people".

Andrew Bosworth, VP Facebook, 2016:

Table of Contents

General Introduction

There are now over four billion Internet users, and this number is growing by hundreds of millions every year (Statista, 2020). Collectively, it is estimated that in 2020, humans spent over a million years on social media every single day (Kemp, 2020). As a result, social media platforms such as Twitter, Facebook, and YouTube have become the dominant public space for people to learn about and discuss a range of topics, from the mundane and humorous, to the complex and potentially divisive (Duggan, Smith, & Page, 2016). How this rise in online communication, and the growing importance of the online world, has changed how groups relate to each other online, and the consequences this has for both their online and offline intergroup relations, is a critical but yet unanswered question.

While Internet and social media use has skyrocketed, it has increased the opportunity for connection and contact between those who would otherwise be separated due to geographical, political, or social barriers. This initially led to an optimism that by increasing intergroup contact, social media could help reduce discrimination, prejudice, and conflict between opposing groups. Many social media platforms have promoted this optimistic 'tech utopian' view. Facebook, the world's largest social media platform, was set up with the stated goal of connecting people. It assumed that this connection could only be a good thing, and that in the long-term increased connection would trump any short-term damage and social harms (Bosworth, 2016; Newton, 2018).

More recently however, this optimism has started to fade, and today societies are more polarised and divided than ever. Perceptions of the impact of social media on society reflect this trend. Instead of being discussed as a tool which allows for positive social change and empowerment (Howard et al., 2011), the role of social media is now discussed in terms of the creation of ideological 'echo chambers' (Pariser, 2011; Sunstein, 2017), incubation of extremism (Ebner, 2020), or the promotion of intergroup conflict and political violence (Bail et al., 2018; Meleagrou-Hitchens & Kaderbhai, 2016).

Opinions are divided on the exact role that the Internet is playing in driving these negative social trends, and this requires further investigation. In particular there is a need to reconcile the idea that while increased intergroup contact can improve intergroup relations offline (Allport, 1954), this does not appear to be occurring naturally online. Initially, researchers resolved these contrasting

perspectives by focusing on the role that social media might play in creating echo chambers and 'filter bubbles' - networks of like-minded people who confirm each other's opinions instead of promoting critical thinking (e.g. Conover, Ratkiewicz, & Francisco, 2011; Pariser, 2011; Sunstein, 2017; Yardi & Boyd, 2010). This idea claims that intergroup relations online are not improving because social media is not leading to the expected increases in diverse social contact, and is instead reducing it.

Evidence does not support this echo chamber view however. Instead, more recent research suggests that people are exposed to more ideologically cross-cutting information online than ever (Bakshy, Messing, & Adamic, 2015; Bright, 2018), and opportunities for intergroup contact online are high (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015). Despite this, polarisation continues to increase (The Pew Research Center, 2017). This suggests that intergroup contact does occur online, but it is not leading to the promised benefits, and instead it may be making things worse (Bail et al., 2018). Additionally, the prevalence of abuse, intolerance, and hate speech online is increasing (Vidgen, Margetts, & Harris, 2020), potentially driving groups further apart (Williams, Burnap, Javed, Liu, & Ozalp, 2019). However, the exact role that this plays in exacerbating intergroup conflict, both online and offline, is unclear.

This continued debate highlights the complexities in studying the role of the online world on intergroup relations. It also demonstrates the need for interdisciplinary research which brings together inputs from social psychology, political science, communication studies, alongside new methodological approaches utilising advances in data science and artificial intelligence to tackle these questions (Lazer et al., 2009). In addition, processes of intergroup conflict online and offline are likely connected, so it is important to avoid thinking of the online environment as an isolated space and instead consider the interplay between online and offline environments in a more integrated way (Slater, 2002).

This thesis examines the role that social media plays in fostering and exacerbating intergroup conflict. As the reliance on Internet communications and social media as a means to stay connected continues to increase, understanding this relationship becomes ever more critical. This work is not an investigation of whether social media and the Internet cause polarisation within previously harmonious societies, but rather how social media use influences the relations between groups which are already in conflict, and in particular how it affects the individuals with the strongest outgroup prejudices within these groups. More specifically, we investigate the role of hostile intergroup contact, extreme outgroup denigration in ingroup discussions, and hostile manipulation from foreign powers,

on the intergroup relations of these groups online, and on their behaviour offline. Together, our research provide novel insight into how the Internet is driving increases in intergroup conflict, extremism and violence.

Thesis Overview

This thesis explores different intergroup and intragroup dynamics which may both explain and drive the role that social media plays in fostering and exacerbating intergroup conflict.

We first introduce the key topics with a review of the relevant literature in intergroup conflict, social polarisation, and extremism, both offline and online, and discuss the major motivations for this work along with important practical, methodological, and ethical considerations.

Over five separate research chapters, this thesis then investigates different aspects of how the Internet and social media may be contributing to increased intergroup conflict. This follows an integrated thesis approach, whereby all of these chapters are written in a manuscript format and can be read independently, although Chapters 3 and 4 rely on methods described in Chapter 2. Each chapter contains an introduction relevant to the set of questions it explores, and as such the opening literature review chapter covers these topics at a more abstract level to avoid repetition.

In Chapter 1, we explore the nature of online interactions between members of opposing political groups and how this affects group behaviour in the real world. Counter to predictions from technological optimists and contact theory proponents, we present evidence that greater intergroup interaction online does not appear to improve group relations but is instead associated with greater intergroup violence offline. These results highlight (i) the role of online communications in driving extremist violence, (ii) that intergroup contact does occur online, but can have negative impacts if quality is low, and (iii) that ingroup discussions are also important, and indeed themselves predictive of future violence in the real world.

In Chapter 2, we develop an automated hate speech classifier using natural language processing and machine learning in order to identify extreme outgroup denigration on large datasets of communication from across multiple social media platforms. The results demonstrate that by leveraging cross-platform data we can improve performance of automatic classification of online hate

and build a more flexible approach which can be easily updated in the future. This is important given the fast-changing nature on the online space, the evolving way that extremist groups use the Internet, and the shifting dynamics of language use itself. It is also practically beneficial to develop classification techniques which can be improved and adjusted easily as larger and more comprehensive training datasets become available. We also discuss the value and importance of considering extreme digital speech on a spectrum and building more nuanced classification schemes which go beyond binary distinctions of 'hateful' and 'clean' online communication.

The work in Chapter 3 uses this novel hate speech classifier to investigate how hate speech expressed by individual users on fringe social media platforms change over time. More specifically, we examine whether users arrive on these social platforms in order to express hate or if this develops through time spent on the platforms. We focus on the ingroup discussions of far-right groups, which the results in Chapter 1 show are associated with group extremism and violence. We present evidence that while many users join fringe far-right platforms with pre-existing prejudices and outgroup discriminatory behaviours, over time users move to adopt the wider outgroup denigration position of the group. This is likely due to high exposure to hateful content on the platforms, from other ingroup members, with hate speech spreading through the network through a process of 'social contagion'.

In Chapter 4 we expand on these results to look again at the connections between the online and offline world as in Chapter 1, but this time using continuous longitudinal data from an entire fringe social media platform popular with the extreme far-right, Gab. We explore the relationships between online hate speech, offline hate crime and terror attacks, and their effects on intergroup conflict and mutual radicalisation. Our results demonstrate a cyclical relationship between opposing extremist groups, whereby online hate from far-right groups both precedes offline violence from these same groups, and spikes following offline violence from opposing Islamic extremist groups. Additionally, far-right Islamophobic violence offline is also followed by increased online interest in Islamic extremist topics. Together, these findings show that the Internet, and specifically hate speech, play a potential key role in driving mutual radicalisation.

Chapter 5 focuses on another factor which could be driving intergroup conflict – the hostile manipulation of online environments from state actors. We present a novel approach for measuring the polarisation within Twitter networks, and novel applications of causal impact modelling, combined with text analysis measures, to measure the effect of hostile activity from the Russian state on the

conversations of genuine users on Reddit. We demonstrate that online activity from the Russian Internet Research Agency had a measurable effect on the subsequent conversations of genuine users and was associated with increased polarisation of controversial topics on Twitter, and led to increases in the level of intergroup antagonism and hostility on Reddit. By developing methods to measure the impact of information operations in online conversations and demonstrating a measurable effect on genuine conversations, this chapter aims to provides an important step in developing effective countermeasures against this type of hostile activity.

We close the thesis with a discussion on how the effects detected in the different chapters may work in combination, how they may each contribute to an increasing cycle of intergroup conflict, which is both mediated by, and reliant on, social media and the Internet. We conclude by discussing what this research teaches us about the likely success of potential mitigation efforts and how these negative effects of intergroup conflict can be prevented, and finish with suggestions for how this research can be taken forward.

Author contributions and publication status

The work presented in this thesis is my own. My supervisors contributed their ideas and feedback throughout the planning, data analysis, and manuscript preparation. More specifically, Miles Hewstone provided guidance on Chapter 1, Marc Heerdink on Chapters 1, 2, 4 and 5, Jonathan Bright and Joss Wright on Chapters 2, 3 and 4.

Chapters 1 and 5 have been published with two of my supervisors as co-authors. Signed contribution statements from them confirming that I was the primary author of these papers, along with the specific contributions of each co-author, are included at the end of these two chapters.

Chapter 1 has been published in the journal 'Social Media + Society'
Gallacher, J. D., Heerdink, M. W., & Hewstone, M., *(2021)* Online contact between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society.* pp. 1–19
https://journals.sagepub.com/doi/pdf/10.1177/2056305120984445

Chapter 5 has been published in the journal 'Defence Strategic Communications'

Gallacher, J. D., & Heerdink, M. W., (2019) Measuring the effect of Russian Internet Research Agency information operations in online conversations. *Defence Strategic Communication., vol. 6, p.155:198*
https://www.stratcomcoe.org/jd-gallacher-m-w-heerdink-measuring-effect-russian-internet-research-agency-information-operations

An earlier version of Chapter 3 was presented at the 2020 European Consortium for Political Research General (ECPR) Conference and published in the conference proceedings.
Gallacher, J, D. (2020) The ontogeny of online hate speech: Do social media platforms drive increased hate or reflect existing prejudices? *Paper prepared for presentation at the ECPR General Conference, 25-28 September 2020*
https://gc.ecpr.eu/Filestore/paperproposal/988e78ed-b4e1-4378-a260-092fd83ae994.pdf

Chapter 2 was submitted to the 2020 Trust and Truth Online Conference, and following useful feedback from reviewers will be re-submitted to a similar conference in the near future.

Chapters 3 and 4 will be submitted to peer-reviewed journals in the field of applied social sciences in due course.

References

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., … Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.17605/OSF.IO/4YGUX

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bright, J. (2018). Explaining the emergence of echo chambers on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33. https://doi.org/10.2139/ssrn.2839728

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *133*(26), 89–96. https://doi.org/10.1021/ja202932e

Duggan, M., Smith, A., & Page, D. (2016). *The political environment on social media*. Retrieved from http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Ebner, J. (2020). *Going dark: The secret social lives of extremists*. Bloomsbury Publishing.

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). *Opening closed regimes: What was the role of social media during the Arab spring? Project on Information Technology & Political Islam*. https://doi.org/10.2139/ssrn.2595096

Kemp, S. (2020). Social media users pass 4 billion: Digital 2020 October statshot report. Retrieved November 11, 2020, from https://blog.hootsuite.com/social-media-users-pass-4-billion/

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., … Alstyne, M. Van. (2009). Computational Social Science. *Science*, *323*, 721–724.

Meleagrou-Hitchens, A., & Kaderbhai, N. (2016). *Research perspectives on online radicalization*. VOX-Pol Network of Excellence. Retrieved from https://icsr.info/wp-content/uploads/2017/05/ICSR-Paper_Research-Perspectives-on-Online-Radicalisation-A-Literature-Review-2006-2016.pdf

Newton, C. (2018). In a leaked memo, Facebook executive describes the consequences of its growth-at-all-costs mentality. *The Verge*, (June 2018). Retrieved from https://www.theverge.com/2018/3/29/17178086/facebook-growth-memo-leak-boz-andrew-bosworth

Pariser, E. (2011). *The Filtter Bubble: What the internet is hiding from you*. New York, New York, USA: The Penguin Press.

Slater, D. (2002). Social relationships and identity online and offline. In *Handbook of New Media: Social Shaping and Consequences of ICTs* (pp. 533–546). SAGE. https://doi.org/10.4135/9781848608245.n38

Statista. (2020). Number of internet users worldwide. Retrieved November 11, 2020, from https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/

Sunstein, C. R. (2017). *#Republic: divided democracy in the age of social media*. Princeton, New Jersey, United States: Princeton University Press.

The Pew Research Center. (2017). *The partisan divide on political values grows even wider*. Retrieved from https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/

Vidgen, B., Margetts, H., & Harris, A. (2020). *How much online abuse is there? A systematic review of evidence for the UK*. Retrieved from https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-monitor

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/0270467610380011

Literature Review

This thesis investigates the role of the Internet, social media in particular, in promoting and exacerbating intergroup conflict.

Researching this question requires an understanding of, and input from, and number of different fields. This literature review first considers some of the traditional literature in social psychology on social identities and intergroup relations, which underpins much of our understanding of why groups come into conflict. This is followed by a discussion of extreme intergroup conflict and how opposition between groups can turn into violence. Subsequently, we outline the current state of computational social science research into the role of the Internet in driving effects of social polarisation and conflict, and how extremist groups use the Internet to purposefully drive division.

We follow this with a discussion of the challenges of defining and detecting online hate speech, and of recent research on its prevalence and impact on online spaces. We outline how the Internet has been used since its inception by hateful extremist groups, in particular recent developments in its use by far-right extremist groups—on which most of this work focuses—on both mainstream and fringe platforms.

The literature review finishes with a discussion of both the opportunities provided and challenges presented by the dramatic increase in the prevalence of social media data for research into online intergroup conflict, along with recent advances in computational methods which allow for new approaches to study this important topic.

Intergroup relations and conflict

Offline social interactions, and the impact that they have on behaviours and attitudes at both the group and individual level, have been studied in depth for decades. One of the most successful results of this body of research is social identity theory (Tajfel, 1974; Tajfel & Turner, 1979; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), which explains intergroup behaviour through the concepts of social identity theory and self-categorization. This theory suggests that individuals' behaviours are partly driven by their social identity: the concept they have of themselves, which is derived from attachment to the social groups they are a part of. According to social identity theory, when group memberships are highly salient then group identities override individual identities, and people's attitudes, emotions, and behaviours are driven by this group membership rather than individual characteristics (Abrams & Hogg, 2004; Tajfel & Turner, 1979). Because this social identity forms an important part of both how people view themselves individually, and how they behave in relation to others, it can therefore help to explain both intergroup and intragroup behaviour (Hewstone, Stroebe, & Jonas, 2016), and should be included in our understanding of intergroup relations and conflict.

The formation and categorisation of social groups

When considering these group dynamics, the groups themselves do not need to be formally demarcated. A group exists whenever two or more individuals define themselves as members of this group (Nijstad & van Knippenberg, 2016). This can therefore encompass groups that span many different aspects, e.g. religion, nationality, friendship, or politics, amongst others. These do not need to include any objective characteristic of the group, but rather they are defined by fact that members of the group view themselves through this common identity of belonging to the group. Once defined, these groups may then persist beyond the presence of the individual group members, and the membership of the groups themselves may be consistently changing over time, while the group itself remains (Simmel, 1898). This definition of the ingroup automatically introduces another, for almost every ingroup, there also exists an outgroup encompassing all those who do not belong to this ingroup. These outgroups can then be further subdivided along lines of common characteristics.

While much of this thesis rests on the social psychology literature surrounding intergroup relations, work from the field of sociology is also relevant to the definitions of group identity and intergroup conflict. Early work on social group memberships discusses how they are not mutually exclusive, and individuals can be members of multiple groups at the same time, and these groups may be highly

disparate or diverse. The exact mix of social groups to which one belongs is suggested to be an important factor in driving that person's distinct individuality, which has been suggested to be the product of the intersection of the multiple group memberships to which they are a part (Simmel, 1972). Participation and interaction in different social groups both influences the nature of these groups, and also heightens an individual's opportunities for connection, and this connection in return drives their own personal traits (Chayko, 2015; Moreland & Levine, 1982).

Competition or antagonism between groups has long been considered an unavoidable aspect of group formation, and may be essential in some cases to a healthy society (Simmel, 1904) as it strengthens ingroup bonds, reduces ingroup dissent, and establishes the distinctiveness of the various social groups which in turn helps to increase the stability of the overall social structure (Coser, 1956). This positive effect is only observed up to a point however, and groups which are engaged in continual conflict with others tend to become increasingly intolerant, overly restricting any internal dissent, and responding with violence to outgroup actions (Stein, 1976).

Within social psychology, self-categorisation theory helps explain how the process of categorising oneself as a group member forms a social identity and brings about behaviours and influence from both the ingroup and the outgroup (Turner & Oakes, 1986). These behaviours include group polarisation (the tendency of groups to take more extreme positions than the sum of its individual group members, Moscovici & Zavalloni, 1969), majority-minority influence (majority groups tend to exert normative social pressures, while minority groups must rely on informational influence, Turner, 1991) and intergroup discrimination behaviours, such as ingroup bias (Sherif et al., 1961). This ingroup bias is represented in behaviours and attitudes which favour the ingroup over the outgroup, i.e. treating the ingroup more positively or disadvantaging the outgroup. Evidence has shown that even arbitrary distinctions between groups, such as preferences for certain paintings, can trigger favouritism for one's own group at the expense of outgroups (Tajfel, Billig, Bundy, & Flament, 1971). This has been found to occur even when it means sacrificing personal gain (Sidanius, Haley, Molina, & Pratto, 2007).

Theories of intergroup dynamics of prejudice and discrimination

In addition to ingroup bias, the psychological separation of society into ingroups of 'us' and outgroups of 'them' can lead to a number of other wider undesirable outcomes, including derogation of the outgroup. Outgroup derogation is typically characterised by two factors: prejudice (a negative, unfair,

or unreasonable opinion, attitude, or feeling towards a group that is not justified by the facts) and discrimination (unjustified actions and treatments of others based on the groups to which they are perceived to belong). There is evidence that the degree to which people see themselves in terms of group membership, and their degree of value and emotional attachment to the group, is associated with greater prejudice and discrimination against the outgroup (Gagnon & Bourhis, 1996). This effect, however, is stronger for ingroup favouritism—such as rewarding ingroup members more favourably or working harder towards ingroup goals—than outgroup derogation, and therefore cannot fully explain intergroup conflict (Mummendey, Otten, Berger, & Kessler, 2000). In addition, attitudes towards social outgroups are often correlated. Those who have negative attitudes towards one outgroup tend to have negative attitudes towards other outgroups (Allport, 1954). While this observation initially prompted research into personality factors (Adorno, Frenkel-Brunswik, Levinson, & Sanford, 1950), more recent evidence has instead pointed towards social interaction and social identity effects as larger driving factors in prejudiced behaviours (Brown, 2010). As prejudice and discrimination involve making negative evaluations of others based on their membership of groups (Spears & Tausch, 2016), investigating the intergroup aspects of outgroup derogation is essential.

To explain prejudices and discriminatory behaviours, theories of intergroup conflict have studied the relations between the ingroup and outgroup, and in particular the threats, real or imagined, that the outgroup presents to the ingroup. Prejudice and discrimination are presented as ways of coping with these threats. Realistic conflict theory explains prejudice and discrimination as reflecting real conflicts of interests between groups who are in competition for valued and scare resources (Sherif et al., 1961). In this theory the negative attitudes arise as groups view the competition over resources as having a zero-sum outcome; what one group wins the other loses (Jackson, 1993). Conversely, intergroup threat theory focuses on perceived threat rather than actual threat (Stephan & Stephan, 2000; Stephan, Ybarra, & Morrison, 2009). These perceived threats to group existence can span across multiple dimensions and are not just related to physical resources but can include threats to group status and self-esteem (Abrams & Hogg, 1988), threats to group values (Branscombe & Wann, 1994), or threats to group distinctiveness (Brewer, 1991). Within this theory, majority groups tend to develop stereotypes about minority groups based on the observation of a small group of deviants among them. Minority groups that experience discrimination from majority groups often feel threatened in turn, because they view this prejudice as irrational or unjustified, leading to anxiety about interacting with the outgroup (Stephan & Stephan, 1985). Together these theories of intergroup relations have helped explain a number of intergroup conflicts, from Muslim-Hindu relations in India (Tausch, Hewstone, &

Roy, 2009), majority-minority group relations in the Netherlands (González, Verkuyten, Weesie, & Poppe, 2008), and Islamophobic attitudes in Europe (Uenal, 2016).

While these theories of intergroup relations are primarily situated in the offline environment, they provide a solid theoretical grounding for the study on intergroup relations online and demonstrate the validity and importance in studying social effects at the group level rather than the individual level.

How intergroup conflict can lead to extremism and radicalisation

While intergroup conflict is a common phenomenon across many areas of society, its damaging effects can vary depending on the groups involved and the level of conflict experienced. For example, intergroup conflict between sports teams can be beneficial to competition, but detrimental if it descends into hooligan violence. Similarly, distinctive ethnic groups can help foster tradition and create intragroup support networks, but when taken to the extreme can also lead to ethnic conflict and even, in the worst cases, ethnic genocide. There is a relationship between these levels of conflict, and countries with high levels of ethnic fractionalization also have higher of intergroup violence (Montalvo & Reynal-Querol, 2005). Understanding how and when intergroup conflict can descend into violence is therefore vitally important, and theories of intergroup relations and intergroup dynamics can help explain this process.

In the most severe cases intergroup conflict can lead to extremism; the belief that the survival of the ingroup is inseparable from some kind of direct or offensive action against the outgroup (Berger, 2018a). Extremism can be brought about through a process of radicalisation, where individuals make increased preparation and commitment to intergroup conflict, driven by a perceived threat to the ingroup arising from the outgroup (McCauley & Moskalenko, 2008). The process of radicalising to a position of extremism can therefore be considered an extreme consequence of group creation. An extreme form of premeditated intergroup violence is terrorism, which can be defined as *"violence – or the threat of violence – used and directed in pursuit of, or in service of, a political aim"* (Hoffman, 2006, pp. 2–3). Individuals are vulnerable to engaging in ideologically motivated violence when they experience changes to their perception of social identities, in particular the prioritisation of the extremist group membership over all others (Morgan, 2001), and the replacement of their personal beliefs, attitudes and behaviours with those of the group norms (Stahelski, 2005). This is supported by

brain-imaging results from supporters of radical extremist groups, who presented lower activity in brain areas associated with deliberative cognitive control and cost-benefit analysis when performing decision making on group-level vs individually relevant topics (Hamid et al., 2019). This suggests that for radicalised individuals highly ingroup normative decisions are reached at the level of the group rather than in terms of direct individual benefit.

Numerous models have looked to describe this process of radicalisation, and the social aspects of ingroup group dynamics and behaviour have consistently been highlighted as being an important driving factor. Intense personal identification with the ingroup is suggested to facilitate members individually adopting extreme ideology which justifies violence, if this is shared with other ingroup members, even if this ideology would otherwise breach more general moral or social norms (Kruglanski et al., 2014). Additionally, becoming motivated to pursue this violence provides a route for members to achieve ingroup authority and personal significance in relation to the group (Kruglanski et al., 2014). Furthermore, furthering existing intergroup conflict is suggested to justify perceptions of ingroup superiority and relative significance (Kruglanski et al., 2013), while the fusion of personal and group identities (Swann, Gómez, Seyle, Morales, & Huici, 2009; Swann, Jetten, Gómez, Whitehouse, & Bastian, 2012) is proposed to generate a collective sense of invincibility and 'special destiny' (Atran, Sheikh, & Gomez, 2014), which promotes the adoption of violence.

In extreme situations of intergroup conflict, and in particular shared traumatic experiences with other ingroup members, this can lead the strength of the relationship between ingroup members to exceed that of familial relationships (Whitehouse, McQuinn, Buhrmester, & Swann, 2014), and this strength can in turn motivate otherwise irrational actions such as self-sacrifice for the group or commitment to further violence (Whitehouse et al., 2017). This does not mean that personal identity is lost and replaced with group identity as has been historically suggested (e.g. Kaplan, 1978; Post, 1984), but rather that extreme behaviour can arise out of an individual's heightened awareness of both personal and group identity and how these relate to one another (Swann et al., 2009). Together, this highlights the importance of studying the role of social identify and group dynamics in processes of radicalisation.

The role of intergroup emotions

These processes of extremism and radicalisation need not be entirely rational, and also implicitly include elements of emotional processing. Intergroup emotions theory argues that if people define themselves in terms of a social identity, they experience emotions not as individuals, but as group members (Mackie, Devos, & Smith, 2000; Smith, 1993). In this way, individuals can experience emotional responses vicariously on behalf of the group. This group emotion approach can explain more extreme and intense forms of prejudice, such as outgroup derogation, because powerful group-level emotions such as anger, fear, or contempt, can lead to derogatory perceptions of the outgroup (DeSteno, Dasgupta, Bartlett, & Cajdric, 2004). If these emotions carry across to outgroup dehumanisation (Haslam, 2006) then this in turn can be predictive of support of violent behaviours and aggressive actions against the outgroup (Kteily, Bruneau, Waytz, & Cotterill, 2015). For example, an individual who identifies as a member of group A, and observes attacks on group A from group B, may experience a negative emotional response as if they were attacked personally. Given this threat to the ingroup, the individual may respond with derogation back against group B to preserve positive intergroup comparison. This derogation can be presented by verbal or physical attacks on the outgroup.

Models of intergroup extremism

These ideas of intergroup dynamics run counter to the early literature on extreme acts of violence, which argued that people commit acts of extreme inhumanity due to a lack of awareness or control (Haney, Banks, & Zimbardo, 1973; Milgram, 1963). Instead, growing evidence suggests that individuals commit these acts because they believe what they are doing is right, morally correct, and to be celebrated (Reicher, Haslam, & Rath, 2008). Reicher et al. (2008) propose a five-step social identity model to explain why groups display hostility to one another and how inhumane acts against an outgroup can come to be celebrated as 'right': i) identification of an ingroup; ii) creation of the outgroup; iii) representation of the outgroup as endangering ingroup identity; iv) championing of the ingroup as (uniquely) good; and v) embracing the eradication of the outgroup as necessary to defend ingroup virtue. This highlights that the creation and definition of the ingroup is just as important as the creation of the outgroup – there can be no 'them' without an 'us'. The final stage of this process is also in agreement with the above definition of extremism (Berger, 2018a).

Building on this model, Smith, Blackwood, & Thomas (2019), propose a group socialisation model of radicalisation where group members develop shared identification with a set of radical norms,

including outgroup hate, through social interactions that leverage their shared perceptions and experiences, whilst also changing the nature of the group as a whole. In this model, individuals gradually adopt norms, ideologies and customs following their involvement within a certain social group (Wiktorowicz, 2005). Shifts towards violence are a function of changes in the content of group discussions and social interactions, i.e. people's radicalisation to violence is inseparable from the social context in which their social interactions take place. Similar to Reicher, Haslam and Rath (2008), they produce a multistage model whereby individuals (i) initially recognise some kind of grievance or injustice, often in relation to an outgroup. This grievance, and the related thoughts, feelings and beliefs about the normative conflict are then (ii) communicated with others. During this social interaction, the opportunity is given for others to (iii) validate this grievance as "real and true," and for members of the group to reach agreement. When ingroup consensus on outgroup stereotypes is reached, it can then (iv) be used by ingroup members to mobilize themselves and justify and legitimize intergroup violence, discrimination, dehumanisation, and conflict. The increase in shared definitions of these ideas creates a sense of shared understanding and shared social norms, which are both a function of, and driven by, individual members social identities.

Importantly, the process of intragroup socialisation is key in both of these models. During the process of socialisation, new group members learn the norms of the group and develop the shared expectations and belief system about how, and how not, to behave (Levine & Moreland, 1994; Moreland & Levine, 1982). Therefore, if the norms of the ingroup include extreme derogation towards the outgroup, then these can be learned through socialisation. Ingroup norms are particularly powerful in driving members to express outgroup prejudice and discrimination in social situations, and public expressions of prejudice are highly correlated with the social approval of that particular expression (Crandall, Eshleman, & O'Brien, 2002). Similarly, outgroup dehumanisation is associated with the view that these outgroups contravene ingroup norms (Kteily et al., 2015). This has been shown in practice in apartheid South Africa where racial prejudice among white South Africans was expressed more due to conformity to social norms than to personality factors or individual beliefs (Pettigrew, 1958). Importantly, this relationship also holds in reverse, and when social norms overtly sanction prejudice, outgroup discrimination becomes less prevalent (Dovidio & Gaertner, 1991).

In this way, people inflict harm on others, and in particular on other groups, not because they are unaware of what they are doing, but because they believe what they are doing is right. This doesn't happen spontaneously, but because of group identification and the existence of dangerous group

norms, which can be driven by those in leadership positions within the group (Haslam, Reicher, & Bavel, 2019). Similar arguments have been made from the study of terrorism; Sageman (2004) argues that radicalisation results from interactions with others in their social circle, and immersion in a social group which reinforces an individual's extremist beliefs. As a consequence, in order to understand violent extremism, it is therefore key to understand the group processes involved, including both the intergroup and intragroup dynamics, and how these lead to intergroup conflict, prejudice, discrimination, and violence.

Online polarisation and its effect on intergroup conflict

The Internet has dramatically changed how groups communicate and interact. Concurrent with the Internet becoming a central part to people's social lives at the turn of the century, there has also been a dramatic rise of populist politics across Western democracies as well as an increase in political polarisation and political violence (Mudde, 2007; Mudde, 2019). Outside of Western environments, similar trends have been observed in countries such as Myanmar, Sri Lanka, and India (Rani, 2018; Samaratunge & Hattotuwa, 2014; Stecklow, 2018). This has led to a wealth of research and conjecture into the role that the Internet, and social media in particular, may be playing in driving increases in intergroup conflict and extremism, however the exact nature of this relationship remains to be understood.

Mechanisms of Social Polarisation

Typically, this increase has been framed in terms of social polarisation, the process of increased segregation into distinct social groups, separated along racial, economic, political, religious or other lines (Castree, Kitchen, & Rogers, 2013). In particular, much research into online social polarisation has focused on segregation along political lines, leading to individuals endorsing more extreme politically ideological positions following discussions with other ingroup members (ideological polarisation; Turner, Davidson, & Hogg, 1990), or may result in an increased dislike of outgroup members without a change in issue position (affective polarisation; Mason, 2015). Greater polarisation is closely linked to intergroup conflict, it decreases the desire for compromise (Nicholson, 2012), and a lack of cross-cutting social identities is associated with an increased risk of political violence (Mason, 2018). Evidence suggests that both ideological and affective polarisation have increased in tandem with the rise in popularity of social media and the Internet, and a large and growing proportion of society has a negative attitude towards those who identify with the opposing political party, and they

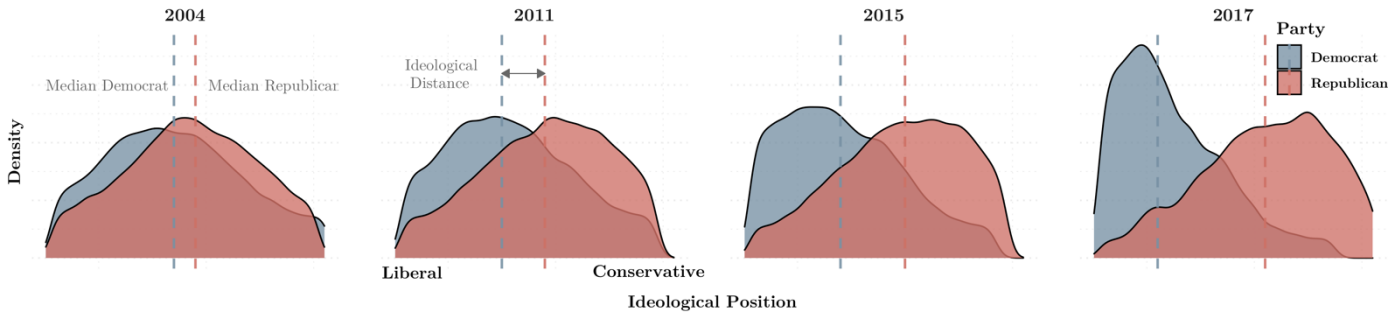


Figure 1 – The distribution of political values from Republican (red) and Democrat (blue) partisans in the USA, from more liberal to the left to more conservative to the right, illustrating increases in political polarisation between 2004 and 2017. Median positions for each party are represented by dotted lines. Data from the Pew Research Centre, 2017.

perceive a growing distance in opinions across political lines (Iyengar, Sood, & Lelkes, 2012; Lelkes, 2016). This trend is illustrated in Figure 1 with the growing ideological distance between Republicans and Democrats in the US since 2004 (The Pew Research Center, 2017). This trend was measured through a consistent set of 10 questions across all years relating to political values (e.g. the role of government regulation on private business, the importance of environmental protection, the importance of racial equality, the role of immigration etc), and the distribution of opinions across self-reported Republicans and Democrats was compared. Overall, the median Republican is more conservative and the median Democrat more liberal today than in 2004, with less overlap in the political values of the two groups (Heltzel & Laurin, 2020). This polarisation also has an affective component, for example in 1994, approximately 20% of those affiliated with a political party in the US had 'very unfavourable' views about members of the other party, while in 2016 this number had risen to over 55% (The Pew Research Center, 2017). Today in both the US and the UK individuals are more likely to consider their own political party as more trustworthy, open, and honest (Duffy, Hewlett, McCrae, & Hall, 2019).

Growing evidence supports a link between social media use and polarisation. Facebook users report more biased evaluations about outgroups than non-Facebook users, this effect is largest in those who use the site the most, and the effect is driven more by negativity towards the outgroup than positivity towards one's ingroup (Settle, 2018). Additionally, there is experimental evidence that reducing social media usage through deactivation of one's Facebook account can reduce polarity of views on political issues (Allcott, Braghieri, Eichmeyer, & Gentzkow, 2019). Although most of these studies have focused on Western environments, there is evidence that these effects also occur more broadly; a study of social media use in South Korea concluded that social media indirectly contributes to polarisation

through increased political engagement (Lee, Shin, & Hong, 2018). Understanding the role that social media use plays in driving polarisation is therefore vital.

The role of echo chambers in driving online polarisation

One pervasive argument for why the Internet, and social media in particular, can lead to increased political polarisation is through the creation of 'echo chambers'—networks of like-minded people who confirm each other's opinions instead of promoting critical thinking (Conover, Ratkiewicz, & Francisco, 2011; Yardi & Boyd, 2010). One firm proponent of this theory, Sunstein (2018), argues that social media platforms allow ingroup members to connect with one another to a greater extent than is possible offline due to geographical constraints, and therefore create more cohesive and ideologically aligned communities. This theory predicts that in these online environments, individuals only interact with those with whom they agree, and are therefore only exposed to information that reinforces their existing opinions, while being isolated from those with different or opposing views. This builds on social psychological theories advocating that individual attitudes are a function of the information that is available in their immediate environment (Hinsz & Davis, 1984; Vinokur & Burstein, 1974). Social media platforms may also actively promote echo chambers through effects of algorithmic filtering, which pushes content most likely in line with users' existing opinions and demote counter-attitudinal material (Noble, 2018). This creates 'filter bubbles' which further drive the consumption of differential information across ideological lines, thereby increasing both ideological and affective polarisation (Pariser, 2011).

Some evidence supports the echo chamber theory of social media use. Firstly, users who are more similar to one another are more likely to connect online and form communities through a process of homophily (McPherson, Smith-Lovin, & Cook, 2001), leading to situations where users online connections are homogenous (Colleoni, Rozza, & Arvidsson, 2014; Conover et al., 2011; Quattrociocchi, Caldarelli, & Scala, 2014; Williams, McMurray, Kurz, & Hugo-Lambert, 2015). Secondly, through processes of confirmation bias and selective exposure, users are motivated to seek out information which supports their existing opinions, while avoiding that which challenges their opinions (Bakshy, Messing, & Adamic, 2015; Knobloch-Westerwick, 2015; Wason, 1960). This leads to situations where false information that confirms users' pre-existing beliefs is clicked on more often than true, but counter-attitudinal, information (Garrett, Weeks, & Neo, 2016). Finally, social media recommender systems (automatic tools which recommend media content such as news articles, e-commerce products, etc to users) have been shown to increasingly promote greater levels of polarising

extremist material (Reed, Whittaker, Votta, & Looney, 2019). For example, YouTube's automatic system for playing videos has been shown to promote gradually more extreme material, starting on more benign topics, before progressively promoting more divisive and inflammatory content (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015; Ribeiro, Ottoni, West, Almeida, & Meira, 2019; Tufekci, 2018).

The echo chamber theory of online polarisation is attractive partly because it presents a simple solution to the problem. If polarisation is caused by only experiencing information which reinforces our existing views, then by 'bursting' the echo chamber and introducing more counter-attitudinal content then polarisation can be avoided, mitigated or even reversed (e.g. Cambre, Klemmer, & Kulkarni, 2017; Cota, Ferreira, Pastor-Satorras, & Starnini, 2019; Liao & Fu, 2014). Exposure to diverse views and opinions is considered necessary for healthy social debate, and this idea has support in the contact hypothesis (Allport, 1954), one of the most successful theories in social psychology. This theory suggests that direct contact and communication between members of different groups provides one of the best ways to improve relations between these groups, reduce conflict, and increase social cohesion. Intergroup contact has been shown to effective in reducing prejudice and discrimination by numerous studies in a wide range of offline conflict situations, it improves perceptions of social distance, and thereby also reducing polarisation (Rupert Brown & Hewstone, 2005; Davies, Tropp, Aron, Pettigrew, & Wright, 2011; Pettigrew & Tropp, 2006; Ramiah & Hewstone, 2013).

Unfortunately, however, the echo chamber theory has not held up fully to scrutiny. The observed increases in social polarisation have been highest amongst those who use the Internet the least (Boxell, Gentzkow, & Shapiro, 2017), raising questions of whether this effect is truly driven by the online world. Additionally, evidence suggests that online echo chambers may not be forming as often as first expected. While most political exchanges on social media take place among people with similar ideas, cross-cutting interactions are more frequent that commonly believed, forming up to 30% of interactions (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015). As such, social media users are exposed to more diverse opinions online than through other traditional types of media (Barnidge, 2017) and more than they would select purely based on choice (Bakshy et al., 2015). This may be partly explained by the fact that users of social media platforms are exposed to information shared by weak ties (friends of friends), who are more likely to hold counter-attitudinal views compared to one's immediate contacts (Barberá et al., 2015; Granovetter, 1973). There is also evidence of greater

connection between moderates of opposing groups than moderates and extremes within the same ideological groups (Bright, 2018), and substantial diversity of discussion even in the most extreme online spaces (Bright, Marchal, Ganesh, & Rudinac, 2020).

The nature of online intergroup contact

However, although exposure to cross-cutting information may be occurring more than originally thought on social media, it does not appear to be having a positive effect on intergroup relations. In fact, artificial exposure to opposing views online can increase polarisation rather than decrease it, possibly caused by the exposure to other opinions highlighting areas of disagreement and further entrenching ones' opinion (Bail et al., 2018). Similarly, natural exposure to political disagreement in online settings is shown to increase political polarization (Suhay, Bello-Pardo, & Maurer, 2018). This is potentially due to the nature of the online environment as online intergroup communication is often hostile (Kumar, Hamilton, Leskovec, & Jurafsky, 2018), and social media users report that interacting with outgroup members online is a stressful and frustrating experience (Duggan, Smith, & Page, 2016). The reason for this apparent online intergroup hostility, and the effect that hostile intergroup contact online has on real-world group relations remains to be understood.

One reason for this negative experience of online intergroup contact may be that the online world is a hyper-partisan political environment, and political identities may be particularly likely to cause intergroup friction (Monroe & Hankin, 2000; White, 2001). Most social media users report that they see at least "a little" political content when they log onto social media (94% on Facebook and 89% on Twitter; Duggan & Smith, 2016). Many of the features of social media platforms make it easy to infer the political and social identities of other users, therefore making group boundaries highly salient (Settle, 2018). These features include linguistic traits and specific ingroup language, and visual cues to group membership (Bruchmann, Koopmann-Holm, & Scherer, 2018; Rule & Ambady, 2010). When these highly salient group identities are combined with an absence of the interpersonal information which would typically occur in offline communication, social media users tend to form impressions of each other based on group memberships rather than individual characteristics (Lea & Spears, 1991; Postmes, Spears, & Lea, 1998). Importantly, when social media users infer the group identities of other users in this way, they are found broadly accurate in doing so even when this is based on relatively little information, or information which is not directly group relevant (Settle, 2018). Users do however typically overestimate the extremity of the positions of others within these inferred

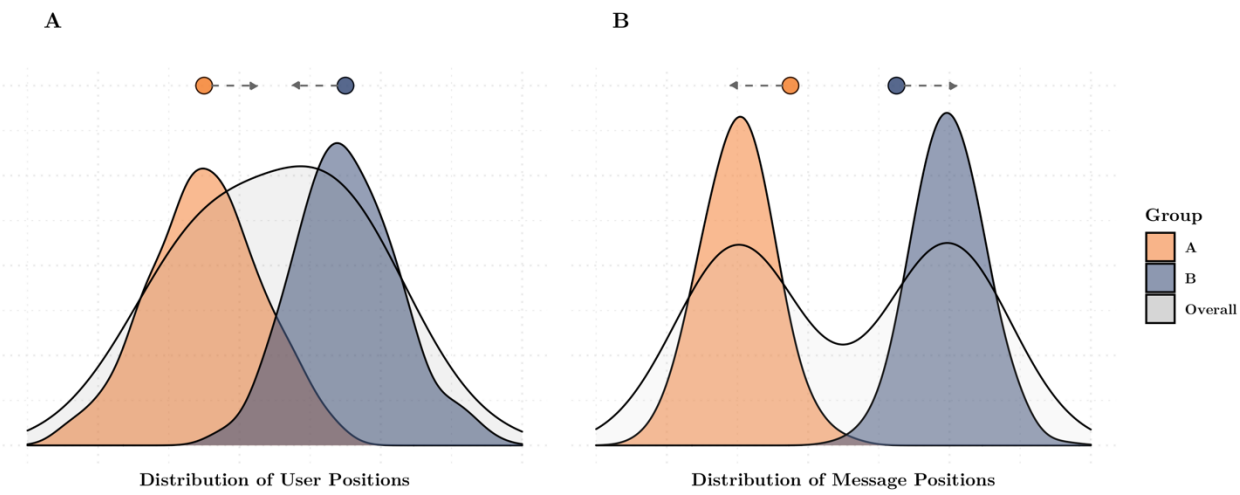


Figure 2 – How a polarised conversation can arise from a diverse set of users on social media.

Panel A shows an example distribution of ideological positions across users for a given topic for two groups; A and B. Group A and group B take different group positions on this topic, but there is a high degree of overlap across groups, leading to a normally distributed range of positions overall. Panel B shows an example where the most active users are also the most extreme users and there is a low degree of overlap between groups, with the conversation therefore becoming more polarised. The dots above each peak represent a new group member joining and the direction they will likely be directed/attracted to, based on the overall distribution.

outgroups, view the outgroup as overly homogenous, and are overly confident in these judgements (Settle, 2018). This aligns with predictions from social identity theory; once a group is recognised, intergroup interactions will lead to polarised evaluations which maximally differentiate the ingroup from the outgroup. In this way the use of social media can serve as a group-identity manipulation – leading to group level evaluations, group level emotions and group level behaviours. This increases the perceived differences between individuals' own position and where they perceive the outgroup to be and contributes to negative evaluations of the outgroup – both in terms of ideological estimates of intergroup distances and affective intergroup stance (Settle, 2018).

The structure of online environments

In addition to these biased estimates, other more structural factors increase perceptions of intergroup distance on social media which can drive polarisation (Brady, Crockett, & Van Bavel, 2020). Users with more extreme ideological positions are more active than those with moderate positions (Barberá & Rivero, 2015; Preoţiuc-Pietro, Liu, Hopkins, & Ungar, 2017), users inflate the extremity of their opinions to match perceived group norms (Rösner & Krämer, 2016), and the most extreme content gets the most attention and spreads further (Brady, Wills, Jost, Tucker, & Van Bavel, 2017). Conversely, less active social media users ("lurkers"), who follow discussions passively without participating, score higher in conflict avoidance and ambivalence than active users. These latter traits

more accurately reflect the average social media user, but due to their lack of activity these users are not reflected in the online discussions (Sun, Rau, & Ma, 2014). Consequently, a polarised online environment can arise even from a diverse group of overall participants and can create the perception of division even when there is none. This process is demonstrated in Figure 2. Panel A represents a distribution of users with differing opinions but an overall normally distributed range of positions on a given topic. If social media were a system where all users independently post one message which accurately represents their own position on this topic, then the distribution of messages will also be normally distributed. In this scenario if new prototypical members of group A and B were to join the conversation, then they would perceive a situation where the average position is less extreme than their own, and they would be drawn towards the centre. However, this scenario is unrealistic and does not represent how social media functions. For the reasons above conversation distributions do not match user distributions and are instead more polarised (bimodal). This situation is shown in panel B. If these same prototypical group members again join this conversation, they will perceive two groups with positions more extreme than themselves and be drawn to one of these extremes, likely that of their own group, or instead drop out from the conversation, which itself will also increase polarisation (Sasahara et al., 2020).

This situation has two key implications. Firstly, the perception of the online world is more polarised and divided than it is in reality, therefore when users are exposed to counter attitudinal information and outgroup members' opinions it is most likely to be those at the extremes of the outgroup rather than moderates, and these extremes will be perceived as a greater threat to the ingroup. Secondly, due to this greater social distance, users are more likely to experience negative intergroup contact compared to positive intergroup contact (MacInnis & Page-Gould, 2015). Negative intergroup contact is likely to have more negative than positive effects on prejudice, discrimination, and intergroup conflict (Paolini, Harwood, & Rubin, 2010), and while negative intergroup contact is rarer than positive intergroup contact offline (Graf, Paolini, & Rubin, 2014), this may not be the case online (Shugars & Beauchamp, 2019).

Overall, while the echo chamber theory has proposed a compelling way to explain why the Internet and social media may lead to greater social polarisation, and some of this theory's predictions about the dangers of highly concentrated intragroup interactions were correct, it has not been fully supported by empirical evidence. Social identity theory suggests that the intergroup context, and intergroup communication, are also key. While together the evidence discussed here presents a strong

argument that social media use can increase social polarisation, much remains to be understood about these processes. In particular, outstanding questions remain about the role that social media plays in intergroup conflict, and more specifically what happens when groups already in conflict interact online, and how interactions between intragroup communication and the perception of the outgroup affect this intergroup conflict. We address these questions in Chapter 1 by analysing the prevalence and nature of online conversations that occur between members of protest groups from opposite sides of the political spectrum, and investigate how this online communication relates to intergroup behaviour when these groups subsequently meet in the real world.

The online nature of intergroup hate

In the models of group radicalisation and extremism discussed so far, there is a notable consistency at the most extreme level - social processes have led to extreme denigration of the outgroup. This is either presented as 'embracing the eradication of the outgroup as necessary to defend ingroup virtue' (Reicher et al., 2008) or 'spreading discrimination, and dehumanisation against the outgroup to legitimize intergroup violence' (Smith et al., 2019). Both are expressions of intergroup hate.

Intergroup hate is closely related to the notions of prejudice and discrimination (Pearson, Dovidio, & Pratto, 2007), but also invokes important attitudinal and emotional concepts. However, despite the significance of hate at the intergroup level, it has typically not received as much attention from the field of emotional psychology as related emotions such as anger or contempt (Fischer, Halperin, Canetti, & Jasini, 2018). Allport (1954) considered hate as an emotional attitude or sentiment which overlapped with prejudice, but one which rose in prominence as individuals and groups move up through his proposed scale of prejudice and discrimination (Mullen & Leader, 2005) from 'lower level' antilocutions and avoidance of the outgroup, to more overt discrimination, attacks, and seeking to eradicate the outgroup. Intergroup hate does also appear to contain strong emotional aspects in addition to attitudinal position (Baldwin, 1960), including the moral emotional triad of contempt, anger, and disgust (Rozin, Lowery, Imada, & Haidt, 1999), although when combined and directed towards an outgroup, hate may be considered as a single emotion in itself (Sternberg, 2003). More recently, intergroup hate has been suggested to differ from emotions such as anger or disgust which occur more commonly in direct response to events (and dissipate once this event has passed, e.g. Seip, 2016), by being more stable and sustained over time (Halperin, 2008), being targeted more towards groups rather than individuals (Ben-ze'ev, 1992), and including the attribution of malicious intent to

the target group (Sternberg, 2003). This inclusion of malice attributed to the outgroup is suggested to lead to the intentions to destroy the target of this hate (either physically or symbolically), in order to ensure ingroup survival (Fischer et al., 2018). While an in-depth exploration of the exact role of these overlapping intergroup emotions in driving intergroup hate is beyond the scope of this thesis –here what we focus on is how hate impacts upon intergroup relations and conflict—it is important to keep in mind the complex nature of hate in relation to intergroup attitudes, emotions, and moral positions. The exact definition of hate speech used in this thesis, and how it is presented online, are discussed later in the literature review and in Chapter 2.

Online, this hate can be considered a form of 'extreme digital speech' (Ganesh & Bright, 2020), which amplifies the group identity of the perpetrator by attempting to create an antagonistic relationship between ingroup and outgroup members, while aspects of 'othering' serve to reinforce existing group boundaries or create new ones (Pohjonen, 2018). The level of impact that online hate speech has on group relations is currently unclear, but it is vital to understand as it has the potential to substantially drive groups apart.

The prevalence and impact of online hate speech

Formally, online hate speech (hereafter hate speech) is presented as messages which contain deliberate attacks against (or about) a specific group of people, motivated by (or focused on) aspects of that group's identity (de Gibert, Perez, García-Pablos, & Cuadros, 2018). These messages act to signal group membership to other ingroup members through a shared identity and adversary (the outgroup), whilst also cementing one's own group position. These messages also diminish members of the targeted outgroup, express the threat posed by this group, boost intergroup distinctiveness, and call for direct action (Gagliardone, Gal, Alves, & Martinez, 2015). Intergroup emotion theory demonstrates that when powerful emotions such as hate or anger are present at the group level, they can inspire attacks from ingroup members against outgroup members (Smith, Seger, & Mackie, 2007). Hate speech therefore does not only represent a reduction in intergroup relations, but also causes it to reduce further, instigating a spiral where intergroup conflict continually increases.

Understanding the nature of online hate speech, its prevalence, and its effect on intergroup relations is therefore vital to mitigate this spiral of intergroup conflict. Evidence suggests that hate speech is becoming increasingly common online. Up to 1% of content on mainstream social media platforms now contains some form of hate speech, while as many as 40% of online social media users report

having seen it (Vidgen, Margetts, & Harris, 2020). Hateful content therefore receives disproportionate attention compared to other forms of content (Kaakinen et al., 2018), and this reach is increasing, with online hate crimes in the United Kingdom rising 40% since 2017 (Williams & Mishcon de Reya, 2019). Minority groups are most frequently the targets of online hate speech, with 25% of African-Americans in the USA receiving harassment online and 25% of young women reporting being sexually harassed online (Duggan, 2017). Similar statistics are given for ethnic minorities, religious minorities, and LGBTQ minorities (Ischinger, 2020).

This prevalence of online hate has led to 27% of adults in the USA refraining from posting messages online, and 13% leaving social media altogether (Duggan, 2017). In addition, it has a number of wider negative impacts, including on the psychological well-being of individuals who are exposed to it, as well as negative impacts on intergroup relations (Tynes, Giang, Williams, & Thompson, 2008). Exposure to online hate causes communal fear (Hinduja & Patchin, 2007), particularly in historically marginalized or disadvantaged populations, and may have many of the same consequences as being targeted by offline hate crimes, such as psychological trauma (Gerstenfeld, 2003). Frequent and repetitive exposure to hate speech leads to desensitization to violence, and subsequently to lower evaluations of the victims of this hate, increasing outgroup prejudice against these targeted groups (Soral, Bilewicz, & Winiewski, 2018). A growing body of empirical evidence also suggests that online hate speech can incite people to offline violence and may be playing a role in fuelling attacks on immigrants, refugees and minority groups in Western countries (Müller & Schwarz, 2020; M. L. Williams, Burnap, Javed, Liu, & Ozalp, 2019).

Given the considerable negative consequences of online hate speech, a lot of recent research has focused on it. Exactly defining hate speech has proven difficult however, and as yet there is no single agreed upon definition of hate speech. Indeed, the concept of hate speech may have different meanings for different fields, and across policy, practice, and research. One of the few areas of consensus separating hate speech from other forms of harmful speech, is that hate speech targets groups or individuals as they relate to a group, compared to person-directed abuse which is focused solely on personal characteristics. In this way, broad categorisations of hate speech typically present it to be "*bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics"* (Sellars, 2016).

Strict legal definitions of hate speech are especially varied, and range from Canada's broad classifications of hate speech as language that which *"wilfully promotes hatred against any identifiable group,"* to the European Union's more focused description of: *"Public incitement to violence or hatred directed against a group of persons or a member of such group defined on the basis of race, descent, religion or belief, or national or ethnic origin"* (Nathaniel & Tucker, 2020). Laws preventing hate speech also greatly differ between countries. In England and Wales there is no single hate speech law, but a number of overlapping pieces of legislation which aim to forbid hatred or discrimination against individuals and groups. Broadly, hate speech is legally defined as any communication which is threatening or abusive, and is intended to harass, alarm, or distress an individual who possesses one (or more) protected characteristics: colour, race, disability, nationality, ethnic or national origin, religion, gender identity, or sexual orientation (Criminal Justice and Immigration Act, 2008; Public Order Act, 1986). Language that 'encourages terrorism' or 'glorifies terrorism acts' has also more recently been outlawed (Terrorism Act, 2006). In addition, section 127 of the Communications Act, (2003), makes it illegal to send electronic communications that are considered 'grossly' offensive. This is balanced with article 10 of the Human Rights Act, (1998), which guarantees all citizens in the UK the "right to freedom of expression", including the right to say things that might *"offend, shock or disturb the state or any sector of the population"*, although this is qualified in that this right may be restricted *"for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others"*. Posting messages to social media constitutes 'electronic communications' in the Communication Act, and these messages may be considered to commit offence at the point when the message is sent, regardless of whether it is then actually received by any intended recipient or not (Starmer, 2012). However, the threshold for what is considered 'grossly' offensive in comparison to simply 'offensive' is contested and open to substantial interpretation (De Guzman, 2012).

These conceptions differ in the United States, another country which we include in some of the studies in this thesis. In the US, a similar legal definition of "hate" is given, as bias against people or groups with specific characteristics, however the list of protected groups varies between states, as does the exact definition on which hateful acts against these groups would constitute a hate crime (The Matthew Shepard and James Byrd, Jr., Hate Crimes Prevention Act, 2009; The United States Department of Justice, 2021). With regard to hate speech in particular, there is no overall hate speech law, and indeed the US Supreme Court has ruled multiple times that criminalising hate speech would violate freedom of speech, a right guaranteed in the First Amendment of the Constitution (e.g. Matal

v. Tam, US, 2016; R.A.V. v. St. Paul, US, 1992). In specific instances hate speech can be prosecuted under the 'fighting words' exceptions to the First Amendment, defined as words which *"by their very utterance, inflict injury or tend to incite an immediate breach of the peace*", or if they are true threats of illegal conduct, such as an immediate threat to kill (Legal Information Institute, 2021). However, these cases can be prosecuted regardless of the specific motivation of violence or characteristics of the victim, and so these legislations should not be considered as focusing specifically on criminalising or preventing hate speech.

Like countries, social media platforms have also taken a diverse approach to defining hate speech for the purpose of moderating online content, and often list protected characteristics (e.g. race or religion) against which attacks are classified as hate speech (Facebook, 2020; Twitter, 2020). Listing protected characteristics has the benefit of narrowing the scope of hate speech, however it has the substantial limitation of missing hate against groups not on these lists and fails to account for attacks against multiple characteristics. In addition, strict guidelines within content moderation policies can lead to sometimes illogical conclusions. These include Facebook's recent policy decision that death threats to public figures should not constitute hate speech, regardless of whether or not these public figures are a member of a protected group or what the consequences of these death threats may be (Hern, 2021), while criticism of political decisions, such as calls to boycott certain countries, could be removed from Facebook if they contain violent religiously linked imagery (Oversight Board, 2021).

For this reason, the broader group-level definition given above (Sellars, 2016) is more adaptable. In order to obtain a more complete view of how hate develops and spreads online, in this thesis we consider hate speech at this wider level, and include targets which fall both within groups with protected characteristics, but also groups which have other characteristics which are not always considered as protected, such as immigration status, gender, political association or profession. Hate targeting the unemployed for example would not necessarily be considered legally defined hate speech for many countries, or indeed for social media platforms, however violence against the homeless has risen over the last 20 years and understanding any role of social media in this trend would require a broader definition of online hate speech. This particular example is not a topic addressed in this thesis, but it demonstrates the utility of a wide definition. Similarly, recent violence in the United States targeting politicians, for which strong evidence suggests an online facilitatory effect (Evans, 2021), may be missed using hate speech definitions which do not include political affiliation or professions, and which also exclude hate speech which has public figures as it's targets.

Because of these varied definitions, differentiating hate speech from other types of harmful 'extreme digital speech' is difficult (Ganesh & Bright, 2020), and poses an ongoing challenge. This is particularly problematic as much of the language in both cases uses similar words but with different meanings (Davidson, Warmsley, Macy, & Weber, 2017; Rossini, 2019). Exacerbating this issue, the nature of online conversations is rapidly changing, with slurs often used in benign conversations and more subtle language or code-words used to attack outgroups (Duarte, Llanso, & Loup, 2018; Gagliardone et al., 2015). This is particularly common amongst online far-right communities where symbolic communication has been commonly used to denote outgroup membership and harass these groups (Smith & Fleishman, 2016). Developing nuanced approaches to identify online hate speech, which take into account the context and rapidly changing nature of the online world, is therefore important. We address this challenge in Chapter 2 by presenting an approach for combining multiple hate speech detection models trained on data from different social media platforms. We demonstrate how this technique can both improve performance in the short-term as well as allow for a more flexible approach to hate speech detection over longer periods of time.

Studying the development of hate online

These challenges, combined with an increased requirement to analyse ever greater amounts of online content, has encouraged a burst of research into automated machine-learning driven approaches to detect hate speech it in online spaces (e.g. Burnap & Williams, 2016; Fortuna & Nunes, 2018; Waseem, Davidson, Warmsley, & Weber, 2017). While early approaches achieved relatively poor performance due to the complex nature and context of online language, more recent performance of hate speech detection models has improved due to advances in natural language processing, context aware general-purpose language models, and machine learning (Vidgen et al., 2019).

Understanding how online hate develops and the impacts it has on users who are exposed, both within groups and across groups, is a vital and ongoing challenge. Despite the increased attention to online hate from the academic research community, legal community, and social media platforms, relatively little research has been done to date to investigate the prevalence, causes, or consequences of different forms of harmful language across diverse social media platforms (Siegel, 2020). Evidence suggests that online hate communities mirror offline communities, and producers of hate speech on social media tend to be young, male, and very active online (Costello & Hawdon, 2018). However, whether participating in these online discussions negatively impacts users' relations with, and perceptions of,

other groups is currently unclear. Users who spend more time in online communities where hate speech is common tend to express more hate speech themselves, however this is not the case if users simply spend time online overall, which demonstrates that connection with hateful communities is key to hate speech expression (Costello & Hawdon, 2018). This is reflected in findings that networks of hateful users are highly connected (Mathew, Dutt, Goyal, & Mukherjee, 2019; Ribeiro, Calais, Santos, Almeida, & Meira, 2018). Producers of hate speech on Twitter tend to start out expressing more indirect hateful language then later gradually express more explicit hate (Beauchamp, Panaitiu, & Piston, 2018), suggesting that a degree of socialisation occurs within these communities. However, alternative evidence suggests that Twitter accounts that express hate speech tend to be newer (ElSherief, Nilizadeh, Nguyen, Vigna, & Belding, 2018; Ribeiro et al., 2018), supported by findings that users seek out online hate communities to express their pre-existing opinions, and therefore arrive into an online space with established prejudices (Pauwels & Schils, 2016).

More research is therefore needed to provide a robust answer to the question of whether participating in online hateful ingroup discussions drives trends in outgroup prejudice, discrimination, and hate from the users themselves. We explore these questions in Chapter 3, and investigate how hate speech expressed by individual users on fringe social media platforms changes over time, and in particular whether users arrive on these social media platforms in order to express hate, or if instead this behaviour develops through time spent interacting with other users on the platforms through a process of social contagion.

Online hate groups

Since its inception, the Internet has been adopted by extremist groups to organise, mobilise, recruit and communicate, both internally and externally (Conway, 2006). Social media platforms in particular have provided extremist groups with a centralised space to facilitate interactive communications and to connect with like-minded individuals on a global scale (Scrivens & Conway, 2020). The majority of early research into this phenomena and its effects on extremism has focused on violent Jihadi groups and Islamic extremism (e.g. Bermingham, Conway, Mcinerney, Hare, & Smeaton, 2009; Brachman & Levine, 2011; Meleagrou-Hitchens, Alexander, & Kaderbhai, 2017), mirroring the growing online presence of Jihadi groups following the 9/11 US terror attacks and the emergence of the Islamic State of Iraq and Syria (ISIS) in 2014. This latter group has been dubbed the world's first 'online terror group' (Braddock, 2020) due to their extensive use of social media platforms as a tool for spreading

propaganda, as well as the planning and preparation of terror attacks (Berger, 2014; Berger & Morgan, 2015).

Defining far-right extremism

However, Islamic extremist groups are not the only extremist groups using the Internet and social media to promote their cause. In the West, the far-right has an extensive history of Internet-related activity, including a particularly prominent spread of online hate. Broadly, far-right extremist groups can be considered as those on the political right who are 'anti-system' and hostile to liberal democracy (Mudde, 2019). For far-right groups, the ingroup and outgroup are typically framed in ethnic or racial terms (Bobbio, 1996), with the ingroup(s) presented in terms of white heterosexual male social identity (Conway, Scrivens, & Macnair, 2019), and the outgroups including immigrants, ethnic and religious minorities, left-wing groups, LGBTQ groups and feminists, amongst others (Ravndal, 2016; Jupskås & Segers, 2020). These outgroups form a set of "enemies", which are seen as a threat against the survival of the ingroup. As a result, far-right extremism is often associated with white nationalism/supremacy, anti-Semitism, racism, xenophobia and misogyny. The far-right is not a single cohesive ideology however, and can be divided into the "extreme right" which rejects the essence of democracy (the most extreme of which is fascism), and the "radical right "which accepts the essence of democracy, but opposes fundamental aspects of liberal democracy such as the rights of minority

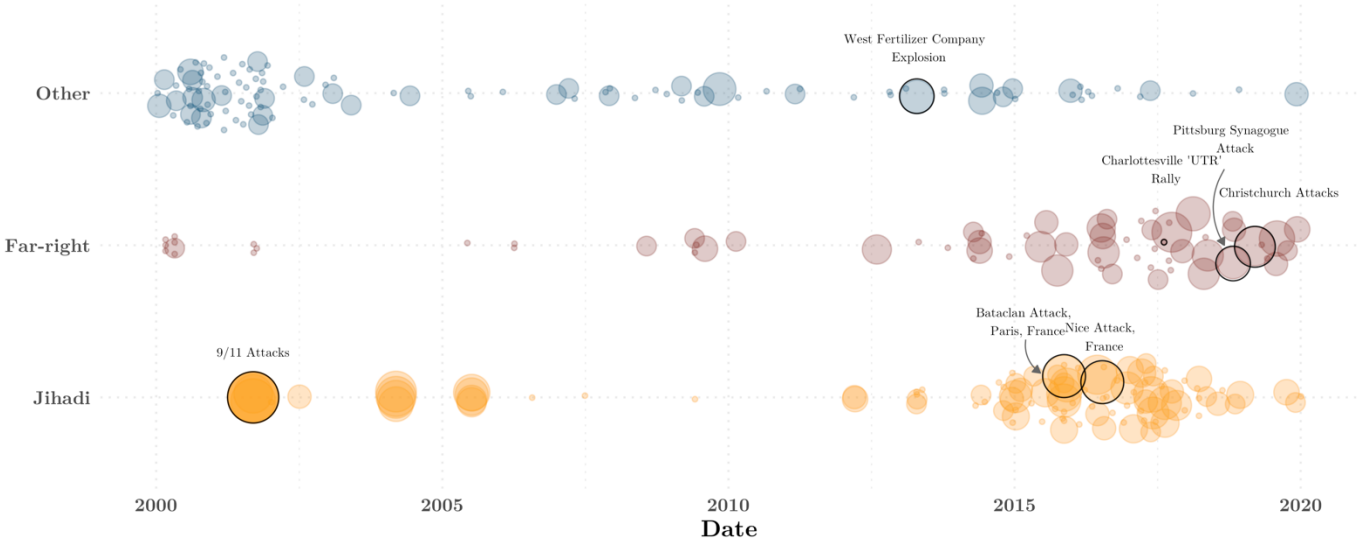


Figure 3 –The temporal distribution of terror attacks conducted in the West between 2000 and 2019. Attacks are coloured by motivation (red = far-right, yellow = Islamic Extremist/Jihadi, blue = Other) and sized by number of victims. Data sourced from the Global Terrorism Database (GTD): National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland (2019).

Figure 4 –The number of annual far-right terror attacks since 2011 in the West (Blue = outlier), with the increasing trend represented by the dotted line. Data sourced from the Global Terrorism Database (GTD): National Consortium for the Study of Terrorism and Responses to Terrorism (START), University of Maryland (2019).

groups (Mudde, 2019). A third, and more recent, counter-jihad movement is also sometimes considered a distinct subset (Koch, 2017). The 'alt-right', a loosely connected far-right and white nationalist movement which grew on the Internet since the early 2010's with a distinct adoption of 'Internet culture', is also growing in prominence (Phillips & Milner, 2020; Wendling, 2018). Discussing the differences between these types of far-right extremism is beyond the scope of this thesis, and generally hereafter we use the term "far-right extremism" to cover all these subsets.

Membership of far-right extremist groups has reached an all-time high in the last few years (Southern Poverty Law Center, 2020). These groups have been associated with an increasing number of attacks since 2010, including responsibility for 90% of extremist-related killings in the United States in 2019 (ADL Center on Extremism, 2020). This threat has become even more worrying in 2020 and has fractured, leading to a wide range of overlapping groups within the same broad ideological community but driven by a diverse mix of conspiracy theories, motivations, and online connectivity via social media (Hoffman & Clarke, 2020). These highly online phenomena have also spilled into offline violence, with direct connections between fringe online activity linked to fatal terror attacks (Evans, 2018a, 2019).

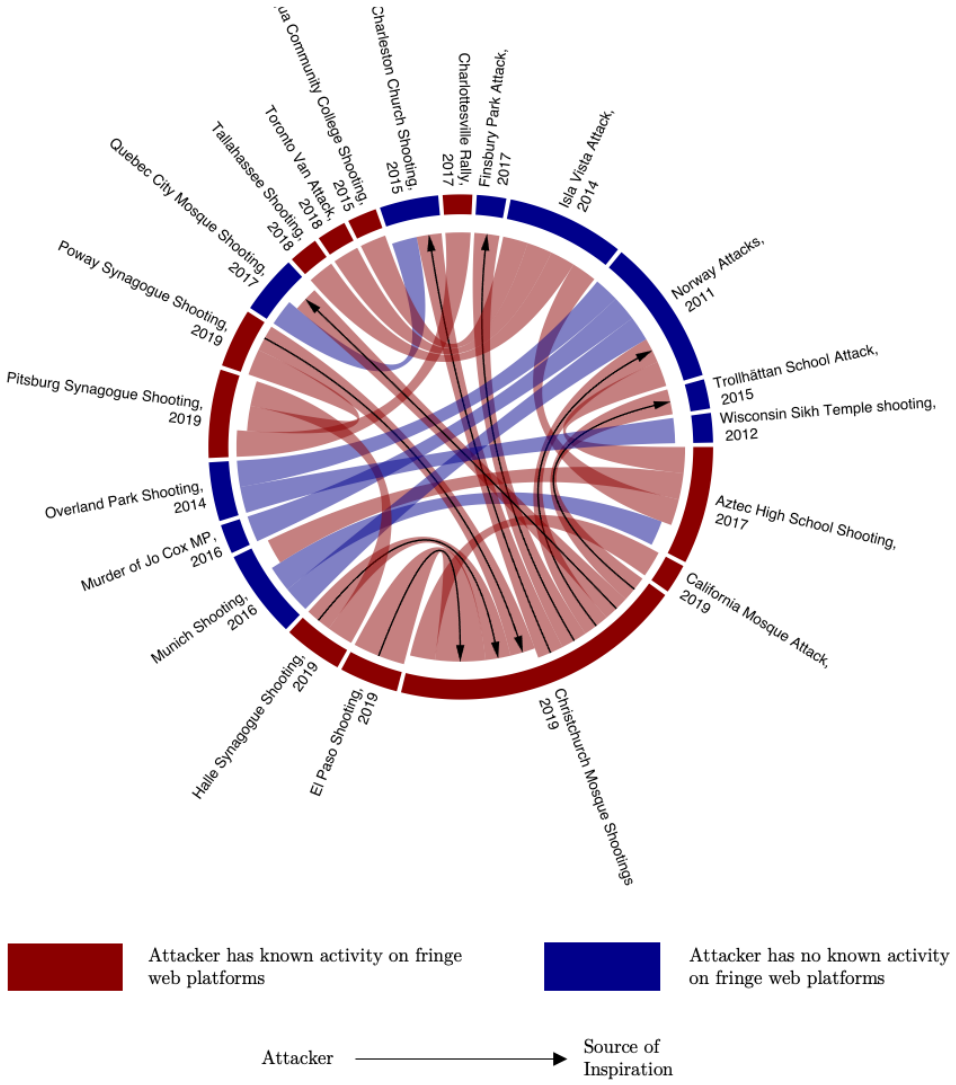


Figure 5 – The connections between notable far-right attacks conducted between 2004 and 2019 coloured by whether the attacker(s) had known activity on fringe social media platforms (red = active on fringe platforms, blue = not active on fringe platforms). Ribbon bands link the attacker (long end) to their source of inspiration (short end), as illustrated by the black arrows. Data sourced from the Global Terrorism Database, with prior analysis into motivation and connection published by the New York Times (Dodd & Grierson, 2019).

The use of the Internet by far-right extremist groups can be traced back to the early 1980s and has evolved as the Internet has developed (Conway, Scrivens, et al., 2019). This use has progressed from early pre-web bulletin board systems, to static websites, and then online forums. Most notably in the latter category is the website 'Stormfront', founded in 1996, and considered the web's first major racial hate site (Brian, 2003) because of its promotion of Holocaust denial and more recently the propagation of Islamophobia. Today, Stormfront still has over 300,000 registered users and a post history which contains over 13 million entries (Potok, 2015). More recently, with the advent of Web 2.0 and interactive web platforms which are characterized by an emphasis on user-generated content and social networking, far-right extremist groups have adopted social media as the primary mode of

online interaction (Conway, Scrivens, et al., 2019). They have had a substantial presence on mainstream social media platforms, including Facebook (Burke, 2017) and Twitter (Berger, 2018b; Davey & Ebner, 2017), and have been highly successful in building a following on these mainstream social media platforms. For example, UK-based group 'Britain First' had over 2 million Facebook followers prior to their eventual removal from the platform in 2018 (Wendling, 2017).

Concurrent with this increase in online popularity and reach for the extreme far-right, there has been an increase in offline far-right terror attacks and hate crimes in the West. Far-right extremist terror attacks have increased year-on-year in Europe, North America and Australia since 2011 (with the exception of 2016) (Figure 4). In fact, since 9/11 far-right extremism has killed more people in the United States than Islamic extremism, with less deadly but more frequent attacks (Munich Security Report) (Figure 3; The National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2020). In the UK, far-right terror plots accounted for around a third of all terror cases between 2016 and 2018, and a quarter of all terrorism arrests during this period were linked to far-right violence (Dodd & Grierson, 2019). The attackers themselves have also become increasingly interconnected, in part because of social media (Ahmed & Pisoiu, 2020). Evidence suggests that at least a third of far-right extremist attackers since 2011 were inspired by, or connected to, others who perpetrated similar attacks through direct communication, the sharing of manifestos, or simply providing inspiration (Figure 5) (Cai & Landon, 2019). This increase prevalence is also reflected in policing and counter-terror activities, for example the far-right caseload of counter-terrorism police in the UK nearly doubled between 2016 and 2018 (Dodd & Grierson, 2019).

Many far-right offline attacks have had clear links with the online far-right environment. Since the early 2000s, there have been numerous cases of users of white supremacist Internet sites committing offline hate crimes, and in many cases these sites contained explicit promotion of this violence (Cohen-Almagor, 2018). Users of the white supremacist forum Stormfront have been associated with the murders of nearly 100 people since the late 2000s, including immigrants and individuals from minority groups (Beirich, 2014). A series of extreme far-right attacks committed in 2018 and 2019 have also exemplified this trend. The attacker at the 2018 Tree of life Synagogue shooting in Pittsburgh, USA, was highly active on far-right fringe social media platforms prior to the attack (Evans, 2018b), while in 2019 the attacker at a synagogue shooting in Halle, Germany, posted a manifesto prior to the attack to online message boards steeped in Internet slang, 'meme culture', and distinct imagery often found on far-right Internet forums (Der Spiegel, 2019). Similar events occurred prior to the 2019 El

Paso shooting in the USA, while attacks at two mosques in Christchurch, New Zealand in 2019 were livestreamed over mainstream social media platforms (Evans, 2019).

Part of the apparent 'success' these far-right extremist groups have experienced on social media can be explained by social-identity processes. Those who identify more strongly with their ingroup in a given context are more likely to behave in ways driven by social-identity categorisations (Tajfel & Turner, 1979). Group identification is heightened during periods of intergroup conflict and intergroup threat, and this greater identification leads to greater conformity to specific group norms (Reicher, 1982; Reicher, 1984). Therefore, contexts in which group identification is high (such as when the group is under perceived threat) and in which the sanctioning of outgroups is highly normative, are particularly effective at driving group identification. All of these factors (group identification, perceived threat, and normative derogation) are present and utilised by far-right extremism groups online, particularly when the presence of hate speech is high.

Changes in the use of the Internet by far-right groups

In recent years however, far-right extremist groups have had more difficulty in preserving their presence on online platforms, due to increasing pressures from governments and civil society on social media companies to moderate their online spaces and remove or prevent the dissemination of online hate speech. The 'Unite the Right' rally in August 2017 in Charlottesville, USA, acted as a catalyst for this trend. This event occurred amidst controversy generated by the removal of Confederate monuments and with the stated goal of unifying the American far-right and white nationalist movements. The rally was met with counter-protestors from anti-racist and anti-fascist groups, and descended into violence, with clashes between the opposing groups, and a self-identified white supremacist deliberately driving his car into a crowd of counter-protesters, killing one person. The event had been organised, planned, and promoted via a Facebook event page, which had been removed in the days before the event due to the quantity of hate speech and calls for violence (Heath, 2017). This event is seen as a turning point for the online far-right (Deverell, Berntsson, Leman, & Valencia-García, 2020; Ebner, 2020; Wendling, 2018) and led to both an increased interest in, and criticism of, these groups; extreme far right Google searches increases by 400% in the weeks following the event (Moonshot CVE, 2018; Tien, Eisenberg, Cherng, & Porter, 2020).

Since the Charlottesville event, both Facebook and Twitter have expanded their anti-harassment and hate speech moderation policies. On Twitter, this includes the option to report an account for hateful

conduct as well as content that *"dehumanises others based on their membership of an identifiable group",* and also includes proscription against the glorification of violence and violent extremist groups (Twitter, 2020). Facebook has taken similarly steps against far-right extremist groups, including removing the pages of several high-profile groups and their leaders (Facebook Newsroom, 2018). Together, these changes have 'purged' far-right extremist accounts and material from mainstream platforms, and while it is still easily available it is not prominent to the same levels. However, these groups have not been removed from the Internet as a whole. Instead, much content has been dispersed onto smaller and more 'fringe' platforms with more lax moderation policies, including peripheral areas of Reddit, newer specialised platforms such as Gab or Voat, as well as message boards such 4Chan and 8Chan. Semi-private messaging apps such as Telegram or Discord are also increasingly popular. Studies suggest that these fringe platforms contain a greater proportion of explicit hateful content than mainstream platforms (Papasavva, Zannettou, De Cristofaro, Stringhini, & Blackburn, 2020; Zannettou et al., 2018) and are also responsible for seeding much of the hateful content which later appears on mainstream platforms (Zannettou et al., 2017). The majority of research still focuses on mainstream platforms however, and the impact that this shift to alternative fringe social media platforms on dynamics of hate speech and intergroup conflict is unclear. In this thesis we focus on both mainstream and fringe platforms.

Focus on far-right extremism

Much of the research in this thesis is done by looking specifically at far-right extremism online, and the way that far-right groups interact online, both within themselves and with other groups. Although the far-right is not the only extremist ideology present online, and despite recent moves to more fringe platforms, it is currently much more active on mainstream social media platforms than other prominent extremist ideologies, such as Islamic extremism, which was more active over the past decade. Because of this prominence, Islamic extremism has been a heavy focus for Law Enforcement and Social Media platforms' content enforcement in the past, and consequently the proportion of Islamic extremist content on mainstream social media platforms fell considerably in 2015 (Conway, Khawaja, et al., 2019). This has not yet happened to the same extent with far-right extremist content, therefore there is more readily available data.

Additionally, the history of far-right prevalence on the Internet and increasing connection between their online presence and offline violence, make far-right extremist groups ideal to investigate the dynamics of online intergroup conflict, and the role of both intragroup and intergroup online

communication in driving this extremism. However, the recent history of shifting presence from mainstream social media platforms to more fringe platforms highlights the importance for research to follow this movement and to investigate the impact that this change has on the presence of online hate across the web as a whole, and not just on mainstream platforms. The work in this thesis tracks some of this shift from mainstream to fringe social media platforms. Initially Chapter 1 focuses on the contact between opposing groups on an open mainstream social media platform (Facebook), demonstrating potential for inciting intergroup violence, and spanning the period over which the Charlottesville rally occurred. Subsequently, Chapters 3&4 investigate the nature of internal communications within a more closed and focused fringe platform (Gab), the role of intragroup interactions within these spaces on outgroup attitudes and hostility, and how this translates into offline actions and violence.

This thesis is not intended to be a study of the far-right however, nor of the way that the far-right uses the Internet. Rather, it is a study of intergroup conflict, and therefore our results should be broadly applicable to wider contexts. These effects may occur across any number of opposing groups, from rival political parties, members of conflicting nations, sectarian groups, or extremist groups from across the spectrum. This work focuses on far-right extremist groups for three primary reasons: firstly, the upwards trajectory of far-right violence over the period studied, secondly the availability of data in comparison to other extremist organisations, and thirdly, the fact that these groups operate in both the United Kingdom and the US allows us to perform analysis on English language material.

Hostile manipulation of the online environment

In the discussion in this literature review up to now the Internet has been considered a 'neutral space' where people and groups interact honestly and genuinely, and where intergroup conflict arises as a consequence of genuine interactions between real users. This is not always the case however. A key feature of the Internet is that its anonymity allows users to take on personas and group memberships that they would not otherwise be able to adopt offline, or to engage in deliberate manipulation of the environment. This allows those who wish to actively drive intergroup conflict to do so with much greater ease than they could offline.

Tactics of hostile online manipulation

One way in which this online manipulation may occur is through deliberate antagonism of opposition groups, which incites or aggravates intergroup conflict. Intergroup schadenfreude – pleasure at outgroup members' negative emotions – may motivate this type of online behaviour (Cikara, 2015; Leach, Spears, Branscombe, & Doosje, 2003), particularly when the opposing group is easily accessible online. Similarly, 'trolling', where users try to deliberately upset others and provoke an emotional response through posting inflammatory, provocative, or off-topic messages into online communications, can also exacerbate online intergroup conflict (Indiana University, 2008)

Another tactic used to deliberately drive intergroup conflict online is through organised harassment campaigns. These are common online, in particular with far-right extremist groups, and often take place as a result of organized efforts by ad-hoc mobs coordinating from fringe social media platforms (Mariconti et al., 2019), which aim to disrupt other platforms and undermine users who advocate for issues and policies they do not agree with. Minority groups and women are particularly targeted (Chatzakou et al., 2017). Often these styles of attacks use fake social media profiles and multiple accounts orchestrated at once to retain anonymity and amplify negative effects. Organised harassment campaigns have recently increased, and during the 2020 Black Lives Matter protests far-right groups

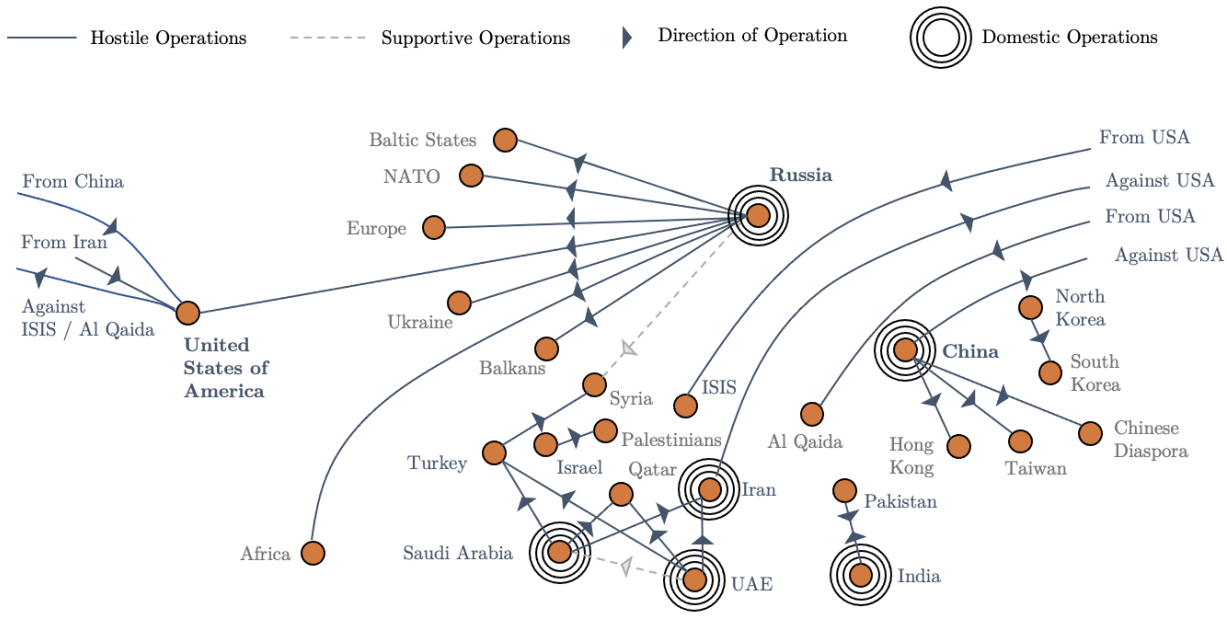


Figure 6 – The direction of targeting of major state-sponsored Information Operations as of 2019. Arrows show the direction of operations from instigator to target, with domestic operations given as concentric rings. Graphic designed by Oxford Analytica based on our analysis with data sourced from social media company public reports (Twitter Transparency Centre 2020, Facebook Newsroom 2020), republished here with permission.

have been found to have created fake far-left 'Antifa' accounts, imitating opposition groups, which gained substantial popularity online and posted messages calling for violence at the protests (Collins, Zadrozny, & Saliba, 2020). Similarly, far-right extremist individuals have been shown to have posed as Black Lives Matter supporters offline when committing acts of vandalism in an effort to 'accelerate' the intergroup conflict and spark further violence, with these activities organised on 4Chan (Davey, 2020).

This manipulation of the online environment is not limited to far-right extremist groups. In recent years, the number of nation states who have engaged in information operations on social media has dramatically increased (Bradshaw & Howard, 2019). Information operations are a form of 'weaponised trolling' with the goal to manipulate the online environment to help achieve domestic or foreign policy goals. This type of activity has more traditionally been conducted by authoritarian regimes such as Russia and China (Gallacher & Fredheim, 2019; King, Pan, & Roberts, 2017), but in 2019 Government organized social media manipulation campaigns took place in at least 70 countries, an increase of over 30% from 2018 and over 200% from 2017 (Bradshaw & Howard, 2019). These activities were often reciprocated, and the targets spanned across the globe (Figure 6).

The Russian Internet Research Agency Campaign

The most well-known example of hostile social media manipulation as part of a Government orchestrated information operation is that of the Russian Internet Research Agency (hereafter Russian IRA). Analysis of public data associated with this activity showed that since 2012 the organisation targeted both domestic Russian audiences and foreign audiences. A notable campaign that targeted the 2016 US Election, seeking to divide online groups along racial, ethnic, social, and political lines, and continued long after the election was decided (DiResta et al., 2018; Howard, Ganesh, Liotsiu, Kelly, & Camille François, 2018). More generally, both sides of numerous controversial debates were inflamed by Russian IRA activity, especially conversations surrounding provocative race issues such as the Black Lives Matter movement (Gallacher & Fredheim, 2019). Additionally, more than 100 million American citizens were exposed to political ads paid for by the Russian IRA that derogated political candidates and political groups, seeking to reinforce intergroup boundaries (Allcott & Gentzkow, 2017; Timberg, Dwoskin, And, & Demirjian, 2017).

In addition to these online manipulation attempts, the Russian IRA's activities also included attempts to ignite intergroup conflict offline, including setting up physical protests via Facebook event pages to

try and instigate violence between opposing far right and counterprotest groups (Bertrand, 2017). Russian IRA operatives were reportedly active in stoking racial divisions in the build-up to the 2017 Charlottesville 'Unite the Right' rally (Birnbaum, 2018). Similarly, Russian IRA accounts were highly active following Islamic inspired terror attacks in the UK (Innes, 2017), promoting the spread of Islamophobic hate speech and seeking to further the social division these events can cause (Pintak, Albright, Bowe, & Shaheen, 2018).

Given the widespread nature of these online manipulation attempts from both foreign and domestic actors, it is vital to understand the effect that they have on online communications and intergroup relations. Whether these efforts are having success in increasing the spread of intergroup conflict is unclear, and more research into understanding these effects is therefore needed to inform counter measures. We look to address this gap in Chapter 5 by measuring the effects of information operations from the Russian Internet Research Agency on the online behaviours on genuine social media users.

Research opportunities and challenges provided by social media data

The research presented in this thesis is motivated by three broad underlying factors. The first is the growing appreciation of the damage that the Internet and social media is doing to some elements of society. This has become increasingly apparent over the last five years that this work covers. Secondly, if we develop a better understanding on when positive and negative effects occur, we may still have the opportunity to use the Internet as a liberating force with the power to reduce intergroup conflict, sectarian violence, and break down prejudice between those with opposing political views by promoting connection between those with diverse backgrounds and experiences. Finally, there have been substantial advances in data science, machine learning and network analysis in recent years, which are now much more widely available thanks to the open-source data science community. Combining these tools with publicly available social media data provides a unique opportunity to address novel and fundamental questions on important topics such as extremism and intergroup conflict which could not have been addressed before.

This explosion of online data available for research is best demonstrated by the striking growth in the number of monthly active users across the largest social media platforms since the early 2000's (Figure 7; Ortiz-Ospina, 2019). Social media data can provide observational information on real

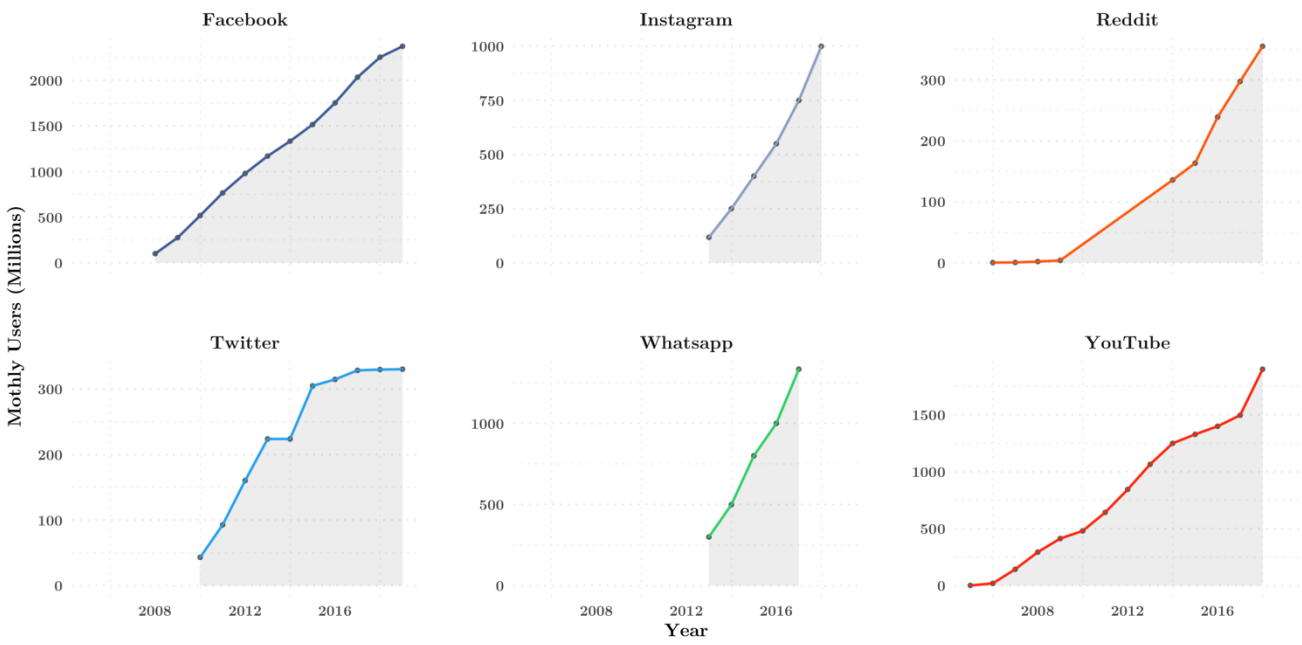


Figure 7 – The number of active monthly users across six popular social media platforms over the previous 10 years. Data from Our World in Data (Ortiz-Ospina, 2019).

behaviour and can be captured in real time or collected historically. This can provide considerable benefits both in terms of data completeness and accuracy. For example, early studies of social networks have typically relied on people's self-report of those with whom they are friends and have regular contact (e.g. Valente, 1995). This causes issues with historical accuracy due to memory limitations (primacy/recency effects), social desirability effects, as well as frequency limitations. Conversely, social media data does not require self-report and can bypass these limitations, providing a much richer and accurate picture of the nature of social interactions (Crawford & Finn, 2015; Tufekci, 2014). Furthermore, social media data can allow access to environments previously unattainable for researchers, including politically sensitive topics, unpredictable exogenous events, and diverse geographical areas, therefore allowing for a more global outlook outside of Western settings (Lazer et al 2009).

Not all social media data is equal however, and not all platforms are equally as open with sharing data with researchers. The majority of research into the social science of the Internet has been focused on a single platform, Twitter, primarily because of its very accessible API for data collection, and an abundance of activity across many topics. However, insights arising from Twitter data may not be applicable to the wider web ecosystem as a whole (Cihon & Yasseri, 2016; Vidgen & Derczynski, 2020). It is therefore important to take a multi-platform approach, which this thesis promotes by studying data from Facebook, Twitter, Reddit, Gab, Google Search Trends, and online discussion

forum Stormfront. This is challenging however, especially as some platforms are moving to restrict data access, e.g. Facebook has recently made it harder for researchers to access public conversations on its platform (Schroepfer, 2018). Data collection is therefore a bottleneck in the academic research process, and while third party organisations are making good efforts to combat this (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020; Mozilla, 2020), current social media data sharing policies are hampering this effort.

Despite the growth in available online data, substantial challenges remain when researching the interplay between the online and offline worlds. Datasets of offline social behaviour, including political violence, have become more widely available recently and are a useful resource (e.g. Google's Global Database of Events Language and Tone). However, the reliability of some of these datasets is questionable—e.g. they may contain high levels of false positives and duplicates of notable events, especially when they rely on automated data collection methods or on information repurposed from other areas. Human verification is therefore required, and additional manual collection of offline behaviours from open-source areas remains necessary. Equally, while machine learning techniques are more accessible than ever, expertise and experience are needed to apply them correctly and avoid biases and errors (Boyd & Crawford, 2012). In this work we use a combination of automated collection techniques for social media data, and combine this with manual collection of real-world event datasets, published statistics from police and government databases and crowd-sourced datasets from non-governmental organisations in order to investigate the relationship between online and offline phenomena. Nonetheless, challenges remain on how to robustly and reliably combine these disparate data sources in order to help us elucidate the relationship between online and offline activity.

The challenge of causality

A further challenge with using observational social media data for social science research is causality. Social science traditionally places a premium on the causal identification of effects, and the scientific processes of evaluating falsifiable hypotheses, usually obtained by performing carefully designed experiments. Observational data makes this process challenging – it is difficult to create manipulation and control conditions under otherwise similar circumstances. Some argue, and indeed is the traditional approach of experimental psychology, that the answer is to recreate conditions from the real world in a laboratory environment and run an artificial experiment while closely controlling all variables. This is extremely difficult however, and even when achieved, these artificial conditions often

greatly differ from 'real world conditions', both in terms of the creation of the artificial environment itself which is unlikely to replicate the highly specialised online space (Guess & Lyons, 2020), and the skewed selection of participants (Henrich, Heine, & Norenzayan, 2010). This therefore limits the applicability of such findings. Ethical limitations also exist, especially when dealing with topics such as extremism, political violence, or disinformation; it would simply not be ethical to expose participants to some of the egregious material they may see online. Researchers must therefore reach a trade-off between controlling all the desired aspects of the environment they study (internal validity) and keeping this environment true to reality (external validity).

The emphasis of much of the work presented in this thesis is placed on external validity and observing real social interactions as they occur online. This type of observational research has previously been criticised as it cannot give the same insight into causality. Indeed, correlation does not imply causation and randomised control trials are best in order to demonstrate true causation. However, this antagonism between correlational and causal evidence is something of a false dichotomy (Bradford-Hill, 1965; Cochran, 1963). Recent arguments have been made for considering evidence on a scale ranging from correlational to causal, with a number of intermediate steps which provide increasing evidence towards causality, without being definitive (Pearl, 2018).

Elements from this idea of a *'ladder of causality'* are included throughout this thesis. In an effort to provide as much robust causal evidence as possible despite using observational data, we make use of several analytical tools beyond correlation to infer the nature of relationships between variables. Chapter 1 uses temporal directionality between the online and offline worlds to strengthen evidence of the association, while also showing that a greater change in variable X leads to a greater change in variable Y. Chapter 4 goes a step beyond this and uses two-directional Granger causality to show that while a change in variable X leads to a subsequent change in variable Y, this change is not reversible. Chapter 5 goes further still, and relies on the creation of counterfactuals to measure differences between real observations and those expected in an absence of intervention. This allows us to say that besides an association between variables X and Y, we would not expect to observe this change in variable Y in the absence of a change in variable X. Together, these do not conclusively prove causal relationships beyond doubt, and nor do they claim to, but rather they add evidence to our scientific understanding.

Ethical Considerations

Conducting research using social media data raises a number of important ethical considerations, especially when investigating sensitive topics such as political extremism, intergroup conflict, and violence.

A key consideration surrounds the practical procedures of ensuring (i) anonymity of the users within the datasets and (ii) platform data agreements are adhered to. This requires only using data collected via social media platforms' official API (application program interface) for publicly available information and prohibits using web-scrapping tools to collect anything not available from the API. This also includes anonymising datasets at the point of capture by taking a cryptographic hash of usernames and identifiable information, and only publishing results in aggregated formats so that nothing is personally identifiable. In accordance with these considerations, all the work contained in this thesis has been approved by the Oxford University Ethics Committee and best practice guidelines have been adhered to throughout.

Wider considerations around the use of predictive modelling for sensitive topics are equally important. Throughout this thesis, analytical models are used to explore the relationship between social media, political extremism, and intergroup violence. These inferential models help us to build scientific evidence underpinning complex relationships. However, these models would not be suitable for predicting violence in a law enforcement or intelligence capacity. Biases and inaccuracies can occur at the level of data collection, data processing, model building, and model application (Olteanu, Castillo, Diaz, & Kıcıman, 2019), all of which could lead to disastrous consequences if these models were released 'into the wild' without sufficient validation. Currently, this validation process is not possible without access to much wider datasets, and continuous improvements and updating to the models, a process which does not fit easily into the academic cycle. In addition, the trade-offs around false-positive vs false-negative model results greatly differ in situations of academic research and law enforcement. While a false positive prediction of violence in a research setting would lead to a reduction in model accuracy and performance, in a law enforcement setting this could lead to anything from a waste of resources to a threat to life, if people are wrongly targeted.

Additionally, considerations around specific use cases and individual event level variance are different depending on whether models are used to infer generalities or make individual predictions. As such, models which are designed to explain behaviour at the collective level should not be applied to the

individual level, and certainly not without considerable care and input from subject matter experts to understand the nuances of the specific example (Patton, 2020). For example, in Chapter 1 we present evidence that offline political violence at a physical event can be 'predicted' from prior online activity surrounding this event. Prediction in this sense is statistical association which is above the level of random chance. This allows us to make inferences about underlying data distributions, but not prediction in the sense that it can be used as foresight into the future.

Finally, it is important to stress that any analysis of the relationship between social media and political protests should not distract from the genuine grievances that the movements contain (Nathaniel & Tucker, 2020). For example, in Chapter 5 we investigate the involvement from Russian-state information operations in the Black Lives Matter movement. Detecting this manipulation activity does not remove legitimate grievances held by these movements. Equally, evidence in Chapter 4 that online activity can invoke hostility does not legitimise any subsequent violence or remove agency from those involved. The role of the internet and social media in processes of extremism, intergroup conflict and violence are likely to be just one aspect of much wider trends, and results should be interpreted as such.

Summary of thesis objectives

As detailed throughout this literature review, this thesis investigates some of the unresolved research questions highlighted so far. Firstly, we investigate the nature of unstructured online intergroup communication of groups in conflict, and its effects on intergroup relations. Secondly, we explore the impact that participating in hateful ingroup conversations online has on users' outgroup attitudes and hateful online behaviours, and on their actions offline. Finally, we study the effect of hostile manipulation of online communications on the nature of these discussions.

References

Aasland, J. R. (2016). Right-wing terrorism and violence in Western Europe: Introducing the RTV dataset. *Perspectives on Terrorism*, *10*(3), 2–15. Retrieved from http://www.terrorismanalysts.com/pt/index.php/pot/article/view/508

Abrams, D., & Hogg, M. A. (1988). Comments on the motivational status of self-esteem in social identity and intergroup discrimination. *European Journal of Social Psychology*, *18*(4), 317–334. https://doi.org/10.1002/ejsp.2420180403

Abrams, D., & Hogg, M. A. (2004). Metatheory: Lessons from social identity research. *Personality and Social Psychology Review*, *8*(2), 98–106. https://doi.org/10.1207/s15327957pspr0802_2

ADL Center on Extremism. (2020). *Murder and Extremism 2019*. Retrieved from https://www.adl.org/media/14107/download

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D., & Sanford, N. (1950). *The authoritarian personality*. Harper & Brothers.

Ahmed, R., & Pisoiu, D. (2020). Uniting the far right: how the far-right extremist, new right, and populist frames overlap on Twitter–a German case study. *European Societies*, 1–23. https://doi.org/10.1080/14616696.2020.1818112

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2019). The welfare effects of social media. *SSRN Electronic Journal*, 1–116. https://doi.org/10.2139/ssrn.3308640

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. Retrieved from http://www.nber.org/papers/w23089%0Ahttp://www.nber.org/papers/w23089.ack

Allport, G. (1954). *The nature of prejudice*. (K. Clark & T. Pettigrew, Eds.). Addison-Wesley Publishing Company. https://doi.org/10.1002/9780470773963

Atran, S., Sheikh, H., & Gomez, A. (2014). For cause and comrade: Devoted actors and willingness to fight. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution*, *5*(1), 41–57. https://doi.org/10.21237/c7clio5124900

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., … Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.17605/OSF.IO/4YGUX

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Baldwin, J. M. (1960). *Dictionary of philosophy and psychology*. The Macmillan Company.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Barnidge, M. (2017). Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political Communication*, *34*(2), 302–321. https://doi.org/10.1080/10584609.2016.1235639

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *ArXiv*. Retrieved from https://arxiv.org/pdf/2001.08435.pdf

Beauchamp, N., Panaitiu, I., & Piston, S. (2018). *Trajectories of hate: Mapping individual racism and misogyny on Twitter. Unpublished working paper*.

Beirich, H. (2014). *White homicide worldwide*. Retrieved from http://www.splcenter.org/sites/default/files/intelligence_report_154_homicide_world_wide.pdf

Ben-ze'ev, A. (1992). Emotional and moral evaluations. *Meta Philosophy*, *23*(3), 214–229.

Berger, J. M. (2014). How ISIS games Twitter. *The Atlantic*. Retrieved from https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/

Berger, J. M. (2018a). *Extremism*. Cambridge, Massachusetts: The MIT Press.

Berger, J. M. (2018b). *The Alt-Right Twitter census: Defining and describing the audience for alt-right content on Twitter*. VOX-Pol Network of Excellence. Retrieved from https://www.voxpol.eu/download/vox-pol_publication/AltRightTwitterCensus.pdf

Berger, J. M., & Morgan, J. (2015). *The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter. The Brookings Project on U.S. Relations with the Islamic World*. Retrieved from

https://www.brookings.edu/research/the-isis-twitter-census-defining-and-describing-the-population-of-isis-supporters-on-twitter/

Bermingham, A., Conway, M., Mcinerney, L., Hare, N. O., & Smeaton, A. F. (2009). Combining social network analysis and sentiment analysis to explore the potential for online radicalisation. *IEEE Social Network Analysis and Mining*, 231–236. Retrieved from https://ieeexplore.ieee.org/abstract/document/5231878

Bertrand, N. (2017). Russia organized 2 sides of a Texas protest and encouraged "both sides to battle in the streets." *Business Insider*. Retrieved from https://www.businessinsider.com/russia-trolls-senate-intelligence-committee-hearing-2017-11

Birnbaum, E. (2018). GOP lawmaker: FBI told me Russia contributed to last year's violence at Charlottesville rally. *The Hill*. Retrieved from https://thehill.com/homenews/house/401403-gop-lawmaker-fbi-told-me-russia-contributed-to-what-happened-in

Bobbio, N. (1996). *Left and right: the significance of a political distinction*. University of Chicago Press. https://doi.org/10.5860/choice.34-5919

Boxell, L., Gentzkow, M., & Shapiro, J. M. (2017). Greater Internet use is not associated with faster growth in political polarization among US demographic groups. *Proceedings of the National Academy of Sciences*, *115*(3), 201706588. https://doi.org/10.1073/pnas.1706588114

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Brachman, J., & Levine, A. (2011). The world of holy warcraft: How al Qaeda is using online game theory to recruit the masses. *Foreign Policy*. Retrieved from http://www.foreignpolicy.com/articles/2011/04/13/the_world_of_holy_warcraft%5Cnhttp://www.foreignpolicy.com/articles/2011/04/13/the_world_of_holy_warcraft?page=full

Braddock, K. (2020). *Weaponized words: The strategic role of persuasion in violent radicalization and counter-radicalization*. Cambridge University Press.

Bradford-Hill, A. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, *58*, 295–300. Retrieved from https://www.edwardtufte.com/tufte/hill

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Retrieved from https://comprop.oii.ox.ac.uk/research/cybertroops2019/

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Branscombe, N., & Wann, D. L. (1994). Collective self-esteem consequences of outgroup derogation when a valued social identity is on trial. *European Journal of Social Psychology*, *24*, 641–657. Retrieved from http://doi.wiley.com/10.1002/ejsp.2420240603

Brewer, M. B. . (1991). The social self: on being the same and different at the same time. *Personality and Social Psychology Bulletin*, *17*(5), 475–482.

Bright, J. (2018). Explaining the emergence of echo chambers on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33. https://doi.org/10.2139/ssrn.2839728

Bright, J., Marchal, N., Ganesh, B., & Rudinac, S. (2020). Echo chambers exist! (But they're full of opposing views). *ArXiv*, (312827), 1–40. Retrieved from http://arxiv.org/abs/2001.11461

Brown, R. (2010). *Prejudice: Its Social Psychology* (2nd Edition). John Wiley & Sons.

Brown, Rupert, & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Bruchmann, K., Koopmann-Holm, B., & Scherer, A. (2018). Seeing beyond political affiliations: The mediating role of perceived moral foundations on the partisan similarity-liking effect. *PLoS ONE*, *13*(8), 1–20. https://doi.org/10.1371/journal.pone.0202101

Burke, S. (2017). *Anti-Semitic and Islamophobic discourse of the British far-right on Facebook. Loughborough University*. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/27177/1/Thesis-2017-Burke.pdf

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0072-6

Cai, W., & Landon, S. (2019). Attacks by white extremists are growing. So are their connections. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2019/04/03/world/white-extremist-terrorism-christchurch.html

Cambre, J., Klemmer, S. R., & Kulkarni, C. (2017). Escaping the echo chamber: Ideologically and geographically diverse discussions about politics. *Conference on Human Factors in Computing Systems - Proceedings*, 2423–2428. https://doi.org/10.1145/3027063.3053265

Castree, N., Kitchen, R., & Rogers, A. (2013). *A dictionary of human geography*. Oxford University Press.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Mean birds: Detecting aggression and bullying on Twitter. *ArXiv*. https://doi.org/10.1145/3091478.3091487

Chayko, M. (2015). The first web theorist? Georg Simmel and the legacy of 'The web of group-affiliations.' *Information, Communication & Society*, *18*(12), 1419–1422. https://doi.org/10.1080/1369118x.2015.1042394

Cihon, P., & Yasseri, T. (2016). A biased review of biases in Twitter studies on political Collective Action. *Frontiers in Physics*, *4*(August), 1–8. https://doi.org/10.3389/fphy.2016.00034

Cikara, M. (2015). Intergroup Schadenfreude: Motivating participation in collective violence. *Current Opinion in Behavioral Sciences*, *3*, 12–17. https://doi.org/10.1016/j.cobeha.2014.12.007

Cochran, W. G. (1963). *Sampling Technique* (2nd Editio). New York: John Wiley and Sons Inc.

Cohen-Almagor, R. (2018). Taking North American white supremacist groups seriously: The scope and challenge of hate speech on the Internet. *International Journal for Crime, Justice and Social Democracy*, *7*(2), 38–57. https://doi.org/10.5204/ijcjsd.v7i2.517

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Collins, B. Ben, Zadrozny, B., & Saliba, E. (2020). White nationalist group posing as antifa called for violence on Twitter. *NBC News*. Retrieved from https://www.nbcnews.com/tech/security/twitter-takes-down-washington-protest-disinformation-bot-behavior-n1221456

Communications Act (2003). United Kingdom of Great Britain and Northern Ireland. Retrieved from https://www.legislation.gov.uk/ukpga/2003/21/section/127

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *133*(26), 89–96. https://doi.org/10.1021/ja202932e

Conway, M. (2006). Terrorism and the Internet: New media - new threat? *Parliamentary Affairs*, *59*(2), 283–298. https://doi.org/10.1093/pa/gsl009

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2019). Disrupting Daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, *42*(1–2), 141–160. Retrieved from https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513984

Conway, M., Scrivens, R., & Macnair, L. (2019). Right-wing extremists' persistent online presence: History and contemporary trends. *ICCT Policy Brief*. https://doi.org/10.19165/2019.3.12

Coser, L. A. (1956). *The functions of social conflict*. Free Press. https://doi.org/10.4324/9780203714577

Costello, M., & Hawdon, J. (2018). Who are the online extremists among us? Sociodemographic characteristics, social networking, and online experiences of those who produce online hate materials. *Violence and Gender*, *5*(1), 55–60. https://doi.org/10.1089/vio.2017.0048

Cota, W., Ferreira, S. C., Pastor-Satorras, R., & Starnini, M. (2019). Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Science*, *8*(1). https://doi.org/10.1140/epjds/s13688-019-0213-9

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Crawford, K., & Finn, M. (2015). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, *80*(4), 491–502. https://doi.org/10.1007/s10708-014-9597-z

Criminal Justice and Immigration Act (2008). United Kingdom of Great Britain and Northern Ireland. Retrieved

from https://www.legislation.gov.uk/ukpga/2008/4/pdfs/ukpga_20080004_en.pdf

Davey, B. J. (2020). *Infiltration operations: How 4chan sought to compromise the Black Lives Matter protests*. Retrieved from https://www.isdglobal.org/digital_dispatches/infiltration-operations-how-4chan-sought-to-compromise-the-black-lives-matter-protests/

Davey, J., & Ebner, J. (2017). *The fringe insurgency: Connectivity, convergence and mainstreaming of the extreme right*. Retrieved from http://www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 512–515. Retrieved from http://arxiv.org/abs/1703.04009

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *15*(4), 332–351. https://doi.org/10.1177/1088868311411103

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 11–20. https://doi.org/10.18653/v1/w18-5102

De Guzman, N. (2012). Section 127 of the Communications Act 2003: Threat or menace? *LSE Media Policy Project*. Retrieved from http://ssrn.com/abstract=2200166%5Cnhttp://blogs.lse.ac.uk/mediapolicyproject/2012/10/19/section-127-of- the-communications-act-2003- threat-or-menace/October 22, 2012

Der Spiegel. (2019). Far-right terrorism in Germany: Shooting exposes lapses in security apparatus. *Der Speigel*. Retrieved from https://www.spiegel.de/international/germany/far-right-terrorism-in-germany-shooting-exposes-lapses-in-security-apparatus-a-1291075.html

DeSteno, D., Dasgupta, N., Bartlett, M. Y., & Cajdric, A. (2004). Prejudice from thin air: The effect of emotion on automatic intergroup attitudes. *Psychological Science*, *15*(5), 319–324. https://doi.org/10.1111/j.0956-7976.2004.00676.x

Deverell, F., Berntsson, J., Leman, J., & Valencia-García, L. D. (2020). How Nordic neo-Nazis use the internet. Retrieved November 13, 2020, from https://www.techagainstterrorism.fm/how-nordic-neo-nazis-use-the-internet/

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., … Johnson, B. (2018). *The tactics & tropes of the Internet Research Agency*. Retrieved from https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper-121718.pdf

Dodd, V., & Grierson, J. (2019). Fastest-growing UK terrorist threat is from far right, say police. *The Guardian*. Retrieved from https://www.theguardian.com/uk-news/2019/sep/19/fastest-growing-uk-terrorist-threat-is-from-far-right-say-police

Dovidio, J. F., & Gaertner, S. L. (1991). Changes in the expression and assessment of racial prejudice. In *Opening doors: Perspectives on race relations in contemporary America*. The University of Alabama Press. Retrieved from https://psycnet.apa.org/record/1991-98067-007

Duarte, N., Llanso, E., & Loup, A. (2018). Mixed messages? The limits of automated social media content analysis. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *81*, 106. Retrieved from http://proceedings.mlr.press/v81/duarte18a.html

Duffy, B., Hewlett, K., McCrae, J., & Hall, J. (2019). *Divided Britain. The Policy Institute, Kings College London*. https://doi.org/10.1177/0265813515604528

Duggan, M. (2017). Online harassment 2017. *Pew Research Center*, 1–85. Retrieved from https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

Duggan, M., Smith, A., & Page, D. (2016). *The political environment on social media*. Retrieved from http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Ebner, J. (2020). *Going dark: The secret social lives of extremists*. Bloomsbury Publishing.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *ICWSM*. Retrieved from http://arxiv.org/abs/1804.04649

Evans, R. (2018a). *How the MAGAbomber and the synagogue shooter were likely radicalized. Bellingcat*. Retrieved from https://www.bellingcat.com/news/americas/2018/10/31/magabomber-synagogue-shooter-likely-radicalized/comment-page-2/

Evans, R. (2018b). How the MAGAbomber and the Synagogue Shooter Were Likely Radicalized.

Evans, R. (2019). *Shitposting, inspirational terrorism, and the Christchurch mosque massacre. Bellingcat*. Retrieved from https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/

Evans, R. (2021). How the insurgent and MAGA right are being welded together on the streets of Washinton D.C. *Bellingcat*. Retrieved from https://www.bellingcat.com/news/americas/2021/01/05/how-the-insurgent-and-maga-right-are-being-welded-together-on-the-streets-of-washington-d-c/

Facebook. (2020). Facebook community standards: Hate speech. Retrieved July 9, 2020, from https://www.facebook.com/communitystandards/hate_speech

Facebook Newsroom. (2018). Taking Action Against Britain First. Retrieved April 11, 2018, from https://newsroom.fb.com/news/h/taking-action-against-britain-first/

Fischer, A., Halperin, E., Canetti, D., & Jasini, A. (2018). Why we hate. *Emotion Review*, *10*(4), 309–320. https://doi.org/10.1177/1754073917751229

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization. Retrieved from http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Gagnon, A., & Bourhis, R. Y. (1996). Discrimination in the minimal group paradigm: Social identity or self-interest? *Personality and Social Psychology Bulletin*, *22*(12), 1289–1301. https://doi.org/10.1177/01461672962212009

Gallacher, J. D., & Fredheim, R. E. (2019). Division abroad, cohesion at home: How the Russian troll factory works to divide societies overseas but spread pro-regime messages at home. In *Responding to Cognitive Security Challenges* (p. 60:79). Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Ganesh, B., & Bright, J. (2020). *Extreme digital speech: Contexts, responses and solutions*. Retrieved from https://www.voxpol.eu/new-vox-pol-report-extreme-digital-speech-contexts-responses-and-solutions/

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. *Journal of Computer-Mediated Communication*, *21*(5), 331–348. https://doi.org/10.1111/jcc4.12164

Gerstenfeld, P. B. (2003). *Hate crimes: Causes, controls, and controversies*. Sage Publications, Inc.

González, K. V., Verkuyten, M., Weesie, J., & Poppe, E. (2008). Prejudice towards Muslims in the Netherlands: Testing integrated threat theory. *British Journal of Social Psychology*, *47*(4), 667–685. https://doi.org/10.1348/014466608X284443

Graf, S., Paolini, S., & Rubin, M. (2014). Negative intergroup contact is more influential, but positive intergroup contact is more common: Assessing contact prominence and contact prevalence in five Central European countries. *European Journal of Social Psychology*, *44*(6), 536–547. https://doi.org/10.1002/ejsp.2052

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Guess, A. M., & Lyons, B. A. (2020). Misinformation, disinformation, and online propaganda. In *Social Media and Democracy: The State of the Field and Prospects for Reform*. Cambridge University Press.

Halperin, E. (2008). Group-based hatred in intractable conflict in Israel. *Journal of Conflict Resolution*, *52*(5), 713–736. https://doi.org/10.1177/0022002708314665

Hamid, N., Pretus, C., Atran, S., Crockett, M. J., Ginges, J., Sheikh, H., … Vilarroya, O. (2019). Neuroimaging "will to fight" for sacred values: An empirical case study with supporters of an Al Qaeda associate. *Royal Society Open Science*, *6*(6). https://doi.org/10.1098/rsos.181585

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–97. https://doi.org/10.1037/h0076835

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review*, *10*(3), 252–264. Retrieved from papers2://publication/uuid/014DC7A5-77D8-47BE-9341-D730C2237A67

Haslam, S. A., Reicher, S. D., & Bavel, J. J. Van. (2019). Rethinking the nature of cruelty: The role of identity leadership in the Stanford prison experiment. *American Psychologist*, *74*(7), 809–822. https://doi.org/10.1037/amp0000443.supp

Heath, A. (2017). Facebook removed the event page for white nationalist "Unite the Right" rally in Charlottesville one day before it took place. *Business Insider*. Retrieved from https://www.businessinsider.com/facebook-removed-unite-the-right-charlottesville-rally-event-page-one-day-

before-2017-8?op=1&r=US&IR=T

Heltzel, G., & Laurin, K. (2020). Polarization in America: two possible futures. *Current Opinion in Behavioral Sciences*, *34*(January), 179–184. https://doi.org/10.1016/j.cobeha.2020.03.008

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*. https://doi.org/10.1017/S0140525X0999152X

Hern, A. (2021). Facebook guidelines allow users to call for death of public figures. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2021/mar/23/facebook-guidelines-allow-for-users-to-call-for-death-of-public-figures

Hewstone, M., Stroebe, W., & Jonas, K. (2016). *Introduction to Social Psychology* (6th Editio). John Wiley & Sons.

Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, *6*(3), 89–112. https://doi.org/10.1300/J202v06n03_06

Hinsz, V. B., & Davis, J. H. (1984). Persuasive arguments theory, group polarization, and choice shifts. *Personality and Social Psychology Bulletin*, *10*(2), 260–268. https://doi.org/10.1177/0146167284102012

Hoffman, B. (2006). *Inside Terrorism*. Columbia University Press.

Hoffman, B., & Clarke, C. (2020). The next American terrorist. *The Cipher Brief*. Retrieved from https://www.thecipherbrief.com/the-next-american-terrorist

Howard, P. N., Ganesh, B., Liotsiu, D., Kelly, J., & Camille François, G. (2018). *The IRA, social media and political polarization in the United States, 2012-2018*. Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf

Human Rights Act (1998). United Kingdom of Great Britain and Northern Ireland. Retrieved from https://www.legislation.gov.uk/ukpga/1998/42/contents

Indiana University. (2008). What is a troll? Retrieved from https://kb.iu.edu/d/afhc

Innes, M. (2017). *Russian influence and interference measures following the 2017 UK terrorist attacks. Cardiff University Crime and Security Research Institute*. Retrieved from https://crestresearch.ac.uk/resources/russian-influence-uk-terrorist-attacks/

Ischinger, W. (2020). *Munich Security Report 2020*. Retrieved from https://securityconference.org/en/

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jackson, J. . (1993). Realistic group conflict theory: A review and evaluation of the theoretical and empirical literature. *The Psychological Record*, *43*(3), 395–413. Retrieved from https://psycnet.apa.org/record/1994-05428-001

Jupskås, A. R., & Segers, I. B. (2020). What is right-wing extremism? *C-Rex Compendium*. Retrieved from https://www.sv.uio.no/c-rex/english/groups/compendium/what-is-right-wing-extremism.html

Kaakinen, M., Räsänen, P., Näsi, M., Minkkinen, J., Keipi, T., & Oksanen, A. (2018). Social capital and online hate production: A four country survey. *Crime, Law and Social Change*, *69*(1), 25–39. https://doi.org/10.1007/s10611-017-9764-5

Kaplan, A. (1978). The psychodynamics of terrorism. *Terrorism*, *1*(3–4), 237–254. https://doi.org/10.1080/10576107808435411

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(03), 484–501. https://doi.org/10.1017/s0003055417000144

Knobloch-Westerwick, S. (2015). The selective exposure self- and affect-management (SESAM) model: Applications in the realms of race, politics, and health. *Communication Research*, *42*(7), 959–985. https://doi.org/10.1177/0093650214539173

Koch, A. (2017). The new crusaders: Contemporary extreme right symbolism and rhetoric. *Perspectives on Terrorism*, *11*(5), 13–24.

Kruglanski, A. W., Bélanger, J. J., Gelfand, M., Gunaratna, R., Hettiarachchi, M., Reinares, F., … Sharvit, K. (2013). Terrorism-A (Self) love story: Redirecting the significance quest can end violence. *American Psychologist*, *68*(7), 559–575. https://doi.org/10.1037/a0032615

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, *35*(SUPPL.1), 69–93. https://doi.org/10.1111/pops.12163

Kteily, N., Bruneau, E., Waytz, A., & Cotterill, S. (2015). The ascent of man: Theoretical and empirical evidence

for blatant dehumanization. *Journal of Personality and Social Psychology*, *109*(5), 901–931. https://doi.org/10.1037/pspp0000048

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the Web. *ArXiv*, 1–11. https://doi.org/10.1145/3178876.3186141

Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies*, *34*(2), 283–301. https://doi.org/10.1016/0020-7373(91)90045-9

Leach, C. W., Spears, R., Branscombe, N. R., & Doosje, B. (2003). Malicious pleasure: Schadenfreude at the suffering of another group. *Journal of Personality and Social Psychology*, *84*(5), 932–943. https://doi.org/10.1037/0022-3514.84.5.932

Lee, C., Shin, J., & Hong, A. (2018). Does social media use really make people politically polarized? Direct and indirect effects of social media use on political polarization in South Korea. *Telematics and Informatics*, *35*(1), 245–254. https://doi.org/10.1016/j.tele.2017.11.005

Legal Information Institute. (2021). Fighting Words. Retrieved from https://www.law.cornell.edu/wex/fighting_words#:~:text=Fighting words are%2C as first,immediate breach of the peace.&text=Fighting words are a category,unprotected by the First Amendment

Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly*, *80*(Specialissue1), 392–410. https://doi.org/10.1093/poq/nfw005

Levine, J. M., & Moreland, R. L. (1994). Group socialization: Theory and research. *European Review of Social Psychology*, *5*(1), 305–336. https://doi.org/10.1080/14792779543000093

Liao, V. Q., & Fu, W.-T. (2014). Can you hear me now? Mitigating the echo chamber effect by source position indicators. *CSCW: Proceedings of the 17th ACM Conference on Computer Supported Co-Operative Work & Social Computing*. Retrieved from https://dl.acm.org/doi/10.1145/2531602.2531711

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Mackie, D. M., Devos, T., & Smith, E. R. (2000). Intergroup emotions: explaining offensive action tendencies in an intergroup context. *Journal of Personality and Social Psychology*, *79*(4), 602–616. https://doi.org/10.1037/0022-3514.79.4.602

Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., … Stringhini, G. (2019). "You know what to do": Proactive detection of YouTube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3359309

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mason, L. (2018). *Uncivil Agreement: How Politics Became Our Identity*. University of Chicago Press.

Matal v. Tam, US, Pub. L. No. 582 (2016). Retrieved from https://www.supremecourt.gov/opinions/16pdf/15-1293_1o13.pdf

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 173–182. https://doi.org/10.1145/3292522.3326034

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

Meleagrou-Hitchens, A., Alexander, A., & Kaderbhai, N. (2017). The impact of digital communications technology on radicalization and recruitment. *International Affairs*, *93*(5), 1233–1249. https://doi.org/10.1093/ia/iix103

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Monroe, K. R., & Hankin, J. (2000). The psychological foundations of identity politics. *Annual Review of Political Secience*, *3*, 419:447.

Montalvo, J. G., & Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *American Economic Review*, *95*(3), 796–816. https://doi.org/10.1257/0002828054201468

Moonshot CVE. (2018). *Searching for hate in America*. Retrieved from http://moonshotcve.com/searching-for-

hate-in-america/

Moreland, R. L., & Levine, J. M. (1982). Socialization in small groups: Temporal changes in individual-group relations. *Advances in Experimental Social Psychology*, *15*(C), 137–192. https://doi.org/10.1016/S0065-2601(08)60297-X

Morgan, S. J. (2001). *The mind of a terrorist fundamentalist: The psychology of terror cults* (1st Editio). Institute Spiritus Vitus.

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135. https://doi.org/10.1037/h0027568

Mozilla. (2020). SMAT - Social Media Analysis Toolkit. Retrieved November 11, 2020, from https://www.smat-app.com/

Mudde, C. (2007). *Populist radical right parties in Europe*. Cambridge: Cambridge University Press.

Mudde, Cas. (2019). *The far right today*. Cambridge: Polity.

Mullen, B., & Leader, T. (2005). Linguistic factors: Antilocutions, ethnonyms, ethnophaulisms, and other varieties of hate speech. In *On the Nature of Prejudice: 50 years after Allport* (pp. 192–207). https://doi.org/10.1002/9780470773963.ch12

Müller, K., & Schwarz, C. (2020). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *00*(0), 1–37. https://doi.org/10.1093/jeea/jvaa045

Mummendey, A., Otten, S., Berger, U., & Kessler, T. (2000, November 2). Positive-negative asymmetry in social discrimination: Valence of evaluation and salience of categorization. *Personality and Social Psychology Bulletin*. SAGE Publications Inc. https://doi.org/10.1177/0146167200262007

Nathaniel, P., & Tucker, J. A. (2020). *Social media and democracy: The state of the field and prospects for reform*. Cambridge University Press. https://doi.org/10.1016/j.solener.2019.02.027

Nicholson, S. P. (2012). Polarizing cues. *American Journal of Political Science*, *56*(1), 52–66. https://doi.org/10.1111/j.1540-5907.2011.00541.x

Nijstad, B. A., & van Knippenberg, D. (2016). Group Dynamics. In *An Introduction to Social Psychology* (pp. 379–405). Wiley.

Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*(July). https://doi.org/10.3389/fdata.2019.00013

Ortiz-Ospina, E. (2019). The rise of social media. https://doi.org/10.4324/9781315563312-21

Oversight Board. (2021). *Oversight Board overturns Facebook decision: Case 2020- 007-FB-FBR*. Retrieved from https://oversightboard.com/news/719407085355044-oversight-board-overturns-facebook-decision-case-2020-007-fb-fbr/

Paolini, S., Harwood, J., & Rubin, M. (2010). Negative intergroup contact makes group memberships salient: Explaining why intergroup conflict endures. *Personality and Social Psychology Bulletin*, *36*(12), 1723–1738. https://doi.org/10.1177/0146167210388667

Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., & Blackburn, J. (2020). Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. *ArXiv*. Retrieved from http://arxiv.org/abs/2001.07487

Pariser, E. (2011). *The Filtter Bubble: What the internet is hiding from you*. New York, New York, USA: The Penguin Press.

Patton, D. (2020). *Turing Lecture: AI for innovative social work. The Alan Turing Institute*. Retrieved from https://www.turing.ac.uk/events/turing-lecture-ai-innovative-social-work

Pauwels, L., & Schils, N. (2016). Differential online exposure to extremist content and political Violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence*, *28*(1), 1–29. https://doi.org/10.1080/09546553.2013.876414

Pearl, J. (2018). *The book of why: The new science of cause and effect*. Allen Lane.

Pearson, A. R., Dovidio, J. F., & Pratto, F. (2007). Racial prejudice, intergroup hate, and blatant and subtle bias of whites toward blacks in legal decision making in the United States. *International Journal of Psychology and Psychological Therapy*, *7*(2), 145–158.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of*

Personality and Social Psychology, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Phillips, W., & Milner, R. M. (2020). *You are here: A field guide for navigating polarized speech, conspiracy theories, and our polluted media landscape*. The MIT Press. https://doi.org/10.1038/scientificamerican1112-18a

Pintak, L., Albright, J., Bowe, B. J., & Shaheen, P. (2018). *#Islamophobia: stoking fear and prejudice in the 2018 midterms*. Retrieved from https://doi.org/10.35650/MD.2006.a.2019

Pohjonen, M. (2018). *Horizons of hate: A comparative approach to social media hate speech*. Retrieved from https://www.voxpol.eu/download/vox-pol_publication/Horizons-of-Hate.pdf

Post, J. M. (1984). Notes on a psychodynamic theory of terrorist behavior. *Terrorism*, *7*(2), 241–256. https://doi.org/10.1080/10576108408435577

Postmes, T., Spears, R., & Lea, M. (1998). Building or breaching social boundries? SIDE effects of computer mediated communication. *Communication Research*, *25*(6), 689–715.

Potok, M. (2015). The Year in Hate and Extremism. *Southern Poverty Law Center Intelligence Report*, (160), 1–73. Retrieved from https://www.splcenter.org/sites/default/files/ir160-spring2016-splc.pdf

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. *Proceedings Ofthe 55th Annual Meeting Ofthe Association for Computational Linguistics*, 729–740. https://doi.org/10.18653/v1/p17-1068

Public Order Act (1986). United Kingdom of Great Britain and Northern Ireland. Retrieved from https://www.legislation.gov.uk/ukpga/1986/64/contents

Quattrociocchi, W., Caldarelli, G., & Scala, A. (2014). Opinion dynamics on interacting networks: Media competition and social influence. *Scientific Reports*, *4*, 1–7. https://doi.org/10.1038/srep04938

R.A.V. v. St. Paul, US, Pub. L. No. 377 (1992). Retrieved from https://supreme.justia.com/cases/federal/us/505/377/

Ramiah, A. Al, & Hewstone, M. (2013). Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: Evidence, limitations, and potential. *American Psychologist*, *68*(7), 527–542. https://doi.org/10.1037/a0032603

Rani, N. (2018). Social media in India: A human security perspective. *The Research Journal of Social Sciences*, *9*(10), 43–52.

Reed, A., Whittaker, J., Votta, F., & Looney, S. (2019). Radical filter bubbles and extremist content. *Global Research Network on Terrorism and Technology*, (8). Retrieved from https://rusi.org/sites/default/files/20190726_grntt_paper_08_0.pdf

Reicher, S. . (1982). The determination of collective nehaviour. In *Social Identity and Intergroup Relations* (pp. 41–83). Cambridge University Press. Retrieved from https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=2192593

Reicher, S. D. (1984). Social influence in the crowd: Attitudinal and behavioural effects of de-individuation in conditions of high and low group salience. *British Journal of Social Psychology*, *23*(4), 341–350. https://doi.org/10.1111/j.2044-8309.1984.tb00650.x

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, *2*(3), 1313–1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W. (2018). Characterizing and detecting hateful users on Twitter. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 676–679. Retrieved from https://arxiv.org/pdf/1803.08977.pdf

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). Auditing radicalization pathways on YouTube. Retrieved from http://arxiv.org/abs/1908.08313

Rösner, L., & Krämer, N. C. (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media and Society*, *2*(3). https://doi.org/10.1177/2056305116664220

Rossini, P. (2019). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. https://doi.org/10.1177/0093650220921314

Rowe, M., & Saif, H. (2016). Mining pro-ISIS radicalisation signals from social media users. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 329–338. Retrieved from http://oro.open.ac.uk/48477/1/13023-57822-1-PB.pdf

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: A mapping between the other-directed moral emotions, disgust, contempt and anger and Shweder's three universal moral codes. *Journal of Personality and Social Psychology*, *76*(4), 574–586. Retrieved from http://proxy.lib.sfu.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=pzhref&AN=JPSP.GF.EGD.ROZIN.CTHMBO&site=ehost-live

Rule, N. O., & Ambady, N. (2010). Democrats and Republicans can be differentiated from their faces. *PLoS ONE*, *5*(1), 1–7. https://doi.org/10.1371/journal.pone.0008733

Sageman, M. (2004). *Understanding terror networks*. University of Pennsylvania Press.

Samaratunge, S., & Hattotuwa, S. (2014). *Liking Violence: A study of hate speech on Facebook in Sri Lanka*. Retrieved from https://www.cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf

Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., & Menczer, F. (2020). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, (0123456789). https://doi.org/10.1007/s42001-020-00084-7

Schroepfer, M. (2018). *An update on our plans to restrict data access on Facebook*. Retrieved from https://about.fb.com/news/2018/04/restricting-data-access/

Scrivens, R., & Conway, M. (2020). The roles of "old" and "new" media tools and technologies in the facilitation of violent extremism and terrorism. In *The Human Factor of Cybercrime* (Vol. 3, pp. 286–309). https://doi.org/10.4324/9780429460593-13

Seip, E. C. (2016). *Desire for vengeance and revenge: An emotion-based approach to revenge*. University of Amsterdam. Retrieved from https://www.semanticscholar.org/paper/Desire-for-vengeance-and-revenge%3A-An-emotion-based-Seip/6ad0bf6aff07614040278c921bee3dbe298d08de

Sellars, A. F. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *20*. https://doi.org/10.1093/jicj/mqaa023

Settle, J. E. (2018). *Frenemies*. Cambridge University Press. https://doi.org/10.1017/9781108560573

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., Carolyn, W., & Green, C. D. (1961). *Intergroup conflict and cooperation: The robbers cave experiment*. Wesleyan University Press.

Shugars, S., & Beauchamp, N. (2019). Why keep arguing? Predicting engagement in political conversations online. *SMaPP-Global*, 1–13. https://doi.org/10.1177/ToBeAssigned

Sidanius, J., Haley, H., Molina, L., & Pratto, F. (2007). Vladimir's choice and the distribution of social resources: A group dominance perspective. *Group Processes and Intergroup Relations*, *10*(2), 257–265. https://doi.org/10.1177/1368430207074732

Siegel, A. A. (2020). Online hate speech. In *Social media and democracy: The state of the field and prospects for reform* (pp. 56–88). Cambridge University Press.

Simmel, Georg. (1972). *On individuality and social forms. University of Chicago Press*. University of Chicago Press.

Simmel, George. (1898). The persistence of social groups. *American Journal of Sociology*, *4*, 167–189.

Simmel, George. (1904). The sociology of conflict. *The American Journal of Sociology*, *9*. Retrieved from https://archive.org/details/jstor-4576614

Smith, A., & Fleishman, C. (2016). (((Echoes))), exposed: The secret symbol neo-nazis use to target Jews online. *Mic*. Retrieved from https://www.mic.com/articles/144228/echoes-exposed-the-secret-symbol-neo-nazis-use-to-target-jews-online#.NZBPQsAaz

Smith, E.R. (1993). Social identity and social emotions: Toward new conceptualizations of prejudice. In *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 297–315). Academic Press.

Smith, Eliot R, Seger, C. R., & Mackie, D. M. (2007). Can Emotions Be Truly Group Level ? Evidence Regarding Four Conceptual Can Emotions Be Truly Group Level ? Evidence Regarding Four Conceptual Criteria, (May 2014). https://doi.org/10.1037/0022-3514.93.3.431

Smith, L. G. E., Blackwood, L., & Thomas, E. F. (2019). The need to refocus on the group as the site of radicalization. *Perspectives on Psychological Science*, *15*(2), 327–352. https://doi.org/10.1177/1745691619885870

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Southern Poverty Law Center. (2020). *The year in hate and extremism: 2019*. Retrieved from https://www.splcenter.org/sites/default/files/yih_2020_final.pdf

Spears, R., & Tausch, N. (2016). Prejudice and intergroup relations. In *An Introduction to Social Psychology*.

Wiley.
Stahelski, A. (2005). Terrorists are made not born: Creating terrorists using social psychological conditioning. *Cultic Studies Review*, *4*(1), 1–10. Retrieved from http://www.homelandsecurity.org/journal/articles/stahelski.html
Starmer, K. (2012). *DPP's guidance on social media prosecutions*. Retrieved from https://www.scl.org/news/2563-dpp-s-guidance-on-social-media-prosecutions
Stecklow, S. (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*. Retrieved from https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/
Stein, A. A. (1976). Conflict and cohesion: A review of the litrature. *Journal of Conflict Resolution*, *20*(1), 143–165. https://doi.org/10.18574/nyu/9780814786390.003.0007
Stephan, W. G., & Stephan, C. W. (1985). Intergroup Anxiety. *Journal of Social Issues*, *41*(3), 157–175. https://doi.org/10.1111/j.1540-4560.1985.tb01134.x
Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In *Reducing prejudice and discrimination* (pp. 23–45). Lawrence Erlbaum Associates Publishers. Retrieved from https://psycnet.apa.org/record/2000-03917-001
Stephan, W. G., Ybarra, O., & Morrison, K. R. (2009). Intergroup threat theory. In *Handbook of prejudice, stereotyping, and discrimination* (p. 44). Psychology Press. Taylor and Francis Group.
Sternberg, R. J. (2003). A duplex theory of hate: Development and application to terrorism, massacres, and genocide. *Review of General Psychology*, *7*(3), 299–328. https://doi.org/10.1037/1089-2680.7.3.299
Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *International Journal of Press/Politics*, *23*(1), 95–115. https://doi.org/10.1177/1940161217740697
Sun, N., Rau, P. P. L., & Ma, L. (2014). Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, *38*, 110–117. https://doi.org/10.1016/j.chb.2014.05.022
Swann, W. B., Gómez, Á., Seyle, D. C., Morales, J. F., & Huici, C. (2009). Identity Fusion: The Interplay of Personal and Social Identities in Extreme Group Behavior. *Journal of Personality and Social Psychology*, *96*(5), 995–1011. https://doi.org/10.1037/a0013668
Swann, W. B., Jetten, J., Gómez, Ã., Whitehouse, H., & Bastian, B. (2012). When group membership gets personal: A theory of identity fusion. *Psychological Review*, *119*(3), 441–456. https://doi.org/10.1037/a0028589
Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204
Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202
Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In M. A. Hogg & D. Abrams (Eds.), *Key readings in social psychology. Intergroup relations: Essential readings* (pp. 94–109). Psychology Press.
Tausch, N., Hewstone, M., & Roy, R. (2009). The relationships between contact, status and prejudice: An integrated threat theory analysis of Hindu–Muslim relations in India. *Journal of Community & Applied Social Psychology*, *19*, 83–94. https://doi.org/10.1002/casp
Terrorism Act (2006). United Kingdom of Great Britain and Northern Ireland. Retrieved from https://www.legislation.gov.uk/ukpga/2003/21/section/127
The Matthew Shepard and James Byrd, Jr., Hate Crimes Prevention Act (2009). Retrieved from https://www.justice.gov/crt/matthew-shepard-and-james-byrd-jr-hate-crimes-prevention-act-2009-0
The National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2020). Global Terrorism Database. Retrieved November 10, 2020, from https://project-iris.app-staging.cloud/
The Pew Research Center. (2017). *The partisan divide on political values grows even wider*. Retrieved from https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/
The United States Department of Justice. (2021). Federal hate crime laws and statutes. Retrieved from https://www.justice.gov/hatecrimes/laws-and-policies
Tien, J. H., Eisenberg, M. C., Cherng, S. T., & Porter, M. A. (2020). Online reactions to the 2017 'Unite the right' rally in Charlottesville: Measuring polarization in Twitter networks using media followership. *Applied Network Science*, *5*(1). https://doi.org/10.1007/s41109-019-0223-3
Timberg, C., Dwoskin, E., And, A. E., & Demirjian, K. (2017). Russian ads, now publicly released, show sophistication of influence campaign. *Washington Post*. Retrieved from

https://www.washingtonpost.com/business/technology/russian-ads-now-publicly-released-show-sophistication-of-influence-campaign/2017/11/01/d26aead2-bf1b-11e7-8444-a0d4f04b89eb_story.html?utm_term=.866bad232bf5

Tufekci, Z. (2014). Big Questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 505–514.

Tufekci, Z. (2018). YouTube, the great radicalizer. *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html

Turner, J. C., Davidson, B., & Hogg, M. A. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

Turner, J., & Oakes, P. (1986). The significance of the social identity concept for social psychology with reference to individualism, interactionism and social influence. *British Journal of Social Psychology*, *25*, 237–252.

Turner, John C. (1991). *Social Influence*. Brooks/Cole.

Twitter. (2020). Hateful conduct policy. Retrieved November 12, 2020, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Tynes, B. M., Giang, M. T., Williams, D. R., & Thompson, G. N. (2008). Online Racial Discrimination and Psychological Adjustment Among Adolescents. *Journal of Adolescent Health*, *43*(6), 565–569. https://doi.org/10.1016/j.jadohealth.2008.08.021

Uenal, F. (2016). The "Secret Islamization" of Europe: Exploring integrated threat theory for predicting Islamophobic conspiracy stereotypes. *International Journal of Conflict and Violence*, *10*(1), 93–108. https://doi.org/10.4119/UNIBI/ijcv.499

Valente, T. W. (1995). *Network models of the diffusion of innovations*. Hampton Press.

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, 1–26. Retrieved from http://arxiv.org/abs/2004.01670

Vidgen, B., Margetts, H., & Harris, A. (2020). *How much online abuse is there? A systematic review of evidence for the UK*. Retrieved from https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-monitor

Vidgen, B., Tromble, R., Harris, A., Hale, S., Nguyen, D., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *3rd Workshop on Abusive Language Online*, 1–14. Retrieved from https://www.aclweb.org/anthology/W19-3509/

Vinokur, A., & Burstein, E. (1974). Effects of partially shared persuasive arguments on group-induced shifts: A group-problem-solving approach. *Journal of Personality and Social Psychology*, *29*(3), 305–315. https://doi.org/10.1037/h0036010

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *Proceedings Ofthe First Workshop on Abusive Language Online*, 78–84. https://doi.org/10.18653/v1/w17-3012

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wendling, M. (2017). Why is Britain First big online? *BBC*. Retrieved from https://www.bbc.co.uk/news/blogs-trending-42170543

Wendling, M. (2018). *Alt-Right: From 4Chan to the White House* . Pluto Press.

White, R. W. (2001). Social and role identities and political violence: Identity as a window on violence in Northern Ireland. *Social Identity, Intergroup Conflict, and Conflict Reduction*. New York, NY, US: Oxford University Press.

Whitehouse, H., Jong, J., Buhrmester, M. D., Gómez, Á., Bastian, B., Kavanagh, C. M., … Gavrilets, S. (2017). The evolution of extreme cooperation via shared dysphoric experiences. *Scientific Reports*, *7*(February), 1–10. https://doi.org/10.1038/srep44292

Whitehouse, H., McQuinn, B., Buhrmester, M., & Swann, W. B. (2014). Brothers in arms: Libyan revolutionaries bond like family. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(50), 17783–17785. https://doi.org/10.1073/pnas.1416284111

Wiktorowicz, Q. (2005). *Radical Islam rising: Muslim extremism in the West*. Rowman & Littlefield Publishers.

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo-Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Williams, M., & Mishcon de Reya. (2019). *Hatred behind the screens A report on the rise of online hate speech*. Retrieved from https://www.mishcon.com/upload/files/Online Hate Final 25.11.pdf

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/0270467610380011

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? *ArXiv*. https://doi.org/10.1145/3184558.3191531

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., … Blackburn, J. (2017). The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. *Proceedings of the 2017 Internet Measurment Conference*. Retrieved from http://arxiv.org/abs/1705.06947

Chapter 1

Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters

Gallacher, J. D., Heerdink, M. W., & Hewstone, M.,

Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters (2021) *Social Media + Society., p.1:16*

https://journals.sagepub.com/doi/full/10.1177/2056305120984445

Abstract

The rise of the Internet and social media has allowed individuals with different backgrounds, experiences, and opinions to communicate with one another in an open and largely unstructured way. One important question is whether the nature of online engagements between groups relates to the nature of encounters between these groups in the real world. We analysed online conversations that occurred between members of protest groups from opposite sides of the political spectrum, obtained from Facebook event pages used to organize upcoming political protests and rallies in the United States and the United Kingdom and the occurrence of violence during these protests and rallies. Using natural language processing and text analysis, we show that increased engagement between groups online is associated with increased violence when these groups met in the real world. The level of engagement between groups taking place online is substantial, and can be characterized as negative, brief, and low in integrative complexity – the latter reflecting one-sided or overly simplistic viewpoints. These findings suggest that opposing groups may use unstructured online environments to engage with one another in hostile ways. This may reflect a worsening of relationships, in turn explaining the observed increases in physical violence offline. These findings raise questions as to whether unstructured online communication is compatible with positive intergroup contact, and highlights the role that the Internet might play in wider issues of extremism and radicalization.

Introduction

The rapid expansion of digital communication technologies and the Internet allows individuals to connect and interact in ways that were previously impossible. Today, people can communicate, share ideas, and participate in political discussions from almost anywhere on the globe. This has the potential to have either positive or negative effects on social cohesion and social integration. When in its infancy, it was hoped that the increased interpersonal connection made possible by social media would bring about global expansions in democracy, highlighted by its role in promoting the Arab spring (Howard et al., 2011). Today this idea has faded, with social media instead seen as posing a fundamental threat to democracy by driving social polarization, disinformation and hostility (Guess, Barber, et al., 2018; Sunstein, 2017). These opposing optimistic and pessimistic views are difficult to reconcile as the exact relationship between online interactions between groups and offline group dynamics is not known.

In this study we explore how opposing groups use the online environment to engage spontaneously with one another, what the nature of this engagement is, and whether it relates to group behaviour in the real world. Specifically, we investigate whether the degree of inter-group engagement on social media is associated with offline violence between rival groups when they subsequently meet in the real world, and whether the qualities of this conversation moderate such effects. To answer these questions, we analyse conversation data from Facebook event pages preceding 25 recent political protests and rallies in the United Kingdom and the United States, all consisting of a right-wing protest group and a left-wing counter protest. We use a text classifier to estimate the level of outgroup engagement (defined as communication between members of opposing groups), and a combination of established and novel text analysis techniques (natural language processing) to calculate four measures of conversation quality on 73,657 posts within Facebook event pages.

Literature review

We organize this brief review of the literature around four key topics. First, we consider the relationship between online communication between members of a group and this same group's members' offline group behavior. Second, we consider evidence for the efficacy of online contact in improving relations between groups. Third, we review evidence that communication via the Internet plays a key role in co-radicalization, making both groups involved hold even more extreme worldviews. Finally, we propose how natural language processing can be used to provide a quantitative and objective measure of the types of online conversations between opposing political groups.

The relationship between online communication and offline group behaviour

Online social media activity within groups has been shown to correlate with offline group behaviour in cases of social mobilization and political change. Multiple studies have found that increased online activity, often on Facebook and Twitter, has been associated with subsequent increases in protest attendance at a later date. These include the pro-democratic movements of the Arab spring (Steinert-Threlkeld et al. , 2015), anti-capitalist and economic inequality protests in the United States and Spain (Bastos et al., 2015), and anti-government protests in Ukraine (Gruzd & Tsyganova, 2015). Digital connectivity has been identified as a driving factor in how these social movements connect, organize and evolve (Tufekci, 2017), as it reduces the costs of organization and allows activity to erupt in spontaneous and unexpected ways (Enikolopov et al., 2016). In addition, the nature of social

media conversations has been shown to be related to the future number of hourly arrests at prolonged one-sided political protests that descend into arson and vandalism (Mooijman et al., 2018); within group conversations which become more moralized may have made this violence appear more socially acceptable (Mooijman et al., 2018), and social media allows both the signalling and gauging of the moral sentiment of others (Barberá et al., 2015). Similarly, violence towards immigrants within western countries has been shown to be related to the degree of anti-refugee sentiment expressed on social media in areas where the violence takes place (Müller & Schwarz, 2020a), while in the United States, anti-Muslim messages disseminated by President Trump over social media correlate with the number of anti-Muslim hate crimes in states where social media usage is high (Müller & Schwarz, 2020b), although the temporal order of online and offline measures here is unclear. These cases show how the online environment may not only affect online behaviours but spreads offline as well.

The contact hypothesis online

Those who espouse the optimistic view that social media can bring about increased social cohesion highlight that by providing new opportunities for individuals from different groups to gather and interact with members from other groups, the Internet could potentially play an influential role in increasing contact, and breaking down barriers between groups (Amichai-Hamburger & McKenna, 2006; Schwab & Greitemeyer, 2015). The contact hypothesis proposes that positive face-to-face contact between members of different groups provides one of the best ways to improve relations between these groups if certain facilitating (rather than essential) conditions are met: equal status between groups, the sharing of common goals, intergroup co-operation, personal interaction and support from authorities (Allport, 1954). The efficacy of offline contact in improving intergroup relations has been demonstrated with groups that differ in terms of, for example, race and religion (Brown et al., 2003), sexual orientation (Herek & Glunt, 1993), and has been confirmed in experimental (Ioannou et al., 2017) and longitudinal (Ramiah & Hewstone, 2013) research and meta-analyses (Davies et al., 2011; Pettigrew & Tropp, 2006). It has also been demonstrated for political views (Sønderskov & Thomsen, 2015); for example, intergroup contact between liberals and conservatives reduced hostility and improves attitudes to opposing parties in the United States (Manbeck et al., 2018), and positive contact with EU nationals was associated with support for EU membership during the 2016 Brexit referendum (Meleady et al., 2017). If these same offline effects carry across to the online world, then positive change offline could follow online intergroup contact.

In highly controlled settings, intergroup contact via the Internet has been shown to have positive effects on intergroup relations (White & Abu-Rayya, 2012). This highlights the opportunities offered by the Internet for positive intergroup contact in cases where physical contact is restricted by geographical, political and economic barriers (Austin, 2006; Hoter, 2009). Amichai-Hamburger and McKenna (2006) go one step further in their "Internet contact hypothesis", and outline not only how all of the conditions for positive intergroup-contact can be satisfied in an online context, but also propose that online contact has an advantage over offline contact as it allows various features to be manipulated in order to create optimal contact conditions. This hypothesis is echoed in the optimistic view of online interaction promoted by the world's largest social media platform, Facebook, which announced that it "is proud to play a part in promoting peace by building technology that helps people better understand each other. By enabling people from diverse backgrounds to easily connect and share their ideas, we can decrease world conflict in the short and long term" (Facebook, 2010).

However, Amichai-Hamburger also warns of the potential risks from misuse of digital platforms, and advocate the careful selection of discussion participants, supervision, and prior agreement to stay on topic and avoid 'flaming', which is the act of posting insults, profanity or offensive language with the intention to seek out a negative reaction from the reader (Amichai-Hamburger, 2008). The open and unsupervised nature of social media platforms cannot provide these controls. Social media platforms instead provide the opportunity for unrestrained interactions, often in an anonymous format (users can select their own usernames or profile images, not necessarily linked to their real identity) with very few limitations on the type of language used or moderation of inflammatory content. In such situations negative rather than positive intergroup contact may occur, and while positive intergroup contact can reduce prejudice, negative intergroup contact may increase it (Graf et al., 2014; Paolini et al., 2010). Negative contact increases the salience of an outgroup individual's group membership, and so any negative effects of contact generalize more strongly to the group as a whole (Brown & Hewstone, 2005). While negative intergroup contact is less common than positive intergroup contact in the real world (Graf et al., 2014), this may not be true online. For example, analysis of Facebook groups about the Israeli-Palestinian conflict found little evidence for positive intergroup contact, but rather evidence of hateful antagonistic positions and intolerance (Ruesch, 2011), Within the Facebook pages dedicated to this conflict analysed by researchers, most content was dedicated towards intragroup mobilization and declaration, and although some pages did self-categorize as 'peace

groups', with the stated goal of promoting intergroup dialogue, these pages were much less popular than highly-partisan pages.

Evidence shows that in large online environments small numbers of communities initiate a large proportion of the intergroup communication, and this communication is often distinctly hostile (Kumar et al., 2018). Social media users in the United States interviewed about interactions with political outgroup members reported that they were stressful and frustrating, and that other users with whom they interacted online were angry and disrespectful (Duggan et al., 2016). This is reflected in evidence that artificial exposure to opposing views on Twitter can increase political polarization (Bail et al., 2018).

Because positive consequences of online engagement with other groups are possible, but certainly not guaranteed, we investigated how naturalistic engagement with outgroup members via social media relates to real world intergroup behaviour. With increases in polarization (Dimock et al., 2014), fragmentation, and extremism (Poushter et al., 2015) throughout the western world, understanding the impact of online connectivity is paramount to inform us about and counter social division. In this study, we provide a unique insight into this question by examining the online relationships between opposing groups at the extremes of the political spectrum. We define this extremism as a belief that ingroup survival is inseparable from a need for hostile action against an outgroup (Berger, 2018a). These hostile actions can vary from discriminatory behaviour to verbal attacks or violence (Berger, 2017). Right-wing extremist violence has recently increased across the world (Muhlhausen & McNeill, 2011; Neiwert et al., 2017), and for right-wing extremists the Internet has become a vital tool used to radicalize, recruit, mobilize, and network (Berger, 2018b; Conway & Courtney, 2017). Over time their language has become more aggressive and is associated with a sharp rise in online hate crimes (EUROPOL, 2017). As a result, one of most prominent far-right groups (Mudde, 2007; Mudde, 2019) in the United Kingdom, Britain First, has recently been banned from the two largest social media sites: Facebook (Facebook Newsroom, 2018) and Twitter (BBC News, 2017). Recognizing the impact of online interactions, Facebook reported that it took this action because the group was sharing hate speech designed to stir up division.

Co-radicalization and the Internet

It has been suggested that groups from opposite ends of the political spectrum 'feed off' one another other in a process of co-radicalization (Knott et al., 2018; Pratt, 2015), a two-way process where different groups reciprocally construct increasingly radicalized worldviews (also referred to as cumulative extremism or mutual radicalization). Often co-radicalizing groups use actions of the other group to justify their own behaviours or prejudices (Ebner, 2017). Offline, this can result in violence by one group being met with violence by the other (Bundesministerium des Innern, 2015).

Evidence for co-radicalization in online spaces is limited, but there is increasing evidence for its occurrence. For instance, areas of the United Kingdom, Germany, Belgium and France with larger far-right communities and greater anti-Muslim hostility offline also have greater levels of pro-Islamic State content online (Mitts, 2019). This remains the case even when accounting for socio-economic factors such as unemployment and income, suggesting that there may be a link between the offline prejudice and online radicalization.

Additionally, there are indications that both far-right and Islamic extremist groups online make sustained references to the other group, with both sides blaming and demonizing the other, and spreading sentiments of victimization and conspiracy theories about the other group (Fielitz et al., 2018). Similarly, there is evidence that in online spaces overtly hostile language is used to counter hate speech in 39.7% of cases, and this is often met unfavourably – leading to a further reduction in relations rather than the intended improvement (Mathew et al., 2019).

While this evidence suggests that contact between opposing groups online may facilitate co-radicalization, little has been done to research the nature of direct interactions between opposing groups, and it is these between-group interactions which are a key feature of how the co-radicalization process occurs (Moghaddam, 2018). Our study investigates the nature of these direct interactions in an unstructured and relatively unmoderated online space.

For far-right groups, opposition groups such as Muslim communities or anti-fascist counter protestors are often framed as 'extreme' and as posing an existential threat to the far-right ingroup in order to legitimize radical responses (Jackson, 2018). The Internet is moreover becoming recognized as an

important facilitating factor in this process (Sirseloudi, 2017), extending the reach of activists, allowing for international cooperation between ideological allies and increasing opportunities for radicalization (Briggs & Strugnell, 2011; Von Behr et al., 2013). This makes far-right groups and counter protest groups an ideal case with which to study the association between online outgroup engagement and digital and real-world group dynamics.

The present research

In the current study we make use of a number of natural language measurements which are particularly useful when studying intergroup relations and online intergroup conflict. Firstly, we measure sentiment, a broad indication of how positive or negative a post/comment is. Evidence from intergroup contact research shows that positive interaction between opposing groups is more likely to lead to a reduction in group hostility, while negative interaction can have the reverse effect (Brown & Hewstone, 2005). Secondly, we measure integrative complexity (IC), which quantifies the ability of an individual to think and reason with input from multiple perspectives (Streufert & Suedfeld, 1965), and has proven successful in measuring cognitive complexity in situations ranging from international relations and electoral competition to political revolutions (Suedfeld, 2010). The level of IC presented in online communication can provide information about the extent to which authors hold radical or extremist views (Smith, Suedfeld, Conway, & Winter, 2008), and changes in IC are predictive of international violence (Guttieri et al., 1995; Suedfeld & Bluck, 1988) as well as intergroup conflict (Tetlock et al., 1993). As such, IC may be an important moderator between online outgroup engagement and subsequent improvements in group relations. Finally, we use a measure of online incivility: toxicity. This is defined as a measure of how likely a comment is to make someone leave a conversation, with comments that are defined as being more rude, disrespectful, or unreasonable being more likely to receive a higher 'toxicity' score (Wulczyn et al., 2017). This is similar to negative sentiment but goes a step further by including the detection of personal attacks and harassment. As such this is likely to be a useful metric when measuring intergroup communication, as it gives an indication of how antagonistic the communication is.

Here we use these metrics to investigate how opposing groups engage online with one another, and test whether the quantity and quality of their online communications is linked to their behaviour when they meet in the real world. Specifically, we look at opposing political groups engaging on Facebook and, based on the idea that hostile communication can further divide already opposed

groups, we hypothesize that more communication online will be associated with more violence offline, when members of these groups meet later in the real world. We further predict that the relationship between outgroup engagement and subsequent violence would be moderated by IC, toxicity, and sentiment whereby lower IC, and higher toxicity and sentiment, are associated with greater violence.

Methods

Sampling

Twenty-five physical events were selected for analysis that occurred between October 2015 and October 2017. Each event consisted of a right-wing protest, march or rally that occurred with a corresponding counter-march or protest by the opposing political side, organized in tandem, on the same day and at the same location. Of these 25 events, 20 occurred in the United Kingdom and five occurred in the United States. Events were selected for the UK from the most active street-protest groups from each side of the political spectrum, and in the US, events were selected which occurred in response to the 'Unite the Right' rally in Charlottesville in August 2017. [See Supplementary Information (SI) 1]

Once the political events were identified, the conversations taking place online were collected. This was done via the Facebook Graph Application Program Interface (API). We collected conversations taking place on the Facebook pages that were set up to promote the event. This collection method gathered 73,632 comments in total with an average of 1,473 comments per event page, and a total of 2,946 comments on average per event. Once collected, the data were cleaned to remove any conversation that occurred after the planned start time of the event. In doing this we can safely assume that any violence at the event had no impact on conversation online. [See SI 1]

Facebook Event Pages are ideal for our study because they are unique in how the online space is linked directly to an offline event, and how they are chronicled and remain accessible for past events. Facebook pages are one of the primary places where online engagement between opposing groups occurs, with the social media platform being used not just for communication within the group but as a primary means for group mobilization. Here we focus specifically on Facebook event pages, which are the primary method through which groups plan and disseminate information about upcoming marches, protests and rallies, and which, crucially, allow for public discussion and hence for members of differing groups to communicate. While it is possible that other social media platforms may also be used to promote and coordinate these events (such as Twitter), the link between online conversation and offline event on these other platforms is more ambiguous, and inferences about this relation would therefore be more difficult to draw.

Text analysis measures

In order to allow for comparison between events all text analysis measures were coded at the comment level, and then aggregated to the event page level. This resulted in 50 data points for each measure in total (25 right-wing pages, and 25 left-wing pages). These were subsequently aggregated to the event level using a BLUP-based method (Croon & Van Veldhoven, 2007).

Conversation tone and sentiment extremity

Sentiment analysis was performed using RSentiment (Bose, 2017) for R (Version 1.1.383, 2017), which classifies each comment into very positive, positive, neutral, negative, very negative categories using a 'parts of speech' tagging system. It first classifies each word in the sentence as one of the above categories, and then calculates the overall classification of the comment. To account for negation, the package checks whether each word has been preceded by any negative quantifier and if so, adjusts the score accordingly (Bose et al., 2017).

While this analytic approach does not fully overcome the limitation that sentiment analysis tools lack context as they look at each message in isolation, we also take the average score for the entire conversation for each event page to reduce the impact of individual misclassified messages. From the classification of individual comments, a single 'tone' value was calculated for each event page. This tone value ranged on a scale from one to five and was calculated by assigning values from one to five for comments from very negative to very positive and calculating the average score per event page. An event page score of one would represent a page with 100% very negative comments, while a score of five would represent a page with 100% very positive comments. In order to account for the fact that positive and negative sentiments are often not mutually exclusive (Berrios, Totterdell, Kellett, & Brose, 2015), a sentiment extremity score for each event page was then calculated. This was done by calculating the percentage of comments within the page that were classified as either very positive or very negative. This second measure, ranging from 0 to 100, therefore gives an indication of the emotional extremity of the conversation.

Integrative Complexity (IC)

We used an automated IC scoring system, AUTO IC (Gideon et al., 2014; Houck, 2014), to generate IC scores for each comment within an event page, from which the mean IC for the entire conversation on the event page was calculated. The Automated IC system produces a score from one to seven for

each comment. This uses the same scoring methodology as human-scored IC. In both systems, scores of one represent a total lack of differentiation (acknowledgement of different viewpoints) or integration (combination/connections of multiple viewpoints). Scores from two to three represent levels of increasing differentiation, but no integration. Scores from four to six represent increasing and moderate to high levels of differentiation and integration. A score of seven indicates high differentiation plus high integration.

Individuals who display higher IC tend to construct more accurate and balanced perceptions of other people, use more information when making decisions, as well as holding less extreme views, and as a result these individuals are shown to be less prejudiced and are better able to resolve conflicts cooperatively with outgroup members. Furthermore, within-group discussions with higher levels of IC have been shown to decrease displays of greed and fear, and reduce the likelihood that a group would decide to take a competitive stance against others (Park & DeShon, 2018). Recently, AUTO IC has been used successfully for the study of online terrorist content, demonstrating the validity of the application to the digital domain (Houck et al., 2017).

Toxicity

We used the Google Perspective API (Google Project Jigsaw, 2018) to measure the level of toxicity within the online conversations. This classification tool was designed by Google's 'Project Jigsaw' and 'Counter Abuse Technology' teams with the aim of promoting better discussions online (Wulczyn et al., 2017). The model gives a toxicity score for each comment on a scale ranging from zero (least toxic) to one (most toxic). In the current study, each comment was sent through the Perspective API, and from this the average toxicity rating for each event page calculated.

Outgroup engagement

In order to identify occurrences of outgroup engagement we trained a neural network (Sebastiani, 1999) to classify comments as either 'within-group', for comments that were directed towards other ingroup members, or 'between-group', for comments that were directed towards a member of the outgroup. A 'between-group' comment could therefore be either a member injecting a comment into the event page of the opposing group, or a reply to this injection from a member of the incumbent page. The proportion of between-group comments for each event page was calculated. *[See SI 5 for further details of how this measure was operationalized].* We define this type of communication

between members of opposing groups as 'outgroup engagement' as it falls short of the level of interpersonal involvement and connection required for traditional intergroup contact but shares some important characteristics with it.

The neural network was trained on a set of 1,000 randomly sampled comments from the overall dataset to ensure that group-specific language and idioms were accurately interpreted, as such elements may be misinterpreted by generic lexicon measures (Omand, Bartlett, & Miller, 2012). Each comment in the training set was human coded. The default coding option was within-group communication, and this was selected in all cases where a decision could not be made (either through a lack of information or clarity). To ensure accuracy of the human coding, all comments were coded by a second coder who was blinded to the hypotheses, and inter-coder reliability (ICR) scores calculated. For the training set the ICR was 97.80% with a Scott's PI of .96. For the test set the ICR was 95.90% with a Scott's PI of .90. *[See SI 5].*

The overall accuracy of the classifier was 89.0%, with a sensitivity of 85.9% and a specificity of 89.9% when checked against representative test set of 1,000 comments. This is therefore a conservative judgement classifier with regard to outgroup engagement classification, reflecting the conservative nature of setting within-group conversation as the default.

As an additional measure of classification consistency, we calculated the proportion of comments for each user within an event page that are given the same label by the classifier (as either within-group or between-group communication). Overall, we found 91.0% consistency in these ratings. We judged this to be a high level, especially as we would not expect consistency to reach 100%; 'home' users with the event page started by their own ingroup should be classified differently when replying to an outgroup member that visits a page compared to when replying to an ingroup member on the same event page.

Similar approaches using machine learning to code digital text have previously been shown to be valid with regard to online comment abuse detection (Chu et al., 2017), machine translation (Wołk & Marasek, 2015), and sentiment analysis (Kim, 2014), but to the best of our knowledge, this is first time such methods have been used to identify cases of online intergroup engagement.

Violence

We developed two measures of violence for each real-world event. In all cases this was based on open source intelligence taken from a range of sources, including professional journalistic reports, citizen journalism, photos, videos, and police reports on arrest statistics *[See SI 2].* The first violence measure is a binary 'absence or presence of violence' (0 = absent, 1 = present) based on whether reports stated that violence occurred at the event. The second measure allocates a degree of violence score based on seven security industry-standard violence indicators. We fit a latent trait model to these indicators, which assumes that each event has an unobserved (latent) true level of violence that manifests itself in the absence or presence of these indicators. The model determines two parameters for each indicator: the 'severity', reflecting the level of violence at which this indicator is likely (>50%) to be present, and 'discrimination', which reflects the sensitivity of the indicator to changes in violence. These parameters are then used to estimate the latent violence score for each event on a continuous scale. *[See SI 2]*

We checked the robustness of this measure using sensitivity analysis, whereby each indicator was removed in turn and the analysis repeated. The results were very similar in all cases, suggesting that no individual indicator is responsible for the observed effects *[See SI 4].*

Statistical methods

All statistical analyses were conducted in R. For each model the optimum combination of predictors (text analysis measures) was selected using Akaike Information Criterion (AIC).

We tested the predictive power of text analysis measures, including outgroup engagement, on violence using a logistic regression (generalized linear model, GLM) *[SI 2].* This analysis was then repeated replacing presence/absence of violence with degree of violence, this time using a linear model (LM). To test for moderation effects of the quality of both within-group and between-group conversations on violence measures we tested for interaction effects using a GLM and LM. We used paired *t*-tests to compare the nature of the comments in between-group and within-group communication within the same event.

The total conversation size within the event page discussion (number of comments) was not found to explain any variance and therefore was not included in these models. All events contained a degree of

intergroup contact; however, five event pages (representing approximately half the event conversation in each case) contained no intergroup contact. Therefore, when comparing the qualities of the intergroup contact with subsequent violence for these events we used the values of the event page that did contain contact and did not aggregate across pages. When comparing the length of continuous chains of comments, a non-parametric Mann-Whitney U independent samples test was performed to account for a negatively skewed nature of the comment chain length distributions.

The models were shown to be robust through the absence of influential data points and multicollinearity [SI 4]. Where we detected multicollinearity (testing for moderation effects) we resolved this by centering the predictors prior to analysis. Throughout the results all measures are shown as estimate ± SE.

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Ethics Reference: R55162/RE001). All data collection was conducted using open source methods and publicly available data, and hence, informed consent was not explicitly obtained. No privacy infringements were made, no private groups were joined, and no accounts 'befriended' in order to access data that are not publicly available.

Results

Overall, 32.0% of comments across all event pages were classified as outgroup engagement. In order to understand better what type of outgroup engagement occurred, we compared the between-group conversation on an event page to the within-group conversation on the same page. We found that compared to within-group conversation, between-group conversation displayed a higher level of toxicity (paired *t*-test, between-group $M = 0.47 \pm 0.02$, within-group $M = 0.20 \pm 0.01$, df = 44, $t = 12.36$, $p < .001$, $d = 1.84$) and a higher level of IC (paired *t*-test, between-group $M = 1.46 \pm 0.04$, within-group $M = 1.18 \pm 0.01$, df = 44, $t = 6.95$, $p < .001$, $d = 1.04$). With regard to tone, between-group conversation was slightly more negative than within-group conversation, but this difference was not significant (paired *t*-test, between-group $M = 3.37 \pm 0.07$, within-group $M = 3.52 \pm 0.04$, df = 44, $t = -1.89$, $p = .065$, $d = 0.28$) (Figure 1). Additionally, in order to identify whether there was a difference in the duration of between-group conversations and within-group conversations, we calculated the average length of a continuous chain of between-group comments, and found this was shorter than the average length of a continuous chain of within-group comments (Wilcoxon signed rank test, between-group: 2.58 ± 0.02, within-group: 3.44 ± 0.03, $U = 45144000$, $p < .001$). This suggests that outgroup engagement often consists of short-lived interjections in the other group's discussions, which invite a prompt response.

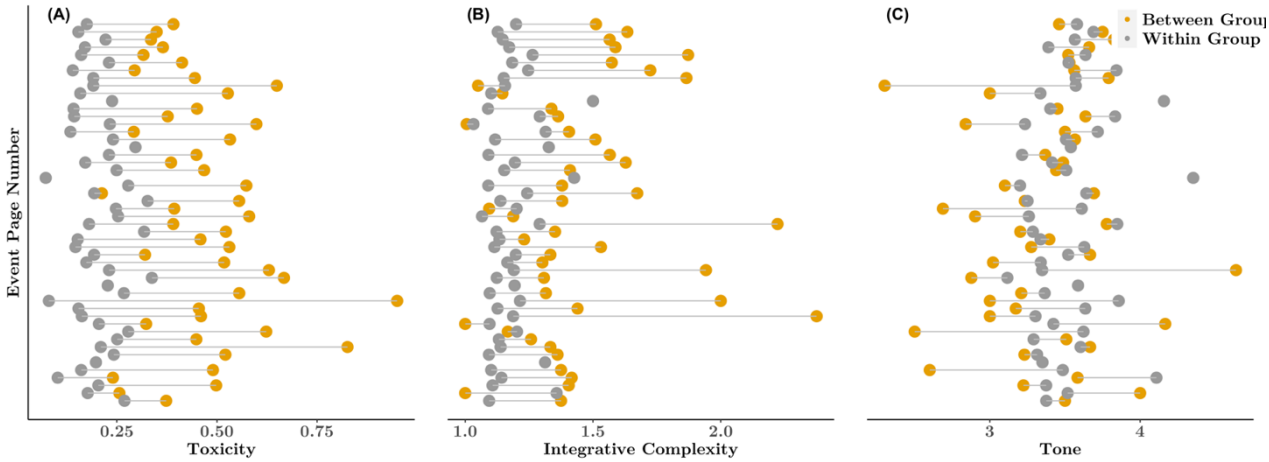


Fig 1. The differences in conversation qualities for between-group conversation and within-group conversation for each event page.

Dumbbell plots show that (A) Toxicity is higher in between-group conversation, (B) Integrative Complexity is higher in between-group conversation, and (C) There is no difference in Tone for between-group communication and within-group conversation.

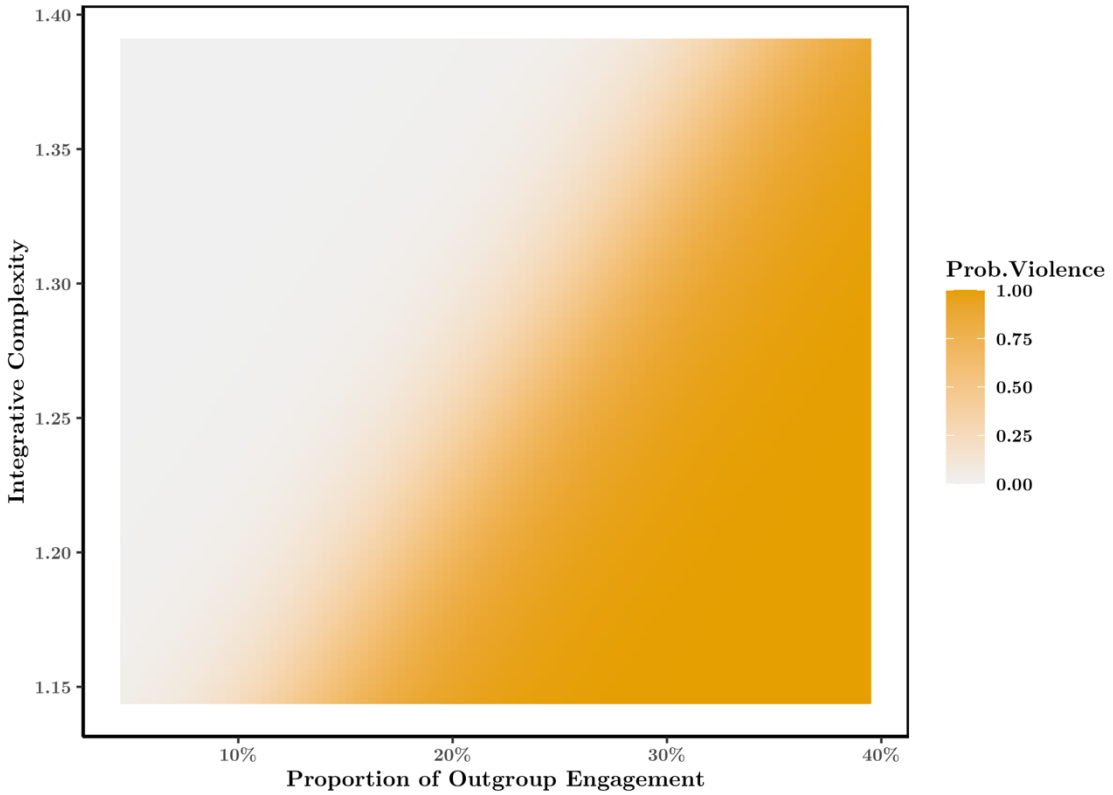


Fig 2. Probability of an event becoming violent for levels of outgroup engagement and Integrative Complexity in preceding conversations on Facebook event pages.

Offline violence is more probable when there is more online outgroup engagement prior to the event, or when the integrative complexity of the online discussion is lower.

In order to test whether conversation qualities and the proportion of outgroup engagement with a Facebook event page are associated with offline violence, we first measured these variables separately on the left-wing and right-wing pages relating to each event [*see SI 7, Figure S4 & Table S7, for differences in conversation metrics between right-wing and left-wing pages*], and then aggregated these scores to the event level by calculating best linear unbiased predictors (BLUPs); we then tested their association with offline violence using a logistic regression *[SI 4]*. We used a stepwise method to compare the ability of different combinations of variables assessing the nature of online conversations to statistically predict offline violence later in time, and selected the best model based on AIC.

Real-word violence was associated with two conversational variables: the level of outgroup engagement (i.e., the proportion of between-group conversation on a page) and the integrative complexity (IC) of the conversations. Violence was more likely if conversations previously had higher levels of outgroup engagement and lower levels of IC (Figure 2) (GLM, $n = 25$, outgroup engagement: $B = 0.38 \pm 0.16$, Wald's $z = 2.33$, $p = .020$; IC, multiplied by 10 to account for the limited range [1.13 to 1.39 on a 1-7 scale]: $B = -3.20 \pm 1.60$, Wald's $z = -2.00$, $p = .046$; notation:

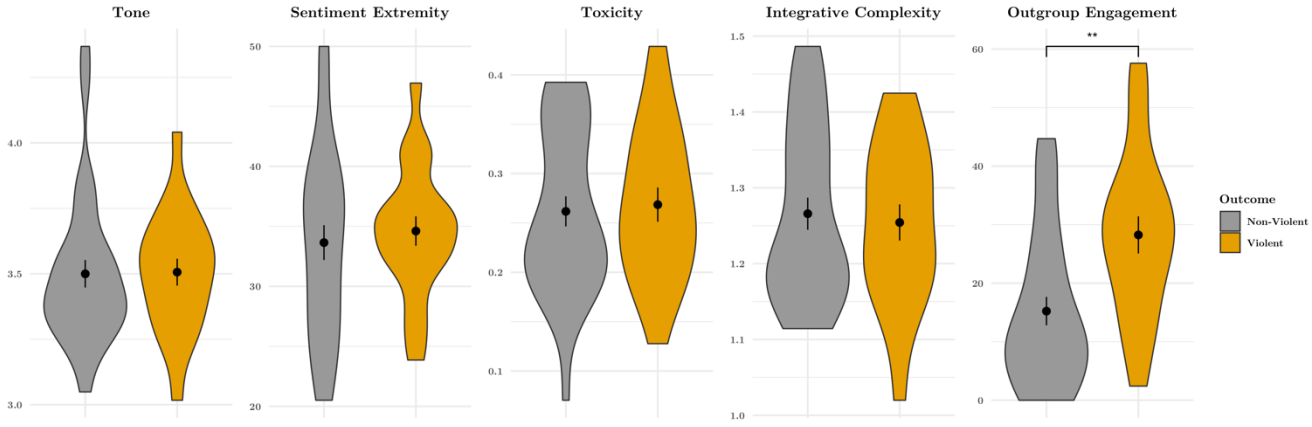


Fig 3. Comparison of violent vs non-violent conversation qualities.
*Differences in Tone, Sentiment Extremity, Toxicity, Integrative Complexity and outgroup engagement occurring on Facebook event pages for events which subsequently became violent vs. those which remained peaceful. Mean and standard error are given by dots and lines, respectively. ** denotes $p < 0.001$*

estimate ± SE). In other words, each 1.0% increase in the proportion of outgroup engagement on an event page increased the odds of the event becoming violent by a factor of 1.46, while each 0.1 unit increase in the average conversation IC decreased the odds of violence occurring by a factor of 24.50. Model selection did not retain tone, sentiment extremity or toxicity as variables associated with subsequent violence. Figure 3 illustrates this result, such that events which became violent displayed higher levels of outgroup engagement (Welch two sample *t*-test, violent $M = 28.25 \pm 2.71$, non-violent $M = 15.23 \pm 2.12$, df = 24, t = -3.78, $p = .001$, $d = 1.56$) and lower levels of IC than events which remained peaceful (Welch two sample *t*-test, violent $M = 1.27 \pm 0.02$, non-violent $M = 1.25 \pm 0.02$, df = 24, $t = 0.40$, $p = .70$, $d = 0.17$), although the latter is only significant in the full GLM.

To determine whether the two variables, outgroup engagement and IC, were also associated with the *degree* of violence, we developed a continuous measure of violence based on the standard indicators of violence used in the security industry *[See SI 2]* and repeated the analysis. In the best model according to AIC, only outgroup engagement was significantly statistically associated with the degree of violence, with more outgroup engagement being associated with more violence (Figure 4, LM, $n = 25$, $B = 0.05 \pm 0.01$, $t_{23}3.30$, $p = .003$). The distribution of these violence metrics across event pages is shown in the supplementary information (SI) table S3 and table S4.

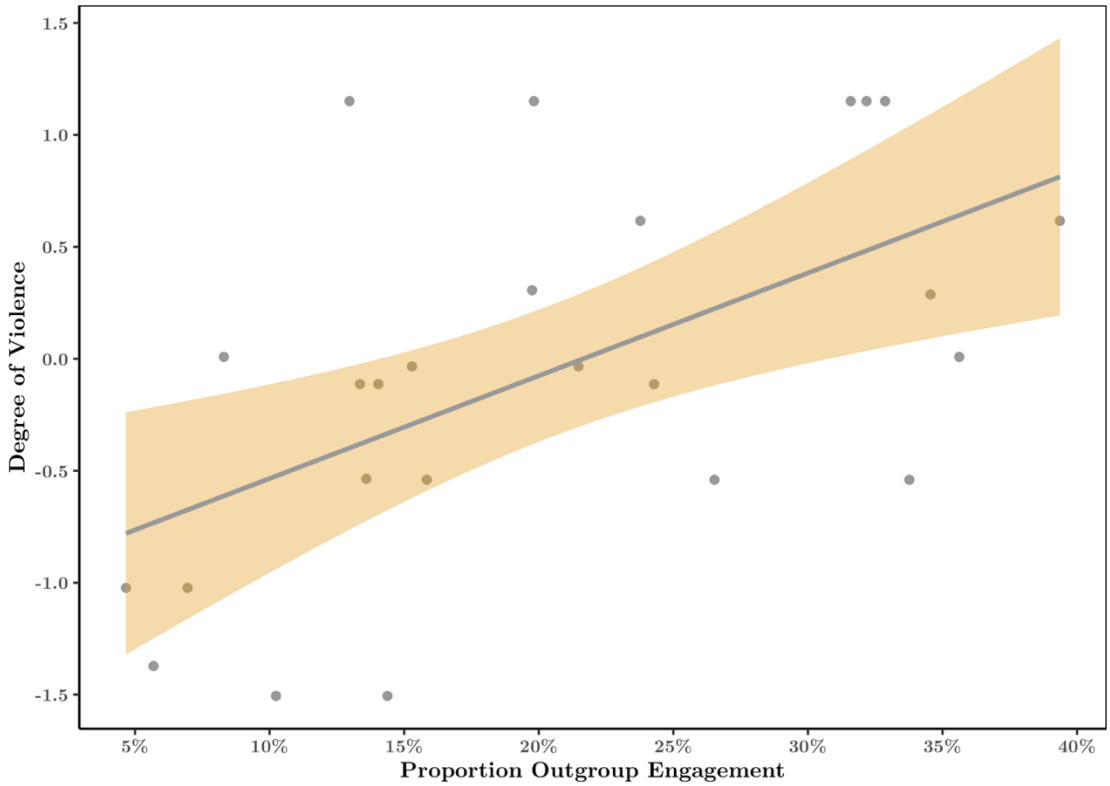


Fig 4. Outgroup engagement and degree of subsequent violence.
The level of outgroup engagement in Facebook event pages preceding a political protest is associated with the degree of violence at that event

We hypothesized that the relationship between outgroup engagement and subsequent violence would be moderated by IC, toxicity, and tone. We therefore tested whether any interactions between outgroup engagement and these metrics in online conversations were associated with the degree of offline violence. We found no significant interactions between outgroup engagement and IC, and outgroup engagement and tone (LM, $n = 25$, outgroup engagement×IC; $B = -0.29 \pm 0.24$, $t_{21} = -1.24$, $p = .230$, outgroup engagement×Tone $B = 0.07 \pm 0.13$, $t_{21} = 0.52$, $p = .607$), but a significant negative interaction between outgroup engagement and toxicity (outgroup engagement×toxicity; $B = -0.81 \pm 0.38$, $t_{21} = -2.15$, $p = .044$). However, this latter model did not account for more variance in the degree of offline violence than outgroup engagement in isolation. We also tested whether interactions between outgroup engagement and conversation quality were associated with the presence rather than degree of violence, and found that they were not (GLM, n=25, outgroup engagement×IC; $B = -0.28 \pm 1.20$, Wald's $z = 0.24$, $p = 0.814$, outgroup engagement×Toxicity; $B = -2.74 \pm 2.21$, Wald's $z = -1.24$, $p = 0.215$, outgroup engagement×Tone; $B = -0.18 \pm 0.68$, Wald's $z = 0.26$, $p = .792$). In order to confirm that the quality of within-group conversations was not masking the role of quality of between-group conversations (outgroup engagement), we tested for interactions between the

quality of only the between-group conversations for each event and degree of violence, and again found no evidence for moderation (LM, $n = 25$, outgroup engagement×between-group IC; $B = 0.002 \pm 0.08$, $t_{21} = -0.03$, $p = .980$, outgroup engagement×between-group toxicity; $B = -0.22 \pm 0.20$, $t_{21} = -1.15$, $p = .265$, outgroup engagement×between-group tone; $B = -0.005 \pm 0.06$, $t_{21} = -0.10$, $p = .923$). Thus, our hypothesis that conversation quality moderates the effect of outgroup engagement was not supported.

Given the negative relationship between IC and the occurrence of violence, and the higher level of IC in between-group conversation compared to within-group conversation, we investigated if either the IC of within-group conversations or the IC of the between-group conversations was more strongly associated with subsequent violence. We tested this while accounting for the level of outgroup engagement, and found that only the IC of the within-group conversations was associated with the presence/absence of violence (within-group IC $B = -7.54 \pm 3.60$, Wald's $z = -2.10$, $p = .036$, between-group IC $B = 0.07 \pm 0.43$, Wald's $z = 0.17$, $p = .864$). In fact, replacing the IC of the overall conversation with IC of the within-group conversations improved the association between the model's specified variables and violence (ΔAIC = -7.70).

Discussion

In this study, we examined whether the frequency and quality of naturally occurring online conversations between opposing political groups on Facebook is associated with offline physical violence at subsequent real-world events. We found that the level of outgroup engagement on Facebook event pages was the variable most consistently associated with both the presence and degree of violence during a subsequent encounter between the groups. Overall, we found that the Facebook event pages used by groups to mobilize and gather support for a march or rally are a place where online outgroup engagement occurs, with 32.0% of the overall conversation across all pages occurring between members of opposing groups. We consider this a substantial percentage, given that previous studies estimated that within right-leaning communities on Twitter up to 93.0% of interactions occur between ingroup members (Conover et al., 2011). The extent of this communication between groups, and the societal impact of the violence associated with this communication, emphasizes the importance of studying such online conversations.

Conversation quality and subsequent offline violence

In addition to the quantity of communication between groups, our findings show that conversation quality is also associated with future violence. Specifically, lower levels of integrative complexity (IC) in the conversation were associated with violence during the offline encounter. This aligns with previous findings that linked decreases in IC with the deterioration of group relations, ranging from more competitive intergroup behaviour (Park & DeShon, 2018) to international violence (Suedfeld & Bluck, 1988). Individuals who display higher IC tend to construct more accurate and balanced perceptions of others, use more information when making decisions, and hold less extreme views. As a result, these individuals are shown to be less prejudiced and are better able to resolve conflicts cooperatively with outgroup members (Tetlock et al., 1993). While high IC is not traditionally held as one of the conditions required for, or mediators of, positive intergroup contact, it measures the ability to think and reason with input from multiple perspectives, and this ability to take others' perspective and empathize with them is a key mediator of how contact improves outgroups attitudes (Pettigrew & Tropp, 2008). It is possible that in an online environment stripped of much individuating and subtle information the ability to overtly demonstrate multiple viewpoints becomes critical.

When comparing the between-group and within-group communication on an event page, we found that between-group communication was more toxic, indicating that it is more rude, aggressive, or

disrespectful to the outgroup. Interestingly, this same communication was also higher in IC, suggesting that while this communication is quite negative, it also engages with opposing views to a greater extent than when ingroup members speak with each other. Furthermore, we found that the IC of the between-group communication was not associated with greater violence, but rather it was the IC of the within-group communication which was, and indeed to a greater extent than the IC of the overall conversation. This suggests that the IC of the conversations taking place between ingroup members is most directly related to group behaviour. It may be that less complex conversations reflect an increased homogeneity of the ingroup, and an increased clarity concerning norms regarding interaction with the outgroup, enhancing group identification, and perhaps increasing the likelihood that a group may be provoked or respond to inflammatory triggers in a group fashion rather than in an individual manner. It should also be noted though that as these conversations are taking place in a public forum, it is possible for outgroup members to 'observe' the opposing groups' ingroup conversations and this may affect behaviour. Together, these findings suggest a dynamic in which different aspects of between-group and within-group communication reflect a group's disposition to engage in outgroup-directed violence.

Nature of outgroup engagement in the conversation

The benefits of intergroup contact are premised on such contact being positive (Brown & Hewstone, 2005); in the current study, however, the outgroup engagement could not be characterized as such. We found that naturally occurring communication between members of opposing groups was more toxic than equivalent within-group conversations, suggesting that positive experiences would not be felt by either group involved in these exchanges. This corroborates previous findings that online discussions between ideologically opposed communities typically carry a negative sentiment (Williams et al., 2015). Additionally, across all event pages, the average length of a continuous chain of conversation between members of opposing groups was significantly lower than for within-group conversations, indicating shorter instances of intergroup than intragroup contact. While some interactions were longer, on the whole the interactions are far too short and fleeting for any level of personal or prolonged contact to occur. The short nature of the conversations suggests that those taking part are not motivated to maintain the conversation for long. This might reflect the negative nature of the conversation pushing participants away, or it could reflect a fact of the social network platform itself promoting short-term conversations in a constantly changing and updating digital environment. Short-term exposure does not prevent negative, stereotype reinforcing contact from

occurring (MacInnis & Page-Gould, 2015). Additionally, it requires more time to develop positive impressions online than offline (Jarvenpaa & Leidner, 1999; Walther, 1996) and so these short exchanges, even when positive, may be too fleeting to lead to positive group outcomes such as a reduction in prejudice and discrimination, and an improvement in relations. It should, however, be noted that these findings do not take into account that the same individual may take part in multiple exchanges. Because we anonymized the dataset and looked solely at the content of messages sent, more developed exchanges may be occurring over time in a number of short bursts of engagement.

The relative anonymity of users, due to a lack of individuating information beyond username and profile image being provided, may have been a further obstacle to positive effects of intergroup engagement (Islam & Hewstone, 1993; Lee, 2007); indeed, anonymity has previously been found to reduce the positive effects of computer-mediated intergroup contact under controlled conditions (Schumann et al., 2017). The abrasive and confrontational nature of the discussions (demonstrated by the high toxicity) may instead have increased the salience of group memberships (an element which could be tested in future research). This salience may lead those communicating online to generalize their predominantly negative experience more strongly to the outgroup as a whole, increasing intergroup anxiety, promoting negative stereotypes and damaging chances for future positive interactions (R. Brown & Hewstone, 2005). In the absence of face-to-face cues and prior personal knowledge about other members of the conversation, then whatever subtle social cues do appear in the online environment take on a much larger weight (Bacev-Giles & Haji, 2017; Postmes et al., 1998). This combination of highly toxic interactions that are short-lived and with low individuality, but highly salient group membership, is a likely explanation for why online outgroup engagement in confrontational situations is associated with negative group outcomes, in this case an increase in offline violence.

Criticism of social media dividing societies has often cited the potential for these platforms to create ideological 'echo chambers' (Bright, 2018; Conover et al., 2011) – networks of like-minded people who confirm each other's opinions instead of promoting critical thought. Furthermore, these criticisms assume that increasing digital connection and 'breaking down echo chambers' such that individuals interact with people from other social groups will naturally lead to positive outcomes (Berke, 2018). Our findings, however, are not only at odds with the notion of echo chambers – with 32.0% of all communication taking place between groups – but also challenge the assumption that breaking down

echo chambers will necessarily improve intergroup relations. Instead, our results align with findings from the offline domain in showing that such improvements may not occur unless at least some of the key conditions for positive intergroup contact are met (Allport, 1954). A growth of recent evidence also suggests, like our findings, that online echo chambers may not be occurring as commonly as expected (Dubois & Blank, 2018; Guess et al., 2018; O'Hara & Stevens, 2015); however, evidence is limited that online intergroup exposure, such as it is, is associated with improvements in group relations (Yardi & Boyd, 2010). Our results provide evidence that may help to explain this apparent paradox. In our sample, online engagement with the outgroup is occurring, but its limited quantity and predominantly negative quality is unlikely to promote positive group outcomes and reduce antagonism between the groups.

Adversarial nature of opposing groups

Given our focus on communication between protest groups and counter-protest groups, the starting point of the contact situation was likely to be adversarial by default, and outcomes of online intergroup contact may be different with more benign or neutral initial positions (Gehlback et al., 2018). Moreover, the online contact environment is prone to attract individuals with stronger outgroup prejudices (Hasler & Amichai-Hamburger, 2014), and given that the structure of online networks facilitates ingroup contact, engaging the opposing side in discussion may well require, or at least be typically associated with, a motivation to engage the outgroup in online intergroup conflict. This competitiveness, however, is a characteristic feature of intergroup relations (Wildschut et al. 2003), and, combined with the fact that online contact is primarily text based (as in the present study, via Facebook pages), which can itself increase the chance of conflict (Schroeder et al., 2017), conflictual online communication – between left-wing and right-wing political groups, or otherwise – is unlikely to be rare. The societal importance of the outcomes studied here (including, in some cases, bodily harm) highlights how important it is to study the association between online intergroup contact and its behavioural correlates in the real world.

It should be noted that some of the observed adversarial conversation may come from 'troll' accounts who act deliberately to inflame or provoke other members of the conversation. Indeed, this type of behaviour from inauthentic accounts run by state proxies has been shown to lead to a worsening of online conversations, including an increase in the level of toxicity and a reduction in the IC (Gallacher & Heerdink, 2019). In the current study we cannot differentiate between genuine social media users

and inauthentic accounts, and it is therefore possible that some toxic outgroup engagement was a result of this type of behaviour. However, this does not affect our results, because regardless of whether an outgroup provocation is issued from a real or troll account, the effect on the recipient in their perception of the outgroup remains the same.

Limitations and future directions

The main limitation of this study is that the evidence we have reported does not allow us to demonstrate that this relationship between online and offline behaviour is causal. There may be wider events which are driving both the increase in online hostility and subsequent offline violence in parallel. These variables may include the wider media environment and a specific focus on far-right related issues, political activity and key leader expressions, as well as highly relevant real-world events such as terror attacks – which have been shown previously to lead to spikes in far-right online activity and hate speech (Williams et al., 2014; Williams & Burnap, 2016; Williams et al., 2019). Regardless of this limitation, we believe that our findings are novel and useful and may be indicative of a wider trend where antagonistic online discussions are associated with offline actions later in time. In this sense, the online activity may be viewed as a measure of the 'temperature' or 'atmosphere' of the group relations at any given time, which only expresses itself physically when the groups subsequently meet in the real world. Future experimental research, which would be ethically demanding in this sphere, could demonstrate that the relationship we found between the online and offline worlds is a causal one. In the absence of such experimental data, data such as we have analysed (where online conversations link with, and temporally precede, the offline behaviour at the relevant event) can make useful contributions towards a better understanding of these issues.

One important and related question to answer in this regard is whether the same people were taking part in the online and offline conversations. For ethical reasons we did not store the names or profiles of those taking part in the online conversations, and we can therefore provide no insights on this. However, given that we studied event pages aiming to coordinate the offline rallies, it is likely that a significant proportion of the online users were also present offline. Besides, for our general argument it is just as important whether more extreme online exchanges are associated with more violence offline by the same *or other* participants. While having more direct evidence in this regard would help further our understanding of how online interaction can translate into offline activities at the individual level, this is not what our study aimed to achieve, and it is difficult to foresee how such

research could be done whilst respecting individual privacy. Instead, we aimed to study group-level effects, acknowledging that the individuals within groups varied in the extremity of their views, and that different individuals will partake at different times, and to test whether groups which had certain online conversation characteristics as a whole were more prone to be involved in more violent events in the real world.

Additionally, as is the case with field studies of intergroup contact, there is a self-selection bias with the participants taking part in the conversations. As discussed above, those who are willing to partake in contact with the opposing group may hold stronger outgroup prejudices (Hasler & Amichai-Hamburger, 2014). Forced intergroup contact has been shown to have larger effects than voluntary contact (Pettigrew & Tropp, 2006); in the current study it can be thought that the individual 'reaching out' to the opposing member is making a voluntary engagement, while the recipient is having this interaction forced upon them. Whether effects of either voluntary or forced contact exist within online environments remains to be seen. Equally, our results are restricted to political extremism in the United Kingdom and United States; future research should seek to replicate our main findings in other countries, as well as in domains other than right- and left-wing politics (e.g., hooliganism, separatist conflicts, or racial divides).

Future research could also focus on identifying cases where positive online intergroup contact does occur, and how to generalize these conditions to the wider ecosystem, as well what structural changes could be made to the current online environments in order to encourage more developed, sustained and positive contact between members of opposing groups. These structural changes could include reducing the level of anonymity such that positive interactions with outgroup members in non-political conversations (which are more diverse (Barberá et al., 2015) carry across to political conversations. Alternatively, algorithmic additions, which a user can control, which suppress hostility and promote civility, may help to counter evidence that moral outrage and aggression spread faster on social media than positive content (Brady et al., 2017; Crockett, 2017) and help to rebalance the perceived hostility of outgroup members (Duggan et al., 2016).

Conclusion

We provide evidence that for adversarial political protest groups online conversations are associated with subsequent offline group behaviour. We show that for highly charged issues spontaneous

engagement with members of opposing groups is fairly frequent, but that social media platforms are failing to facilitate positive outgroup engagement between these antagonistic groups. In fact, the style and nature of such online exchanges is more indicative of negative rather than positive intergroup contact. The superficial and hasty nature of group interactions is likely to reinforce pre-existing prejudices, generate negative affective states and even lead to a situation where groups co-radicalize and polarize their views through unsavoury contact with one other. Our results suggest that Facebook was misguided in the claim that it is decreasing conflict simply through enabling the connection of individuals from diverse backgrounds, at least in the case of political groups.

Exposure to uncivil disagreements online is associated with negative effects, including: withdrawal and isolation from online conversations (Bode, 2016), increased perception of social distance between groups (Iyengar et al., 2012), and increased affective polarization (Suhay et al., 2018). We take this work a step further and demonstrate that uncivil disagreements between groups on social media are associated with, and can statistically predict, violence when these groups meet in the real world. This is not to say that online intergroup contact and digital communication as a whole cannot lead to positive intergroup outcomes, but rather that the 'natural environment' that currently exists within social networking sites is not conducive to nurturing it, especially for those at the extremes of a social, and in this case political, spectrum.

Acknowledgements

This research was supported by grants from EPSRC and the University College Oxford Radcliffe Scholarship. The second author's contribution to this project was supported by a grant from the Netherlands Organisation for Scientific Research (NWO 446-16-015). The funding bodies played no further role in designing or implementing the research, and the authors declare no competing interests. We thank members of the Oxford Centre for the Study of Intergroup Conflict for their helpful feedback.

References

Allport, G. (1954). *The nature of prejudice*. Addison-Wesley Publishing Company. https://doi.org/10.1002/9780470773963

Amichai-Hamburger, Y. (2008). The contact hypothesis reconsidered: Interacting via internet: Theoretical and practical aspects. *Psychological Aspects of Cyberspace: Theory, Research, Applications*, 209–227. https://doi.org/10.1017/CBO9780511813740.010

Amichai-Hamburger, Y., & McKenna, K. Y. A. (2006). The contact hypothesis reconsidered: Interacting via the internet. *Journal of Computer-Mediated Communication*, *11*(3), 825–843. https://doi.org/10.1111/j.1083-6101.2006.00037.x

Austin, R. (2006). The role of ICT in bridge-building and social inclusion: Theory, policy and practice issues. *European Journal of Teacher Education*, *29*(2), 145–161. https://doi.org/10.1080/02619760600617284

Bacev-Giles, C., & Haji, R. (2017). Online first impressions: Person perception in social media profiles. *Computers in Human Behavior*, *75*, 50–57. https://doi.org/10.1016/j.chb.2017.04.056

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., … Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.17605/OSF.IO/4YGUX

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, *65*(2), 320–350. https://doi.org/10.1111/jcom.12145

BBC News. (2017). *Twitter suspends Britain First leaders*. BBC. http://www.bbc.co.uk/news/technology-42402570

Berger, J. M. (2017). *Extremist construction of identity: How escalating demands for legitimacy shape and define in-group and out-group dynamics. Terrorism and Counter-Terrorism Studies*. The International Centre for Counter-Terrorism – The Hague. https://doi.org/10.19165/2017.1.07

Berger, J. M. (2018a). *Extremism*. Cambridge, Massachusetts: The MIT Press.

Berger, J. M. (2018b). *The Alt-Right Twitter census: Defining and describing the audience for alt-right content on Twitter*. VOX-Pol Network of Excellence. https://www.voxpol.eu/new-research-report-the-alt-right-twitter-census-by-j-m-berger/

Berke, J. (2018). *Mark Zuckerberg says the world is much more divided than he ever expected* World Economic Forum. https://www.weforum.org/agenda/2018/02/mark-zuckerberg-says-he-thought-facebook-could-solve-a-lot-of-problems-but-the-world-is-more-divided-than-he-expected

Berrios, R., Totterdell, P., Kellett, S., & Brose, A. (2015). Eliciting mixed emotions : a meta-analysis comparing models , types , and measures. *Frontiers in Psychology*, *6*(April), 1–15. https://doi.org/10.3389/fpsyg.2015.00428

Bode, L. (2016). Pruning the news feed: Unfriending and unfollowing political content on social media. *Research & Politics*, *3*(3), 205316801666187. https://doi.org/10.1177/2053168016661873

Bose, S. (2017). *Package ' RSentiment .'* https://cran.r-project.org/web/packages/RSentiment/RSentiment.pdf

Bose, S., Saha, U., Kar, D., Goswami, S., Nayak, A. K., & Chakrabarti, S. (2017). Rsentiment: A tool to extract meaningful insights from textual reviews. In *Advances in Intelligent Systems and Computing* (Vol. 516, pp. 259–268). Springer, Singapore. https://doi.org/10.1007/978-981-10-3156-4_26

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Briggs, R., & Strugnell, A. (2011). *Radicalisation: The Role of the Internet. A working paper of the PNN.* Institute for Strategic Dialogue. https://www.isdglobal.org

Bright, J. (2018). Explaining the emergence of echo chambers on social media: the role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33. https://doi.org/10.2139/ssrn.2839728

Brown, K. T., Brown, T. N., Jackson, J. S., Sellers, R. M., & Warde, M. J. (2003). Teammates on and off the field? Contact with black teammates and the racial attitudes of white student athletes. *Journal of Applied Social Psychology*, *33*, 1379–1403. https://doi.org/10.1111/j.1559-1816.2003.tb01954.x

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., … Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 1–14. https://doi.org/10.1007/s13278-014-0206-4

Bundesministerium des Innern. (2015). *2015 Annual Report on the Protection of the Constitution. Facts and Trends.* https://www.verfassungsschutz.de/embed/annual-report-2015-summary.pdf

Chu, T., Jue, K., & Wang, M. (2017). *Comment abuse classification with deep learning.* Stanford University. https://web.stanford.edu/class/cs224n/reports/2762092.pdf

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *133*(26), 89–96. https://doi.org/10.1021/ja202932e

Conway, M., & Courtney, M. (2017). *Violent extremism and terrorism online in 2017: The year in review.* VOX-Pol Network of Excellence. https://www.voxpol.eu/vox-pol-year-review-published/

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1–3. https://doi.org/10.1038/s41562-017-0213-3

Croon, M. A., & Van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes: A meta-analytic review. *Personality and Social Psychology Review*, *15*(4), 332–351. https://doi.org/10.1177/1088868311411103

Dimock, M., Kiley, J., Keeter, S., & Doherty, C. (2014). *Political polarization in the American public.* https://doi.org/10.1017/CBO9781107415324.004

Dubois, E., & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information Communication and Society*, *4462*, 1–17. https://doi.org/10.1080/1369118X.2018.1428656

Duggan, M., Smith, A., & Page, D. (2016). *The Political Environment on Social Media.* Pew Research Centre. http://www.pewinternet.org/2016/10/25/the-political-environment-on-social-media/

Ebner, J. (2017). *The rage : the vicious circle of Islamist and far-right extremism.* London: I.B.Tauris.

Enikolopov, R., Makarin, A., & Petrova, M. (2016). Social media and protest participation: Evidence from Russia to understand whether social media indeed promotes protest participation. *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2696236

EUROPOL. (2017). *Europoean Union Terrorism Situation and Trend Report (EU TESAT) 2017.* https://doi.org/10.2813/237471

F.Sebastiani. (1999). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. http://faure.iei.pi.cnr.it/~fabrizio

Facebook. (2010). *Peace on Facebook.* https://www.facebook.com/helppeople/posts

Facebook Newsroom. (2018). *Taking Action Against Britain First.* from

https://newsroom.fb.com/news/h/taking-action-against-britain-first/

Fielitz, M., Ebner, J., Guhl, J., & Quent, M. (2018). *Loving hate. Anti-Muslim extremism, radical Islamism and the spiral of polarization*. https://www.isdglobal.org/isd-publications/hassliebe-muslimfeindlichkeit-islamismus-und-die-spirale-gesellschaftlicher-polarisierung-deutsch/

Gallacher, J. D. & Heerdink, M. W. (2019). Measuring the effect of Russian Internet research Agency information operations in online conversations. *Defence Strategic Commun*ications. 6, 155-198

Gehlback, H., Robinson, C, D., & Vriesema, C, C. (2018). Climate conversations: Seeking a common starting point. *PsyArXiv*. Https://doi.org/10.31234/osf.io/s8a7z

Gideon, L., Iii, C., Conway, K. R., & Houck, S. C. (2014). Automated Integrative Complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Google Project Jigsaw. (2018). *Perspective*. https://www.perspectiveapi.com/#/

Graf, S., Paolini, S., & Rubin, M. (2014). Negative intergroup contact is more influential, but positive intergroup contact is more common: Assessing contact prominence and contact prevalence in five Central European countries. *European Journal of Social Psychology*, *44*(6), 536–547. https://doi.org/10.1002/ejsp.2052

Gruzd, A., & Tsyganova, K. (2015). Information wars and online activism during the 2013/2014 crisis in Ukraine: Examining the social structures of Pro- and Anti-Maidan groups. *Policy and Internet*, *7*(2), 121–158. https://doi.org/10.1002/poi3.91

Guess, A., Barber, P., Vaccari, C., Kingdom, U., Nyhan, B., Seigel, A., … Stukal, D. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature*. William and Flora Hewlett Foundation. https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/

Guess, A., Lyons, B., Nyhan, B., & Reifler, J. (2018). *Avoiding the Echo Chamber about Echo Chambers: Why selective exposure to like-minded political news is less prevalent than you think*. Knight Foundation. https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf

Guttieri, K., Wallace, M. D., & Suedfeld, P. (1995). The Integrative Complexity of American decision makers in the Cuban missile crisis. *Journal of Conflict Resolution*, *39*(4), 595–621. https://doi.org/10.1177/0022002795039004001

Hasler, B., & Amichai-Hamburger, Y. (2014). Online intergroup contact. In book: *The Social Net: Understanding our online behavior*. Edition 2. Chapter 12. Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199639540.003.0012

Herek, G. M., & Glunt, E. K. (1993). Interpersonal contact and heterosexuals' attitudes toward gay men: Results from a national survey. *The Journal of Sex Research*, *30*(3), 239–244. https://doi.org/10.1080/00224499309551707

Hoter, E. (2009). Information and Communication Technology (ICT) in the service of multiculturalism. *International Review of Research in Open and Distance Learning*, *10*(2), 1–15.

Houck, S. C. (2014). Automated Integrative Complexity : Current challenges and future directions. *Political Psychology*, *35*(5), 647–659. https://doi.org/10.1111/pops.12209

Houck, S. C., Repke, M. A., & Conway, L. G. (2017). Understanding what makes terrorist groups' propaganda effective: an integrative complexity analysis of ISIL and Al Qaeda. *Journal of Policing, Intelligence and Counter Terrorism*, *12*(2), 105–118. https://doi.org/10.1080/18335330.2017.1351032

Howard, P. N., Duffy, A., Freelon, D., Hussain, M. M., Mari, W., & Maziad, M. (2011). Opening closed regimes: What was the role of social media during the Arab spring? Project on Information Technology & Political Islam. *SSRN*. https://doi.org/10.2139/ssrn.2595096

Ioannou, M., Hewstone, M., & Al Ramiah, A. (2017). Inducing similarities and differences in imagined contact: A mutual intergroup differentiation approach. *Group Processes and Intergroup Relations*, *20*(4), 427–446. https://doi.org/10.1177/1368430215612221

Islam, M. R., & Hewstone, M. (1993). Dimensions of contact as predictors of intergroup anxiety, perceived out-group variability, and out-group attitude: An Integrative Model. *Personality and Social Psychology Bulletin*, *19*(6), 700–710. https://doi.org/10.1177/0146167293196005

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly*, *76*(3), 405–431. https://doi.org/10.1093/poq/nfs038

Jackson, P. (2018). *The British extreme right: Reciprocal radicalisation and constructions of the other.*

Radicalisation Research. https://www.radicalisationresearch.org/debate/jackson-british-extreme-right-reciprocal-radicalisation/

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science*, *10*(6), 791–815. https://doi.org/10.1287/orsc.10.6.791

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1746–1751. https://doi.org/10.3115/v1/D14-1181

Knott, K., Lee, B., & Copeland, S. (2018). *Briefings: Reciprocal radicalisation*. Centre for Research and Evidence on Security Threats. https://crestresearch.ac.uk/resources/reciprocal-radicalisation/

Kumar, S., Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2018). Community interaction and conflict on the Web. *ArXiv*, 1–11. https://doi.org/10.1145/3178876.3186141

Lee, E. J. (2007). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, *57*(2), 385–403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Manbeck, K. E., Kanter, J. W., Kuczynski, A. M., Fine, L., Corey, M. D., & Maitland, D. W. M. (2018). Improving relations among conservatives and liberals on a college campus: A preliminary trial of a contextual-behavioral intervention. *Journal of Contextual Behavioral Science*, *10*(April), 120–125. https://doi.org/10.1016/j.jcbs.2018.10.006

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., … Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, (Icwsm), 369–380.

Meleady, R., Seger, C. R., & Vermue, M. (2017). Examining the role of positive and negative intergroup contact and anti-immigrant prejudice in Brexit. *British Journal of Social Psychology*, *56*(4), 799–808. https://doi.org/10.1111/bjso.12203

Mitts, T. (2019). From isolation to radicalization: Anti-muslim hostility and support for ISIS in the west. *American Political Science Review*, *113*(1), 173–194. https://doi.org/10.1017/S0003055418000618

Moghaddam, F. M. (2018). *Mutual radicalization: How groups and nations drive each other to extremes* (1st ed., Vol. 163). Washington: American Psychological Association. https://doi.org/10.1037/0000089-000

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, *2*(6), 389–396. https://doi.org/10.1038/s41562-018-0353-0

Mudde, C. (2007). *Populist radical right parties in Europe*. Cambridge: Cambridge University Press.

Mudde, Cas. (2019). *The far right today*. Cambridge: Polity.

Muhlhausen, D. B., & J B McNeill. (2011). *Terror trends: 40 years' data on international and domestic terrorism. Center for Data Analysis & Douglas and Sarah Allison Center for Foreign Policy Studies*. The Heritage Foundation. https://thf_media.s3.amazonaws.com/2011/pdf/sr0093.pdf

Müller, K., & Schwarz, C. (2020a). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *00*(0), 1–37. https://doi.org/10.1093/jeea/jvaa045

Müller, K., & Schwarz, C. (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *SSRN Electronic Journal*, 1–47. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

Neiwert, D., Ankrom, D., Kaplan, E., & Pham, S. (2017). *Homegrown Terrorism*. The Centre for Investigative Reporting. https://apps.revealnews.org/homegrown-terror/

O'Hara, K., & Stevens, D. (2015). Echo chambers and online radicalism: Assessing the Internet's complicity in violent extremism. *Policy and Internet*, *7*(4), 401–422. https://doi.org/10.1002/poi3.88

Omand, D., Bartlett, J., & Miller, C. (2012). Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, *27*(6), 801–823. https://doi.org/10.1080/02684527.2012.716965

Paolini, S., Harwood, J., & Rubin, M. (2010). Negative intergroup contact makes group memberships salient: Explaining why intergroup conflict endures. *Personality and Social Psychology Bulletin*, *36*(12), 1723–1738. https://doi.org/10.1177/0146167210388667

Park, G., & DeShon, R. P. (2018). Effects of group-discussion integrative complexity on intergroup relations in a social dilemma. *Organizational Behavior and Human Decision Processes*, *146*(March), 62–75. https://doi.org/10.1016/j.obhdp.2018.04.001

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*, 922–934. https://doi.org/10.1002/ejsp

Postmes, T., Spears, R., & Lea, M. (1998). Building or breaching social boundries? SIDE effects of computer mediated communication. *Communication Research*, *25*(6), 689–715.

Poushter, J., Wike, R., & Oates, R. (2015). *Extremism concerns growing in west and predominantly muslim countries*. Pew Research Centre. https://www.pewresearch.org/global/2015/07/16/extremism-concerns-growing-in-west-and-predominantly-muslim-countries/

Pratt, D. (2015). Islamophobia as Reactive Co-Radicalization. *Islam and Christian–Muslim Relations*, *26*(2), 205–218. https://doi.org/10.1080/09596410.2014.1000025

Ramiah, A. Al, & Hewstone, M. (2013). Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict: Evidence, limitations, and potential. *American Psychologist*, *68*(7), 527–542. https://doi.org/10.1037/a0032603

Ruesch, M. (2011). A peaceful Net? *First Global Conference on Communication and Conflict*, 1–19.

Schroeder, J., Kardas, M., & Epley, N. (2017). The humanizing voice: Speech reveals, and text conceals, a more thoughtful mind in the midst of disagreement. *Psychological Science*, *28*(12), 1745–1762. https://doi.org/10.1177/0956797617713798

Schumann, S., Klein, O., Douglas, K., & Hewstone, M. (2017). When is computer-mediated intergroup contact most promising? Examining the effect of out-group members' anonymity on prejudice. *Computers in Human Behavior*, *77*, 198–210. https://doi.org/10.1016/j.chb.2017.08.006

Schwab, A. K., & Greitemeyer, T. (2015). The world's biggest salad bowl: Facebook connecting cultures. *Journal of Applied Social Psychology*, *45*(4), 243–252. https://doi.org/10.1111/jasp.12291

Silge, J., & Robinson, D. (2017). *Text mining with R: a tidy approach*. Sebastopol: O'Reilly Media.

Sirseloudi, M. (2017). *Dyadic radicalisation via internet propaganda* [Conference session] Europol Conference on Online Terrorist Propaganda, The Hague, Netherlands.

Smith, A., Suedfeld, P., Conway, L., & Winter, D. (2008). The language of violence: distinguishing terrorist from nonterrorist groups by thematic content analysis. *Dynamics of Asymmetric Conflict*, *1*(2), 142–163. https://doi.org/10.1080/17467580802590449

Sønderskov, K. M., & Thomsen, J. P. F. (2015). Contextualizing intergroup contact: Do political party cues enhance contact effects? *Social Psychology Quarterly*, *78*(1), 49–76. https://doi.org/10.1177/0190272514560761

Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, *4*(1), 1–9. https://doi.org/10.1140/epjds/s13688-015-0056-y

Streufert, S., & Suedfeld, P. (1965). Conceptual structure, information search, and information utilization. *Journal of Personality and Social Psychology*, *2*(5), 736–740. http://www.ncbi.nlm.nih.gov/pubmed/5838772

Suedfeld, P. (2010). The cognitive processing of politics and politicians: Archival studies of conceptual and integrative complexity. *Journal of Personality*, *78*(6), 1669–1702. https://doi.org/10.1111/j.1467-6494.2010.00666.x

Suedfeld, P., & Bluck, S. (1988). Changes in Integrative Complexity prior to surprise attacks. *Journal of Conflict Resolution*, *32*(4), 626–635. https://doi.org/10.1177/0022002788032004002

Suhay, E., Bello-Pardo, E., & Maurer, B. (2018). The polarizing effects of online partisan criticism: Evidence from two experiments. *International Journal of Press/Politics*, *23*(1), 95–115. https://doi.org/10.1177/1940161217740697

Sunstein, C. R. (2017). *#Republic: divided democracy in the age of social media*. Princeton, New Jersey, United States: Princeton University Press.

Tetlock, P. E., Peterson, R. S., & Berry, J. M. (1993). Flattering and unflattering personality portraits of integratively simple and complex managers. *Journal of Personality and Social Psychology*, *64*(3), 500–511. https://doi.org/10.1037/0022-3514.64.3.500

Tufekci, Z. (2017). *Twitter and Tear Gas : the power and fragility of networked protest*. New Haven: Yale University Press.

Von Behr, I., Reding, A., Edwards, C., & Gribbon, L. (2013). *Radicalisation in the digital era: The use of the internet in 15 cases of terrorism and extremism*. RAND.

https://www.rand.org/pubs/research_reports/RR453.html

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal and hyperpersonal interaction. *Communication Research*, *23*(1), 3:43.

White, F. A., & Abu-Rayya, H. M. (2012). A dual identity-electronic contact (DIEC) experiment promoting short- and long-term intergroup harmony. *Journal of Experimental Social Psychology*, *48*(3), 597–608. https://doi.org/10.1016/j.jesp.2012.01.007

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual-intergroup discontinuity effect. *Psychological Bulletin*, *129*(5), 698–722. https://doi.org/10.1037/0033-2909.129.5.698

Williams, H. T. P., McMurray, J. R., Kurz, T., & Hugo-Lambert, F. (2015). Network analysis reveals open forums and echo chambers in social media discussions of climate change. *Global Environmental Change*, *32*, 126–138. https://doi.org/10.1016/j.gloenvcha.2015.03.006

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*. https://doi.org/10.1093/bjc/azz049

Wołk, K., & Marasek, K. (2015). Neural-based machine translation for medical text domain. Based on European Medicines Agency leaflet texts. *Procedia Computer Science*, *64*, 2–9. https://doi.org/10.1016/j.procs.2015.08.456

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. *International World Wide Web Conference*, 1–9. https://doi.org/10.1145/3038912.3052591

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/027046761038001

Supplementary Information (SI) for Chapter 1:

Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters

1 - Sampling, data collection and processing, and events selected for analysis

A two-stage process was conducted to identify cases of events that could be used for analysis. For the United Kingdom sample, the two largest and most active street-protest groups from each side of the political spectrum were identified; these were English Defence League (EDL) and Britain First (BF) from the far-right, and Unite Against Fascism (UAF) and Anti-Fascist Network (AFN) from the left. For each of these groups an initial sample of events where the group participated was selected using a historical open-source search of past events hosted on the groups' Facebook page. Criteria for inclusion were that the events must have a corresponding counter-march or protest by the opposing political side, organized in tandem, on the same day and at the same location. This method resulted in the selection of 18 events. Following this, an additional search was conducted more broadly across reports of political violence in the United Kingdom to identify events organized by alternative groups (often localized events or events held under alternative banners). This resulted in an additional two events being identified. Once identified the same criteria for inclusion were used. Importantly. we have not sought to define groups ourselves into political ideology, but rather relied on existing

formalizations. At the point of collection, this dataset represented the entirety of the publicly available data of this type from the United Kingdom.

For the United States sample five events that took place in the summer of 2017 were selected. All of these cases occurred in response to the 'Unite the Right' rally in Charlottesville (the data from the Charlottesville protest itself is not publicly available as Facebook removed the pages prior to the event. We contacted Facebook to see if they would be willing to share this data, but they did not respond.) These cases were selected as they bore a strong resemblance to the United Kingdom based protests, far-right marches taking place with an opposing left-wing counter protest.

Once the political events were identified, we collected the conversations taking place online using a Python (version 2.7) script connected to the Facebook Graph Application Program Interface (API). The process consisted of two stages. Once a Facebook event page had been identified then the unique identifier for this page was gathered. This identifier was then used to collect the content of each 'status' posted to the page along with the unique identifier for this status. This generated a first dataset. The second stage used this dataset as an input, and collected all the comments posted below each status. This produced a second dataset. Finally, the two datasets were combined to produce the full set. There was a single case of a comment being replicated 95 times in quick succession in the dataset. This was presumed to be due to a glitch in either posting to the social network or in data collection and so only a single version of this comment was included in subsequent analysis.

All comments and statuses were given equal weighting. In some cases, multiple event pages were identified for each side (left or right wing) from multiple organizing groups within the same protest. In these cases, the event page comments were collected separately and the data then compiled into a single set for either the left or the right groups. The final list of events included in the analysis is shown in Table S1. This data collection method gathered an average of 1473 comments per event page, giving a total of 2946 comments on average per event and 73,632 comments in total. Once collected, the data were cleaned to remove any blank entries (posts containing only images or video and no text) and to remove any conversation that occurred after the planned start time of the event. In doing this we can safely assume that any violence at the later event had no impact on earlier conversation online.

Events selected for analysis

Table S1 shows the events which were selected for inclusion in the current study, along with the groups associated with the right-wing and left-wing sides. Table S2 shows the types of data source used to generate violence metrics for each event.

Table S1. Real world protests, marches and rallies selected for analysis

Date	Location	Country	Right Wing Group(s)	Left Wing Group(s)
05/09/2015	Rotherham	UK	Britain First	Rotherham Unite Against Fascism
17/10/2015	Burton	UK	Britain First	East Staffordshire Trades Council
30/01/2016	Dewsbury	UK	Britain First	We Are Dewsbury
16/07/2016	London	UK	English Defense League	Unite Against Fascism (UAF)
06/08/2016	Nottingham	UK	English Defense League	Midlands Anti-Fascist Network
24/09/2016	Newcastle	UK	English Defense League	Newcastle Unites
05/11/2016	Telford	UK	English Defense League	ShropRad
25/02/2017	Rotherham	UK	English Defense League	Rotherham Unite Against Fascism
25/02/2017	Telford	UK	Britain First	Shropshire and Telford Trades Council & UAF
01/04/2017	London	UK	English Defense League & Britain First	Anti-Fascist Network (AFN) & Unite Against Fascism
08/04/2017	Birmingham	UK	English Defense League	Birmingham Unite Against Fascism
15/04/2017	Wishaw	UK	Scottish Defense League	Unite Against Fascism Scotland
03/06/2017	Liverpool	UK	English Defense League	Merseyside Unite Against Fascism
11/06/2017	Manchester	UK	UK Against Hate	Stand-Up to Racism
24/06/2017	Birmingham	UK	Britain First	Birmingham Unite Against Fascism
22/07/2017	Rochdale	UK	Britain First	Unite Against Fascism
29/07/2017	Rochdale	UK	English Defense League	Unite Against Fascism
13/08/2017	Seattle	U.S.	Patriot Prayer	Greater Seattle IWW General Defense Committee Local 24
19/08/2017	Boston	U.S.	Boston Free Speech Coalition	Black Lives Matter Network, Black Lives Matter Cambridge, and Black Lives Matter Boston
26/08/2017	San Francisco	U.S.	Patriot Prayer	Peace Love & Understanding (among others)
27/08/2017	Berkeley	U.S.	No Marxism in America	SAFEBay - Solidarity Against Fascism East Bay
02/09/2017	Keighley	UK	English Defense League	Unite Against Fascism
10/09/2017	Portland	U.S.	Patriot Prayer	Rose City Antifa, Queer Liberation Front
21/10/2017	Peterborough	UK	English Defense League	Peterborough Trades Union Council - PTUC
04/11/2017	Bromley	UK	Britain First	Stand Up To Racism - South East London and Unite Against Fascism

Table S2. Types of data source used for violence measurement at each real-world event.

		Data Sources used to generate offline violence metric					
Date	**Location**	**Police Reports**	**Mainstream National News**	**Local News**	**Citizen Journalism (Blogs)**	**Photo Journalism (e.g. PhotoBucket)**	**Crowdsourced Video (e.g. YouTube)**
05/09/2015	Rotherham, UK	✓	✓	✓		✓	✓
17/10/2015	Burton, UK		✓	✓	✓		✓
30/01/2016	Dewsbury, UK		✓	✓			✓
16/07/2016	London, UK		✓	✓		✓	✓
06/08/2016	Nottingham, UK	✓	✓				✓
24/09/2016	Newcastle, UK	✓	✓	✓			✓
05/11/2016	Telford, UK	✓	✓	✓		✓	✓
25/02/2017	Rotherham, UK	✓	✓	✓			
25/02/2017	Telford, UK	✓	✓	✓			✓
01/04/2017	London, UK	✓	✓			✓	✓
08/04/2017	Birmingham, UK	✓	✓	✓		✓	✓
15/04/2017	Wishaw, UK		✓	✓			✓
03/06/2017	Liverpool, UK	✓	✓	✓			✓
11/06/2017	Manchester, UK	✓	✓		✓		✓
24/06/2017	Birmingham, UK		✓	✓			✓
22/07/2017	Rochdale, UK			✓			✓
29/07/2017	Rochdale, UK			✓		✓	✓
13/08/2017	Seattle, U.S.	✓	✓	✓	✓	✓	✓
19/08/2017	Boston, U.S.	✓	✓	✓		✓	✓
26/08/2017	San Francisco, U.S.		✓	✓			
27/08/2017	Berkeley, U.S.	✓	✓	✓		✓	✓
02/09/2017	Keighley, UK		✓	✓			✓
10/09/2017	Portland, U.S.	✓	✓	✓			✓
21/10/2017	Peterborough, UK		✓	✓		✓	✓
04/11/2017	Bromley, UK			✓	✓	✓	✓

2 - Violence measurements and item response theory

Identifying violent and non-violent events

Our initial measure demonstrates whether or not meaningful violence was sparked, but not the level or degree of this violence. The primary marker for this was whether the event was described afterward in reports as violent. For example: 'Violent clashes occurred at event x' would be classified as violent, while 'a peaceful protest took place', would be classified as peaceful.

Degree of violence

The second measure takes into account the same evidence but uses a range of indicators (Table S3) to classify more specifically how violent an event was depending on how many of these indicators were present. The following indicators were used: heavy police presence (indicating an expectation of violence), physical contact between groups, throwing of projectiles, bloodshed/actual bodily harm, arrest figures (1 indicator for minor arrests, up to 5, and an additional indicator for major arrests of 5+), and violence leading to a fatality. These indicators were selected following advice from analysts at an industry leading security firm familiar with the subject matter. This organization was not involved with analysis subsequently to providing this guidance.

We do not distinguish whether each indicator of violence occurred specifically between members of opposing groups, in contact with the police, or even within a group itself. While delving into more detail on the type and nature of violence at events may be interesting, there are limitations that prevent this, and the information is not always available. In many cases reports after the event did not specify between whom the violence took place and attempts to trace violence observed in photos or videos back to a Facebook profile to determine group membership go far beyond the scope of this study. Instead as violence between protesters and law enforcement, and violence between groups, is likely to be correlated (with one type of violence encouraging the other) we no do distinguish violence based on the recipient of the action, purely based on the presence of the indicator. The number of occurrences of each type of violence are shown in table S3 and the distribution of indicators across events in shown in table S4.

These indicators were weighted using a latent trait model using the 'ltm' package in R (Rizopoulos, 2017), which creates a continuous scale of violence on the z scale ($M = 0$, $SD = 1$). The model is based on the relationship between performance on single item measures compared to overall performance. In the current case, this compares individual indicators of violence to the overall violence score measured with all indicators, without assuming that each item represents an equal level of violence. This results in an estimate of the 'severity' ('difficulty', in latent trait parlance) of each violence indicators on the latent violence scale, with the indicators indicative of more severe violence placed higher on the violence scale.

From the results, we can extract the severity parameter for each indicator, and calculate a per-event violence score on a continuous scale. Fig S1 shows the item characteristic curves for the indicators of violence with coefficients shown in Table S3. The occurrence column gives the total number of times this indicator was reached across all events, with a maximum possible score of 25 if an indicator was present at all events. Clearly, the LTM orders the violence indicators in a sensible way, with heavy police presence being triggered at the lowest level of violence, and bloodshed/bodily harm indicating the most violence. Moreover, the items covered a wide range of latent violence scores (zs from -1.28 to 0.74) and all indicators discriminate very well (discrimination parameters >= 1.77, (Baker, Boston, & Rudner, 2001)), suggesting they yield reliable information about violence. We calculated the latent event-level violence scores using these indicators, which were normally distributed according to a visual inspection of the density distribution. This violence measure was then used as the dependent variable in the linear model with the conversation quality metrics outlined in the main text.

Table S3. Violence Indicators, Violence Coefficients

Reference	Indicator of Violence	Violence *z*-score	Discrimination	Occurrence (max = 25)
g	Heavy Police Presence	-1.28	11.37	22
d	Minor arrests (1-5 Arrests)	-1.05	2.47	20
c	Projectiles / Smoke Bombs	-0.11	2.49	13
a	Physical contact between opposing groups	0.22	3.43	10
e	Major arrests (5+)	0.35	3.09	9
b	Bloodshed / Bodily Harm	0.74	3.42	6
f	Violence leading to a Fatality	n/a	n/a	0

(Note: Indicator 'f', Violence leading to a fatality, did not occur at any sampled events, and so did not receive an associated violence score)

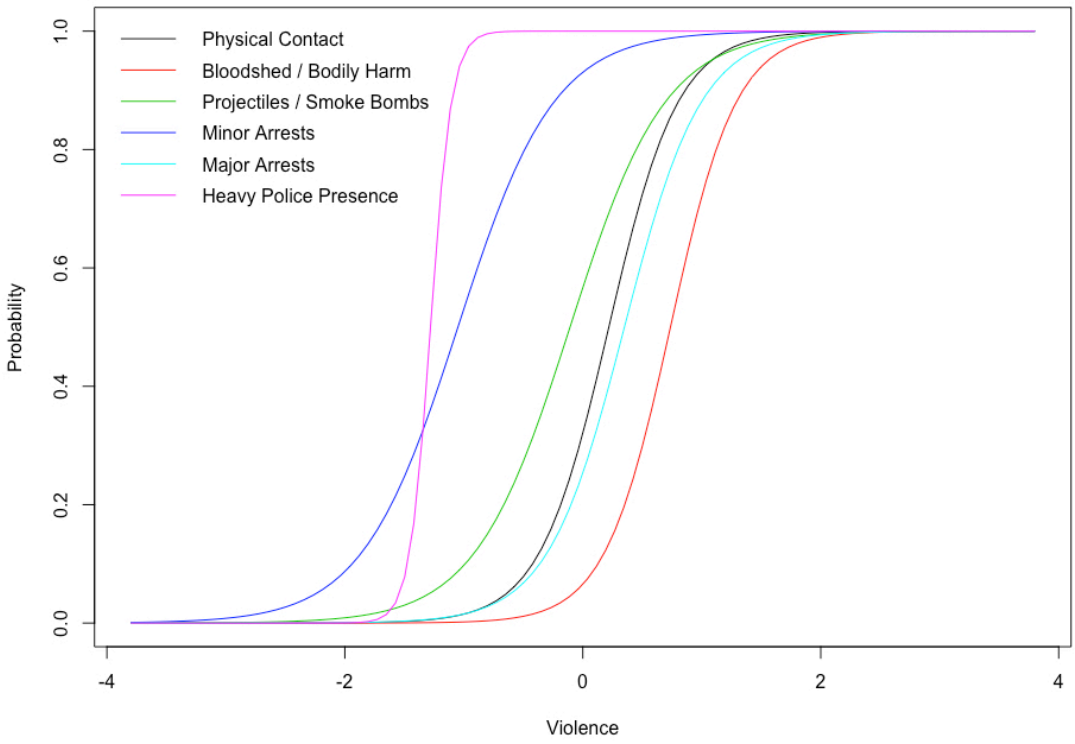


Fig S1. Item Characteristic Curves for the six occurring indicators of violence.
Higher inflection points indicate greater violence is needed for the indicators to occur

Table S4. Indicators of violence observed at each real-world event

			Indicator of Violence						
Date	**Location**	**Country**	**a**	**b**	**c**	**d**	**e**	**f**	**g**
05/09/2015	Rotherham	UK	✓	✓	✓	✓	✓		✓
17/10/2015	Burton	UK				✓	✓		✓
30/01/2016	Dewsbury	UK				✓			✓
16/07/2016	London	UK							
06/08/2016	Nottingham	UK			✓	✓			✓
24/09/2016	Newcastle	UK				✓			✓
05/11/2016	Telford	UK				✓			
25/02/2017	Rotherham	UK				✓	✓		✓
25/02/2017	Telford	UK		✓	✓	✓			✓
01/04/2017	London	UK	✓	✓	✓	✓	✓		✓
08/04/2017	Birmingham	UK	✓			✓			✓
15/04/2017	Wishaw	UK			✓				✓
03/06/2017	Liverpool	UK	✓		✓	✓	✓		✓
11/06/2017	Manchester	UK	✓	✓	✓	✓	✓		✓
24/06/2017	Birmingham	UK			✓	✓			✓
22/07/2017	Rochdale	UK							✓
29/07/2017	Rochdale	UK							✓
13/08/2017	Seattle	U.S.	✓		✓	✓			✓
19/08/2017	Boston	U.S.	✓		✓	✓	✓		✓
26/08/2017	San Francisco	U.S.				✓			✓
27/08/2017	Berkeley	U.S.	✓	✓	✓	✓	✓		✓
02/09/2017	Keighley	UK			✓	✓			✓
10/09/2017	Portland	U.S.	✓	✓	✓	✓	✓		✓
21/10/2017	Peterborough	UK	✓			✓			✓
04/11/2017	Bromley	UK							

3 – Correlations between text analysis measures

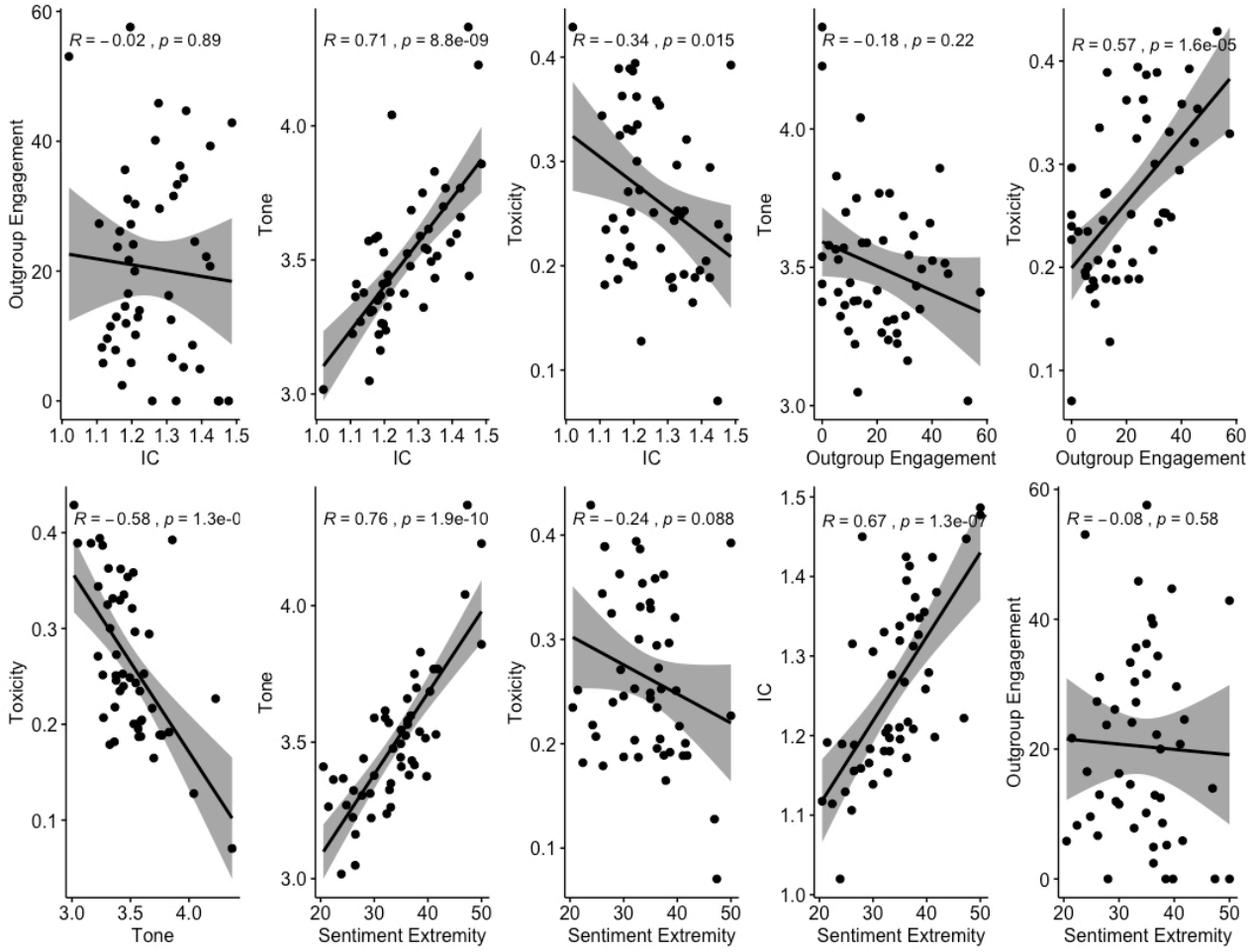


Fig S2. Correlation plots between text analysis measures used in the study.
Significant correlations were found between a number of the measures and therefore multicollinearity in models was checked using variance inflation factor (VIF) . Shaded areas represent 95% confidence interval.

Table S5. Correlation matrix among text analysis measures used in the study

Measure	1	2	3	4	5	6
1 – Sentiment – Tone						
2 – Sentiment – Extremity	0.76***					
3 – Toxicity	-0.58***	-0.24				
4 – Integrative Complexity	0.71***	0.67***	-0.34*			
5 – Outgroup Engagement	-0.18	-0.08	0.57***	-0.02		
6 – Degree of Violence	0.22	0.33	0.15	0.12	0.52**	
Minimum	3.02	20.51	0.07	1.02	0.0	
Maximum	4.37	50.00	0.43	1.49	57.59	
Mean	3.50	34.03	0.26	1.26	20.44	
SE	0.04	0.99	0.01	0.02	2.12	

4 - Best linear unbiased predictor aggregation

In order to account for the fact that the predictors (outgroup engagement, tone, sentiment extremity, integrative complexity and toxicity) were measured at the conversation level but violence was measured at the event level, we aggregated scores for all group variables to the event level. To do this we calculated adjusted means for each variable using the best linear unbiased predictor (BLUP) as in prior research (Croon & Van Veldhoven, 2007). This is required due to the effect that when predicting outcomes at the higher level from predictors nested within a lower level it is statistically biased to directly regress the higher level outcome variable on the unadjusted means of lower level predictors. The correction is shown to yield unbiased estimates of the parameters (Becker, Breustedt, & Zuber, 2017).

This analysis was done using the 'MicroMacroMultilevel' package in R (Becker et al., 2017; Lu, Pagegould, & Xu, 2017). The BLUP aggregations for the event level variables were then used as predictors of the presence/absence of violence within a GLM with a binomial distribution and a logit link function, and as predictors of the degree of violence with a linear model.

5 - Robustness checks

We checked the robustness of models by performing multicollinearity checks though the calculation of variance inflation factors (VIF), checking for the absence of influential data points, and inspecting the linearity, normality of residuals, and homoscedasticity through a visual inspection of residual plots (histogram, Q-Q and rotated fitted values vs residuals).

Sensitivity analysis

To test the robustness of the degree of violence measure, we performed a sensitivity analysis. In this test each indicator of violence (a-f) was removed from analysis in turn, the degree of violence for each event re-calculated using IRT, and the main effect of outgroup engagement on degree of violence tested. In all cases the effect was still found to be significant. These results are shown in Table S6. We conclude from this that the measure is robust and not reliant on any single indicator of violence.

Table S6. Sensitivity Analysis of degree of violence

Indicator Removed	t Value	*p* Value
None	3.20	0.0031 **
Physical Contact	2.99	0.0066 **
Bloodshed / Bodily Harm	3.69	0.0012 **
Projectiles / Smoke Bombs	3.29	0.0032 **
1-5 Arrests	2.99	0.0066 **
5+ Arrests	3.23	0.0037 **
Heavy Police Presence	3.15	0.0045 **

6 - Neural network text classifier for outgroup engagement

In order to classify comments into ingroup directed and outgroup directed, we trained a neural network machine learning classifier using a supervised learning method. The training set for this network consisted of 1,000 randomly selected comments from the entire dataset of available comments. Each comment was classified at either 'within-group' for comments that were directed towards other ingroup members, or 'between-group' for comments that were directed towards a member of the outgroup. Therefore a 'between-group' comment could either be a member injecting a comment into the event page of the opposing group, or a reply to this injection from a member of the incumbent page.

Group membership was identified from self-disclosed information in the content of the comments that a user posted to the Facebook page linked to an explicitly right-wing or left-wing political group. From the content of these comments, it is possible to infer either support for the event or opposition. These comments were manually coded by the first author, who is familiar with the right-wing/ left-wing online environment. Each comment was coded in isolation; however, common themes and patterns were identified and applied throughout the dataset. The default coding option was within-group communication, and so if a decision could not be made then this was the option selected. Additionally, for comments that did not contain enough information (such as photos which would be represented textually as [[PHOTO]]), the default coding option was selected.

This set was evenly balanced such that 500 comments were identified as ingroup and 500 comments were identified as outgroup. The entire dataset contains a higher proportion of ingroup compared to outgroup comments, and so in order to gather this balanced training set, over 1000 comments were

classified until at least 500 ingroup and 500 outgroup comments were identified, and then a random selection of these taken.

The test dataset contained a second random sample of 1000 comments. This set was not manually balanced between ingroup and outgroup communication and so reflects the true balance with the entire sample. These comments were manually coded in the same way.

To ensure accuracy of the human coding, all comments from the training and test set were coded by a second coder who was blinded to the hypotheses, and inter-coder reliability (ICR) scores calculated. For the training set the ICR was 97.8% with a Scott's PI of 0.956. For the test set the ICR was 95.9%% with a Scott's PI of 0.895. These values were deemed high enough to be confident in the original classifications.

Ethical considerations prevent further profiling of individual users beyond the self-disclosed content of each message posted to the event page and all conversation data was anonymized at the point of collection. We made no attempts to link comments to individual Facebook profiles. Additionally, ethical constraints restrict accessing any information that is not made available via the public Facebook API. This information contains only the content of the post, the author, the date/time it was posted and any reactions to it. This means that even if conversations were not anonymized, we would not be able to gather any additional information about the participants beyond this. It is therefore possible, due to this anonymization, that some ingroup critique was coded as outgroup critique as the judgement was made solely on the basis of what was said in the comment. However, in most cases this coding was straightforward for the human coders to complete, and certain words and phrases used only by outgroup members clearly highlighted outgroup engagement (calling someone offensive or derogatory terms for example is unlikely to come from an ingroup member). The high level of agreement between both human coders (97.8% for the training set, 95.9% for the test set) reflects this.

The neural network itself was created using the natural language toolkit (NTLK) library with the Python coding environment and using a 'bag of words' approach (Pinto, 2017). For the entire training set each word was tokenized, stemmed, converted to lower-case, and then added to a dictionary or 'bag of words' with all non-standard characters removed. The model then learns over time the

associated word patterns that occurred during ingroup and outgroup communications. The neural network contained two layers of neurons, with one hidden layer of 20 neurons and was run over 100,000 iterations and an alpha reinforcement learning rate of 0.1.

When classifying comments the network gives a classification along with a confidence judgement for each entry. This confidence value is used to create a threshold below which comments would not be included in further analysis. A confidence threshold of 95% was selected as the best trade-off between minimizing false positives whist avoiding too many unclassified comments (Fig S3). At this threshold, 14.7% of comments were excluded due to low confidence.

Applying this confidence threshold led to an overall accuracy for the classifier of 89.0%, with a sensitivity of 85.9% and a specificity of 89.9%. (Therefore this is a conservative judgement classifier with regard to outgroup classification, reflecting the conservative nature of setting ingroup contact as the default classification setting). The baseline accuracy figures without a confidence threshold applied are 84%, with a sensitivity of 84% and a specificity of 88%. By way of comparison the second human coder in the inter-coder reliability test achieved 97% accuracy, with a specificity and sensitivity both of 97% when compared to the initial coder across both the training and test set.

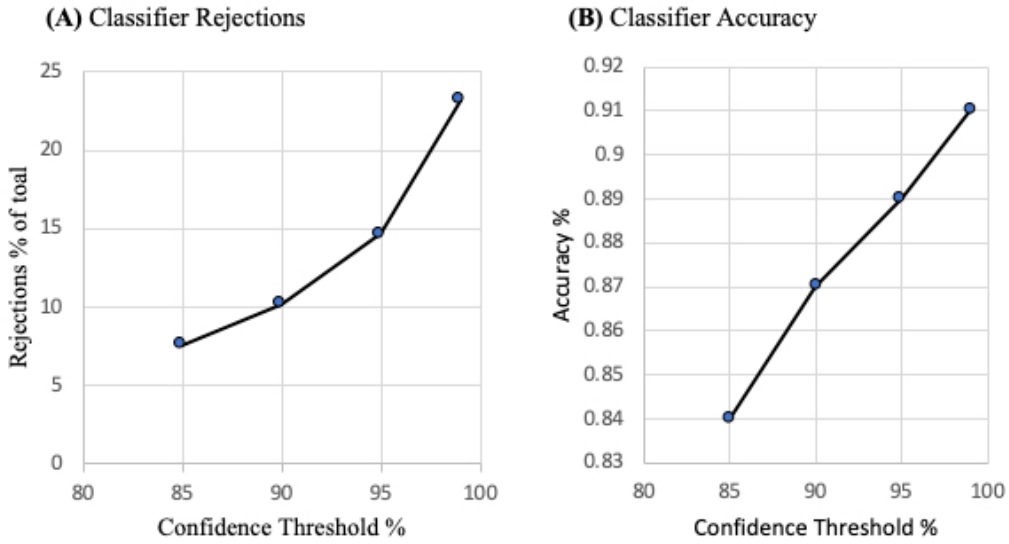


Fig S3 *– (A) The proportion of comments within the test set which were rejected at each confidence threshold. (B) The overall classifier accuracy at each confidence threshold*

It is important to note that the network classifies each comment as if it were occurring in isolation within the dataset – this is then adjusted using two subsequent rules to increase accuracy of the network. Firstly, a proximity rule is applied whereby if a comment is both preceded and followed by a case of outgroup engagement, then its classification is changed to engagement. This was designed to make the classification system more aware of the overall conversation. Secondly, the classification was manually corrected by a human coder (the same human coder from the original classifications). The aim of this was not to reclassify all cases, but rather to check for obvious errors, long strings of replicated comments or cases whereby the initial proximity rule was applied incorrectly. These two rules accounted for only small changes in the contact figures, with only 2.7% of the comments of each event page adjusted on average at stage 1 and 3.0% at stage 2. Running the main analysis (outgroup engagement on degree of subsequent violence) without these two corrections did not alter the conclusions.

7 - Further exploratory analysis

Group comparisons, left vs. right

In order to investigate differences in conversations occurring between left-wing and right- wing group pages we performed comparisons using linear mixed models (LMMs) with the lme4 package. We investigated association of sentiment (tone and sentiment extremity), outgroup engagement, toxicity, and integrative complexity across the two groups, and included event ID as a random effect. Significance levels of fixed effects (and interactions) were obtained by comparing the full model to the null model with a χ^2 test.

Results showed that right-wing pages (n=25) had higher toxicity scores (indicating a more toxic conversation) than left-wing pages (Right-wing: M=0.30±0.02, Left-wing: 0.23±0.01, LMM, χ_1^2 = 12.93, p <0.001) and a lower average value of sentiment extremity (Right-wing: M=30.27±1.13, Left-wing: 37.80±1.25, LMM, χ_1^2 = 17.32, p <0.001). Conversations within right-wing pages displayed lower level of Integrative Complexity (Right-wing: M=1.20±0.02, Left-wing: M=1.32±0.002, LMM, χ_1^2 = 19.19, p <0.001) and lower tone scores (Right-wing: M=3.36±0.03, Left-wing: M=3.64±0.05, LMM, χ_1^2 = 16.94, p <0.001). Finally, right-wing pages showed a larger degree of outgroup engagement than

left-wing pages (Right-wing: M=25.74%±2.61, Left-wing: M=15.13%±3.02, LMM, $\chi_1^2 = 7.39$, $p = 0.007$). These results are shown in Fig S4 and Table S7.

This echoes previous evidence showing that left-leaning politicians tend to display higher levels of IC than right-leaning politicians (Tetlock, 1983). Additionally, integrative complexity in online communication can provide information about the extent to which individuals hold radical or extremist views (Smith, Suedfeld, Conway, & Winter, 2008), therefore suggesting that it may the content on the right-wing pages in the current study which displays more extreme views.

The result that outgroup engagement is more likely to occur in the conversations on the event page hosted by a right-wing group suggests that it is individuals from the left-wing groups who are more likely to 'seek out' this contact by visiting the opposing page and engaging in discussion.

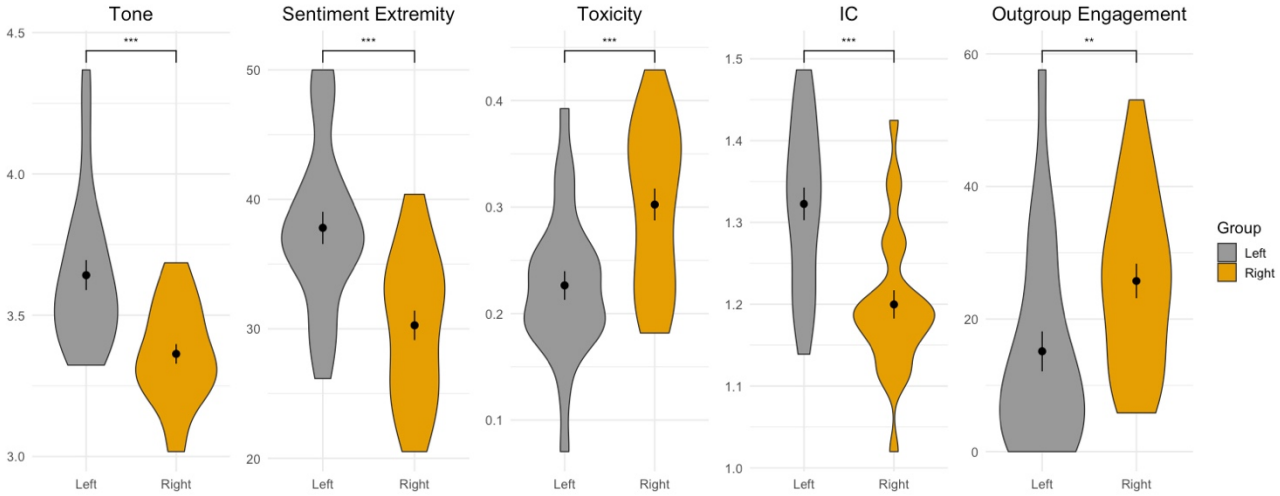


Fig S4. Comparison of right-wing vs left-wing conversation qualities.
*This figure demonstrates the differences in Tone, Sentiment Extremity, Toxicity, Integrative Complexity and outgroup engagement occurring on right-wing and left-wing Facebook event pages. Mean and standard error is given by dots and lines, respectively. *** denotes $p < 0.001$, ** $p < 0.01$.*

Table S7. Differences in conversation metrics between right-wing and left-wing event pages (Mean ± SE)

Group	Tone	Sentiment Extremity	Toxicity	IC	Outgroup Engagement
Left	3.64 ± 0.05	37.8 ± 1.25	0.23 ± 0.01	1.32 ± 0.02	15.1 ± 3.02
Right	3.36 ± 0.04	30.3 ± 1.13	0.30 ± 0.02	1.20 ± 0.02	25.7 ± 2.61

SI References

Baker, F., Boston, C., & Rudner, L. (2001). *The Basics of Item Response Theory* (2nd ed.). ERIC Clearning House on Assesment and Evaluation.

Becker, D., Breustedt, W., & Zuber, C. I. (2017). Surpassing simple aggregation : Advanced strategies for analyzing contextual-level outcomes in multilevel models. *Methods, Data, Analyses*, 1–31. https://doi.org/10.12758/mda.2017.05

Croon, M. A., & Van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Lu, J. G., Page-gould, E., & Xu, N. R. (2017). Package ' MicroMacroMultilevel .' Retrieved from https://cran.r-project.org/web/packages/MicroMacroMultilevel/MicroMacroMultilevel.pdf

Pinto, A. (2017). Text Classification using Neural Networks , Machine Learnings. Retrieved April 12, 2018, from https://machinelearnings.co/text-classification-using-neural-networks-f5cd7b8765c6

Rizopoulos, D. (2017). Package ' ltm .' Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Smith, A., Suedfeld, P., Conway, L., & Winter, D. (2008). The language of violence: distinguishing terrorist from nonterrorist groups by thematic content analysis. *Dynamics of Asymmetric Conflict*, *1*(2), 142–163. https://doi.org/10.1080/17467580802590449

Tetlock, P. E. (1983). Cognitive style and political ideology. *Journal of Personality and Social Psychology.* https://doi.org/10.1037/0022-3514.45.1.118

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Online engagement between opposing political protest groups via social media is linked to physical violence of offline encounters
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2020). Online contact between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society*, 1–44.

Student Confirmation

Student Name:	John Gallacher		
Contribution to the Paper	John Gallacher (J.G) was the primary author on this paper, receiving guidance and supervision from Marc Willem Heerdink (M.W.H) and Miles Hewstone (M.H) while working in the Oxford Centre for the Study of Intergroup Conflict (OxCSIC). J.G conceived the study, collected the data, and developed the inter-group comment classifier. J.G developed the statistical models and analysed the data, with guidance from M.W.H. J.G wrote the manuscript, with M.W.H and M.H providing valuable feedback and comments on early versions and within the peer-review process.		
Signature 		Date	11/01/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Marc W. Heerdink, Ph.D.		
Supervisor comments I hereby certify that the description of my role in this paper is accurate. John Gallacher led all stages of the project and produced the paper largely independently. The paper is therefore primarily his work.		
Signature 	Date	14/01/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 2

Leveraging cross-platform data to improve automated hate speech detection

Abstract

Hate speech is increasingly prevalent online, and its negative outcomes include increased prejudice, extremism, and even offline hate crime. Automatic detection of online hate speech can help us to better understand these impacts. However, while the field has recently progressed through advances in natural language processing, challenges still remain. In particular, most existing approaches for hate speech detection focus on a single social media platform in isolation. This limits both the use of these models and their validity, as the nature of language varies from platform to platform. Here we propose a new cross-platform approach to detect hate speech which leverages multiple datasets and classification models from different platforms and trains a 'superlearner' that can combine existing and novel training data to improve detection and increase model applicability. We demonstrate how this approach outperforms existing models, and achieves good performance when tested on messages from novel social media platforms not included in the original training data.

Introduction

While the Internet has allowed people from across the globe to connect and communicate, it has also given those who wish to spread abuse, incivility, hate, and other forms of extreme digital speech the unprecedented ability to do so, and at greater scale than ever. Up to 1% of content on mainstream social media platforms now contains some form of extreme digital speech, and as many as 40% of online users report having seen some form of extreme online speech (Vidgen, Margetts, & Harris, 2020), suggesting that it receives disproportionate attention. This reach is increasing, with online hate crimes rising by 40% since 2017 (Home Office, 2018; Williams & Mishcon de Reya, 2019).

Given this increasing prevalence, there is a growing requirement to understand both the nature and impact of online hate speech. Analysis of online abuse and hate speech requires robust methods to detect and measure it, and manual qualitative analysis of individual messages cannot scale to the billions of posts made online on social media every day (Internet Live Stats, 2020), limiting its application. In addition, exposing human moderators to the harmful nature of extreme digital speech can be damaging to mental health over the long term (Gillespie, 2018; King, 2018; Roberts, 2019). Because of this, automated techniques to detect extreme digital speech are essential for both research and practical applications.

Automatically detecting extreme digital speech in online communications faces a number of challenges, however (Vidgen, Tromble, et al., 2019). These include how to accurately detect harmful content across diverse social media platforms, how to ensure models remain up to date when the nature online conversations is rapidly changing over time (Florio, Basile, Polignano, Basile, & Patti, 2020; Laaksonen, Haapoja, Kinnunen, Nelimarkka, & Pöyhtäri, 2020), and how to differentiate between the most egregious extreme digital speech and more moderate, but still offensive, language.

In this work we present a novel approach to address these challenges. Using data from across four different social media platforms, combined with an ordinal approach to rank extreme digital speech subtypes, and context-aware semantic embeddings, we build a model for hate speech detection which is more adaptable and comprehensive when compared to training models on data from a single social media platform in isolation. We assess the validity of our approach in three ways. Firstly, we compare performance of our cross-platform approach to models trained on data from a single social media platform to measure improvements in performance gained from leveraging cross-platform information. Secondly, we compare these performance measures to those from the existing literature for identical or similar datasets. Finally, we compare performance on a completely new and unseen dataset from a different social media platform, Reddit, and compare the performance of our proposed approach to the existing 'state-of-the-art' model trained on data from a single social media platform.

Defining online hate speech

Hate speech is a form of dangerous online communication or 'extreme digital speech' (Ganesh & Bright, 2020), which may seek to dehumanise its victims according to their group identity, and amplify the group identity of the perpetrator by attempting to create an antagonistic relationship between ingroup and outgroup members (Pohjonen, 2018). This 'othering' serves to reinforce existing group boundaries or create new ones. Given this potential for hate speech to drive intergroup conflict, many attempts have been made to formally define it, however as yet there is no single agreed upon definition. Various legal definitions of hate speech highlight the requirement for hate speech to cause harm (Delgado, 1982), and that this harm may include the incitement of further hostile actions beyond the speech itself (Marwick & Miller, 2014), whilst serving no redeeming purpose (Ward, 1998). These traits are not definitive however, and a wide array of different forms of speech could fit within a definition of "hate speech," depending on the context (Sellars, 2016). A related challenge is that hate speech is not the only type of extreme digital speech, instead it sits in the middle of a spectrum

of online abuse that ranges from microaggressions, condescension, and insults at the lower end, up to the promotion of self-harm, incitement of violence, and physical threats at the higher end (Jurgens, Chandrasekharan, & Hemphill, 2020). How to efficiently differentiate between elements at the lower and higher end remains an ongoing effort.

One of the few areas of consensus separating hate speech from other forms of harmful speech, is that hate speech targets groups or individuals as they relate to a group, compared to person-directed abuse which is focused solely on personal characteristics (Sellars, 2016). In this way, broad categorisations of hate speech typically present it to be bias-motivated, hostile, and malicious language targeted at a person or group because of their actual or perceived innate characteristics. Offensive language which expresses incivility to an individual, but not a group, would sit below hate and systemic intolerance towards a targeted group on this spectrum of online abuse.

In this work we follow this group-level approach, and use a definition of hate speech as "*messages which express hatred towards a targeted group with the intention to be derogatory, to humiliate, or to insult members of that group*" (Davidson, Warmsley, Macy, & Weber, 2017). Specifically, we make of use the definition proposed by de Gibert, Perez, García-Pablos, & Cuadros, (2019), and expand it slightly to incorporate implicit and indirect attacks. This definition requires three features for a message to constitute hate speech: the message must be (1) a deliberate attack, (2) directed towards, or about, a specific group of people, and (3) motivated by, or focused on, aspects of the group's identity. Importantly, this definition is broad enough to contain any targeted group and we do not differentiate between protected and wider group level characteristics.

Impacts of hate speech

Online hate speech is also often linked with offline hate crime. Violence towards immigrants within western countries is related to the degree of extreme digital speech and anti-refugee sentiment expressed on social media in areas where the violence takes place (Müller & Schwarz, 2020), while in the US, anti-Muslim messages disseminated by President Trump over social media correlate with the number of anti-Muslim hate crimes in states where social media usage is high (Müller & Schwarz, 2020b). In addition, there is evidence for a temporal and spatial association between online race and religion motivated hate speech, and offline racially and religiously aggravated crimes (Williams, Burnap, Javed, Liu, & Ozalp, 2019). Finally, antagonistic discussions between opposing groups

members has been shown to be predictive of offline violence between these same groups later in time (Gallacher, Heerdink, & Hewstone, 2020). The impact of online hate speech is also increasingly felt globally. The UN has highlighted that hate speech on Facebook played a leading role in inciting genocide of the Rohingya community in Myanmar (Stecklow, 2018), and in Sri Lanka anti-Islam hate speech has been linked with deadly mob violence (Samaratunge & Hattotuwa, 2014). Increases in the use of hate speech have also been observed after terror attacks and in particular increases in posts that advocate violence amongst Islamic extremist communities online (Olteanu, Castillo, Boy, & Varshney, 2018).

Online hate speech also causes psychological impacts even when offline violence does not occur. Following receiving online abuse people can feel afraid to leave their homes after (Awan & Zempi, 2016) along with reports of greater feelings of fear, anger, sadness, depression, as well as an increase in animosity and prejudice against the attacker's own group (UK Safer Internet Centre, 2016; Williams & Mishcon de Reya, 2019). In addition, users who display dehumanising language and extreme digital speech also express their desire for psychological and physical separation from the out-group, which increases both ingroup identity and also affective polarisation with the out-group (Harel, Jameson, & Maoz, 2020). Understanding the nature of online hate speech is therefore important in developing any mitigation measures or policy responses.

Current state of the art for automated hate speech detection

Early approaches to automatically detect online hate speech used lexicons and bags-of-words approaches to identify terms associated with hate speech (Warner & Hirschberg, 2012). Performance for these approaches is poor however, and they suffer from missing the context a message is posted in, and thus often miss the subtle or indirect ways in which language is used and lead to large number of false negatives, or even false positives when offensive terms are re-purposed or reclaimed for benign and even positive purposes (Gitari, Zuping, Damien, & Long, 2015; Greevy & Smeaton, 2004). Recently, performance of hate speech detection models has improved due to advances in natural language processing, general-purpose language models, machine learning, and statistical modelling (Vidgen, Tromble, et al., 2019). Key innovations include the use of deep learning and ensemble model architectures, using contextual word embeddings from pre-trained natural language models, and the inclusion of user-level variables (Badjatiya, Gupta, Gupta, & Varma, 2017; Zhang & Luo, 2018). Recent approaches using these techniques have shown promise (Fortuna & Nunes, 2018; Liu, Alorainy, Burnap, & Williams, 2019; Liu, Burnap, Alorainy, & Williams, 2020, 2019; Pitsilis,

Ramampiaro, & Langseth, 2018; Rizoiu, Wang, Ferraro, & Suominen, 2019). Similarly, the inclusion of parts-of-speech tagging, especially typed dependencies and detection of linguistic 'othering' is shown to improve performance of hate speech detection (Alorainy, Burnap, Liu, & Williams, 2018; Burnap & Williams, 2015, 2016). In addition to classification at the message level, extreme users who spread hate speech can also be automatically detected using machine learning approaches (Fernandez, Asif, & Alani, 2018; Ribeiro, Calais, Santos, Almeida, & Meira, 2018).

Limitations of these existing approaches and proposed solutions

Despite these recent improvements, limitations remain and a robust, widely applicable automated method to detect online abusive content remains to be devised (Vidgen, Tromble, et al., 2019). A key issue stems from data sparsity and lack of variability in training data (Schmidt & Wiegand, 2017), which means that classifiers typically struggle to perform well on unexpected inputs or in novel contexts. Three specific key challenges relate and follow from this; (i) how to build classifiers which work across different social media platforms, (ii) how to distinguish between hate speech and other less severe forms of extreme digital speech such as offensive or uncivil language, and (iii) how to keep classifiers up to date in a constantly changing online environment. Here we tackle these challenges with a novel approach leveraging cross-platform datasets, pre-trained natural language models, and an ordinal approach to distinguish hate speech from uncivil/offensive language.

i) Cross-platform classification

The first significant challenge is how to apply hate speech detection systems in a cross-platform manner. Hate speech exists on multiple social media platforms to varying degrees, but it is presented differently on each platform both in terms of structure and content. For example, the presentation of hate speech on a platform such as Twitter, where messages are short and fairly self-contained, will be very different to that on a static webpage, forum, or news website where the comments are posted in response to an article. A successful approach should be able to detect hate speech on multiple platforms without bias.

Most existing methods focus on a single platform, typically Twitter, due to its accessibility (Vidgen & Derczynski, 2020) and have relatively small training sets. These models therefore do not scale well to other platforms (Schmidt & Wiegand, 2017; Cihon & Yasseri, 2016), and risk biasing datasets towards certain types of abuse and framing (Gröndahl, Pajola, Juuti, Conti, & Asokan, 2018). Combining

datasets from multiple platforms is promising (Vidgen & Derczynski, 2020), but cross-platform approaches are scarce and typically join data together in a single cohesive dataset (Corazza, Menini, Cabrio, Tonelli, & Villata, 2019; Karan & Šnajder, 2018; Mishra, Yannakoudakis, & Shutova, 2019; Salminen et al., 2020; Sigurbergsson & Derczynski, 2019), which could lead to interference between datasets as platform-specific nuances may be lost.

We address this limitation by using a 'superlearner' approach. Superlearning is a technique for prediction that involves combining many individual machine learning models into a single prediction (van der Laan, Polley, & Hubbard, 2007). In this way, individual platform-specific supervised models are first trained on datasets from four social media platforms representing a wide range of mainstream and fringe platforms commonly used by online hate groups (Twitter, Facebook, Gab and Stormfront). Then, a superlearner classifier is trained on the predictions of all four models to learn the optimal weighted average of predictions in different situations. This allows us to improve performance by leveraging cross-platform datasets without destructive interference or the need to build large models from all platform data combined. Such ensemble approaches outperform a single algorithm in isolation (Kennedy, 2017; Polley & van der Laan, 2010; van der Laan, Polley, & Hubbard, 2007), and ensemble approaches have previously shown promise in the field of hate speech detection (Liu et al., 2020). When making a prediction on new data, we combine predictions from platform-specific classifiers as features fed into the superlearner. This approach therefore retains platform-specific features and idiosyncrasies.

ii) Spectrum of extreme digital speech

An ongoing challenge in automated detection methods is how to efficiently differentiate between offensive language and hate speech. This is particularly difficult to solve as much of the language in both cases uses similar words but with different meanings (Davidson et al., 2017; Rossini, 2019). Attempts to address this distinction have used multi-class classifiers, with three distinct categorical classes for 'clean', 'offensive', and 'hate speech' (Davidson et al., 2017). This approach improves performance by reducing the conflation in the model between hate speech and offensive language but fails to consider these classes as a continuous spectrum of abuse and instead treats them as distinct unordered categories.

Here, we distinguish hate speech from less severe abuse with a three-class ordinal approach, using the ordinal nature of these classes to improve classifier performance. Our approach aims to detect hate

speech by using the offensive category as a barrier to improve distinction between 'clean' and 'hate' classes.

Furthermore, distinctions between profanity and hate speech are often unclear from surface-level n-grams and heavily rely on context (Zampieri et al., 2019). As such, we use contextual word embeddings—which consider the meaning of words depending on their context—to distinguish between hate speech and offensive language. These can help detect coded language or euphemistic hate terms which traditional approaches would miss (Magu & Luo, 2018). We use embeddings from Google's BERT model, which have previously been successful in detecting hate speech (Kennedy et al., 2020; Salminen et al., 2020).

iii) Model updating

The final challenge is how to keep hate speech detection models up to date as both the nature of the language used online, and the social media platforms themselves, change over time. The specific terms used in hate speech and use of euphemistic language has been to shown change rapidly online (Florio et al., 2020; Laaksonen et al., 2020) and models will need to be updated regularly to account for this. Without regular updating machine learning model performance degrades steadily (Zliobaite, Pechenizkiy, & Gama, 2016). Equally, in practice, when new labelled datasets for hate speech detection become available in the research community it should be possible to update existing models to utilise this new data without the requirement to retain the entire model, which would be both computationally expensive and risk destructive interference with the previous models.

We address this limitation with a two-step approach which allows for the inclusion future datasets into the model in a way which does not involve retraining the entire model. Instead, we propose to train a smaller model on the specific new dataset and then update the superlearner stage to include this new model alongside the existing ones. We discuss how this approach could help address the challenge of model degradation and updating.

Methods

Here we propose an approach to improve hate speech detection in English language social media posts by leveraging cross-platform datasets of training data from Facebook, Gab, Stormfront and Twitter. We combine state-of-the-art natural language representations with syntactic features and an ordinal approach to classify hate and offensive speech to improve classifier performance.

Hate speech and offensive language definitions

As outlined in the introduction, we define hate speech as messages which contain (1) a deliberate attack, (2) directed towards, or about, a specific group of people, and (3) motivated by, or focused on, aspects of the group's identity. Importantly, this definition is broad enough to contain any targeted group, unlike some legal definitions and the definitions used by social media platforms which focus only on explicitly protected groups and specific characteristics (Facebook, 2020). This places our definition more in line with 2020 UK Law Commission recommendation, which gives an 'other' category for group membership in addition to explicitly named groups with protected characteristics (Law Commission, 2020).

Offensive posts are broadly defined as messages likely or with the intention to cause offence (Davidson et al., 2017), this includes uncivil, rude, inappropriate, or overly disrespectful content. It often contains obscenity and profanity towards a recipient in the 2nd or 3rd person but without invoking aspects of the target groups' identity as motivation for the attack (e.g. "You are such a f*****g idiot" / "he is such a f*****g idiot"). It may also include derogatory language (e.g. sexist, homophobic or anti-disability language) to cause offence, but which does not attack the group because of one of its characteristics (Bartlett, Reffin, Rumball, & Williamson, 2014; Orlando & Saab, 2020; Technau, 2018). Although some studies take a less restrictive approach to hate speech classification (Badjatiya et al., 2017; Liu, Alorainy, et al., 2019), this distinction between hate and offensive language that we follow here is common in the literature (Vidgen, Tromble, et al., 2019; Waseem, Davidson, Warmsley, & Weber, 2017; Zampieri et al., 2019). As an example, attacking an individual by using a slur but with no other referral to aspects of the individual's group identity would be classified as offensive (e.g. "haha, you're such a b*tch Brittany", or "stop being a r*tard James"). While these posts may contain offensive terms which do relate to specific groups, they are not used to directly derogate these groups in these instances, and so would be classified as offensive rather than hateful. This is demonstrated in online gaming communities for example, which have been shown to contain high levels of homophobic

language and common use of homophobic slurs, typically directed towards ingroup peers (Elveljung, 2018). Given that these messages are not specifically directed towards someone in the LQBTQ+ community, it does not match the definition above for hate speech, even though the inclusion of this language in 'normal' discourse is dangerous and risks perpetuating prejudice and discrimination. Similarly, when traditionally derogatory terms are reclaimed by a minority group, and used in a positive frame, these would be considered neither offensive nor hateful, e.g. 'I'm the realest b*tch he gone ever come across' (Röttger et al., 2020). Comparatively if a message derogated a group as a whole, even if this group is not typically marginalised (e.g. 'those dirty journalists are all untrustworthy tw*ts'), or called for hostile actions against this group (e.g. 'all journalists should be rounded up and shot, they deserve it') then this would be classified as hate speech.

Importantly, we do not distinguish 'sarcastic' from 'genuine' hate or abuse and classify this content as if it were serious. The high degree of sarcasm, irony, and in-jokes present in these platforms (Zannettou, Caulfield, et al., 2018) has been shown to provide a source of classification error as this distinction is increasingly blurred (Vidgen et al, 2019), especially as this content usually relies on the tacit acceptance of the prejudice and negative tropes even when displayed in a humorous way (Ma, 2014).

Training datasets

We use data from four key social media platforms popular among the far-right: Facebook, Twitter, Gab and Stormfront. Facebook and Twitter are popular with users from across the political spectrum but contain conversations from extremist groups who spread large volumes of hate speech (Burke, 2017; Vidgen, Yasseri, & Margetts, 2019). Gab and white supremacist forum Stormfront are more fringe but popular with extremist organisations and contain considerable hate speech (Kleinberg, Vegt, & Gill, 2020; Mathew, Dutt, Goyal, & Mukherjee, 2019). We selected these platforms because they vary substantially in their structure, the length of messages that can be posted, and specific groups which use them, therefore allowing us to gather a breadth of content and build a nuanced classifier. In total we collated a dataset of ~40,000 messages to build our classifier, split approximately equally across the platforms (Figure 1, Supplementary Information (SI) Table 1).

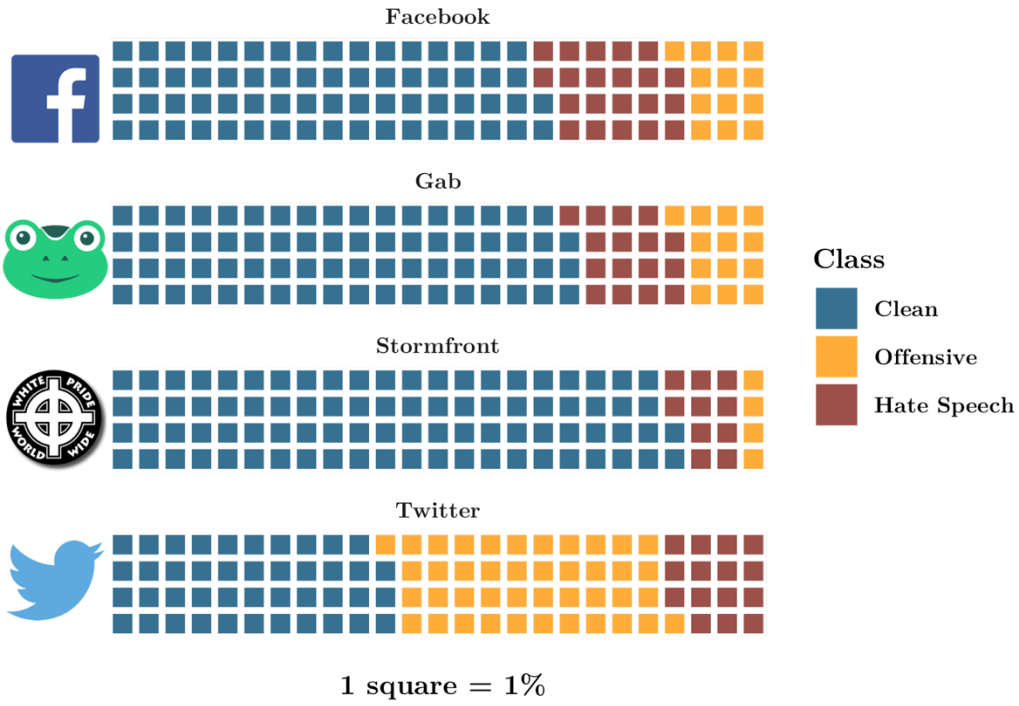


Figure 1 - Proportion of Hate Speech, Offensive Language and Clean Messages in the training data from across the four social media platforms

For Facebook we take a random sample of 10,000 messages from a dataset previously analysed in the context of intergroup conflict (Gallacher et al., 2020). This contains messages posted to 20 event pages from 2015 – 2017 created by two far-right groups in the UK: Britain First and the English Defence League. Both groups have subsequently been banned from the platform for expressing hate speech and spreading extremism.

For Gab we take a random sample of 10,000 messages from across the entire platform posted between its formation on 10th August 2016 and the 29th October 2018, the date at which it was taken temporarily offline. Our data consists of an amalgamation of data shared by Zannettou, Bradlyn, & Cristofaro, (2018) (August 2016 – January 2018) and data from the online repository Pushshift (January 2018 – October 29th 2018).

For Twitter we use a dataset published by Davidson et al. (2017) containing a sample of 24,802 tweets from users who expressed words from a lexicon of hate terms and already coded into "hate speech", "offensive" and "clean" classes (details below). We subsample 10,000 messages from this dataset by randomly down sampling the majority classes.

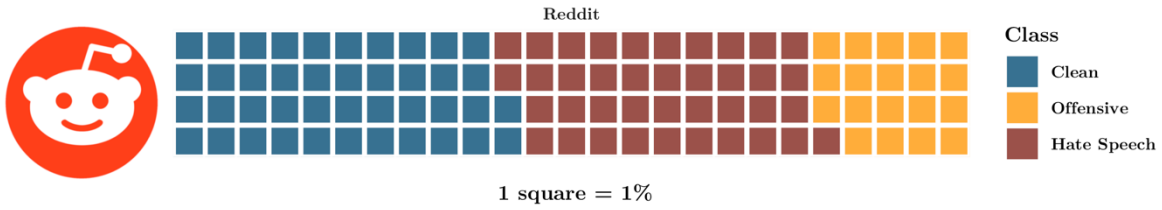


Figure 2 - Proportion of Hate Speech, Offensive Language and Clean Messages for the Reddit validation set

Finally, we use a dataset containing 10,568 messages from the white supremacist forum Stormfront made publicly available by de de Gibert et al., (2018). This dataset contains a random sample of English language posts from any of the 22-sub forums of the platform between 2002 and 2017 and is already classified into "hate" and "non-hate" messages.

The training data from the four platforms contained different proportions of messages from each class, as shown in Figure 1. On all platforms, clean messages formed the dominant class, however while Facebook and Gab datasets contained a fairly even balance of hate speech and offensive messages as minority classes, this was not the case on Twitter and Stormfront, with Stromfront having few offensive messages compared to hate speech, and Twitter having a lot more. This reflects both the sampling strategies used in the collection of this data, with Twitter data sampled using keyword searches which boosts the presence of offensive language for example, and also the underlying platform dynamics, with Stormfront containing less offensive messages relative to hate speech. This does not mean however that Stormfront should be considered a 'cleaner' platform, as this doesn't account for the severity or impact of the hate speech when it occurs on the platform.

Novel platform validation dataset

To test the performance of our modelling approach on a novel social media platform not included in the original training, we use a random sample of 1,000 annotated posts from a different social media platform, Reddit, shared by Qian, Bethke, Liu, Belding, & Wang, (2020). These messages were pre-labelled for presence or absence of hate speech, but we re-annotate them to fit our three categories.

We deliberately chose 1,000 posts as they represents a 'new' dataset becoming available, but of insufficient size to be used efficiently on its own, as it is well below the average for training a completely new hate speech detection model (Vidgen & Derczynski, 2020). However, if combining a model built from this 'small' dataset with the other existing platform-specific models then this would

support the validity our approach. The balance of data across the three classes for the Reddit data is shown in Figure 2.

Data labelling / annotation approach

With the exception of the Twitter dataset, already labelled into 'hate speech', 'offensive' and 'clean' categories by Davidson et al. 2017, all datasets required full (Facebook and Gab) or partial (Stormfront, Reddit) manual labelling. To ensure consistency across datasets we used the same labelling strategy as was used for the Twitter dataset, which is based on the definition of hate speech described above. We first labelled the Facebook and Gab datasets into 'hate' and 'non-hate' categories. Then, we manually split the 'non-hate' speech category into 'offensive' and 'clean'. For Stormfront and Reddit, to allow comparability in performance for hate speech detection with the existing literature, we did not re-classify any messages from within the 'hate speech' class of this dataset but created an 'offensive' class by reclassifying offensive messages previously labelled as 'non-hate'.

All data labelling was conducted by a single native English-speaking coder, familiar with the far-right online ecosystem and language used. To validate the accuracy of this labelling, a sample of 1,000 messages from Facebook and Gab platforms was selected and labelled independently by a second coder who is also familiar with the online environment but has a different nationality, gender and background, who had been trained in data labelling strategy and hate speech definition prior to annotation. Intercoder reliability scores gave a percentage agreement of 89.7%, an ordinal Krippendorff's Alpha of 0.863, and a Cohen's Kappa score of 0.797. The former is at the level of 'good' agreement (George & Mallery, 2003) while the latter is well above the 0.61 threshold for reliable coding and substantial agreement (Glen, 2014), and so we retain our labelling as accurate.

Data pre-processing

After manual labelling, data were pre-processed prior to building the machine learning models. We removed any messages that had fewer than 2 words or fewer than 5 characters in total, as these were too short to extract meaningful features, or which could not be classified by the human coders into a category. These messages were uncommon and included examples such as '*it is'* or '*they are'* which cannot be meaningfully classified, and presumably rely heavily on context not included in the message itself. In this step we also removed messages in languages other than English and any messages which

contained no substantial text (e.g. only URLs or images). We aimed to build a classifier looking at messages in isolation, so we filtered out the 1,001 comments from the Stormfront dataset which had been coded by de Gibert et al. as requiring additional context for human annotators to code – i.e. they could only be labelled by looking back within the comment thread to identify the context of the conversation. Messages which could be classified in isolation, even if embedded in a thread, were still included. This therefore matched with the approach taken on Facebook and Gab labelling, where messages which could not be coded in isolation were removed. The final samples sizes for each platform is given in the Supplementary Information (SI) Table 1.

Following this, all messages were put into lowercase and any non-standard (UTF-8 / ASCII) characters removed. We used regular expressions to extract the title of linked pages or news articles from messages containing URLs in addition to other text (see SI 1.2). The remainder of the URL was discarded. All punctuation, including hashtags, was counted and then removed, although the keyword following a hashtag was retained and treated like any other word within the message.

Features used in machine learning classification

We extracted commonly used and novel features from each message to train the model to classify them into 'clean', 'offensive' or 'hate speech' categories. Broadly, these features can be broken down into semantic features that encode the substance of the message, and syntactic features that encode how the message is presented. In total we extracted 7 types of semantic feature and 6 syntactic features. These are listed in SI Table 2.

Semantic features

For the majority of the semantic features we use word embeddings—a learnt representation of text based on words' meaning but also the context of the entire message—to extract meaning from the text. Specifically, we make use of the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin, Chang, Lee, & Toutanova, 2018) which has been trained on large quantities of unannotated online text, and utilises bidirectional search (forward and backward from every given word) to better understand the use of a word in the context of the whole message. We extracted BERT document vectors (distributed representations of the entire text across multidimensional space) for each message using the 'rBERT' package (Harmon & Bratt, 2020) and took the final layer output vector for the "CLS" token. This creates a representation of each message in 768 features. Extracting

BERT word vectors and feeding these forwards in downstream models performs at near parity with fully retraining the model (Alammar, 2018).

In addition, we encode plural nouns, 'othering' and explicit hate terms; theory-driven semantic features likely to be present with hate speech expressions. Plural nouns may be especially informative for hate speech classification as they may indicate when a message is directed towards an outgroup as a whole – a necessary element of hate speech in our definition. This alone may not indicate hate speech but may be a good indicator when combined with other features such as the content of the message or the presence of offensive language. We used a part of speech (POS) tagging system to extract all plural nouns and identify their relevance for each label category by calculating the weighted-log-odds (details in SI 1.4). Derogatory and pejorative plural nouns, as well as standard terms for groups often receiving abuse, received high scores within the hate speech category; in contrast top terms in the 'clean' category are not offensive (SI Figure 1).

We measure the presence of 'othering' by identifying the use of two-sided pronouns that contain a distinction between the in-group and out-group in a single message (e.g. your/our, them/us, they/we), using dictionaries of popular personal pronouns for both the in-group and out-group. This can improve hate speech detection, especially when the group being targeted is not explicitly named (Alorainy et al., 2018).

Additionally, we measured the presence and severity of explicit hate words/symbols using a crowdsourced dictionary of 883 'hate terms' rated by severity on a 1-100 scale (Quinn, Tuckwood, & Boyd, 2019). For each message we calculated the number and cumulative hate severity score of these hate terms. Furthermore, as extremist groups are increasingly using hate symbols to express coded hate we included a measure of the number of hate symbols in a message, using a dictionary of hate symbols (Anti-Defamation League, 2019), (e.g. '(((echo)))' is used by alt-right online communities to target Jewish individuals).

Additional features included overt obscenity, counted as the number of swear words per message using an online dictionary used to block profanity in online message boards (Fontgate Media, 2014), and the sentiment / polarity of the message on a -1 to +1 scale using the polarity function from the 'qdap' package (Rinker, 2020), as both hate speech and offensive language are likely to have a more negative valence.

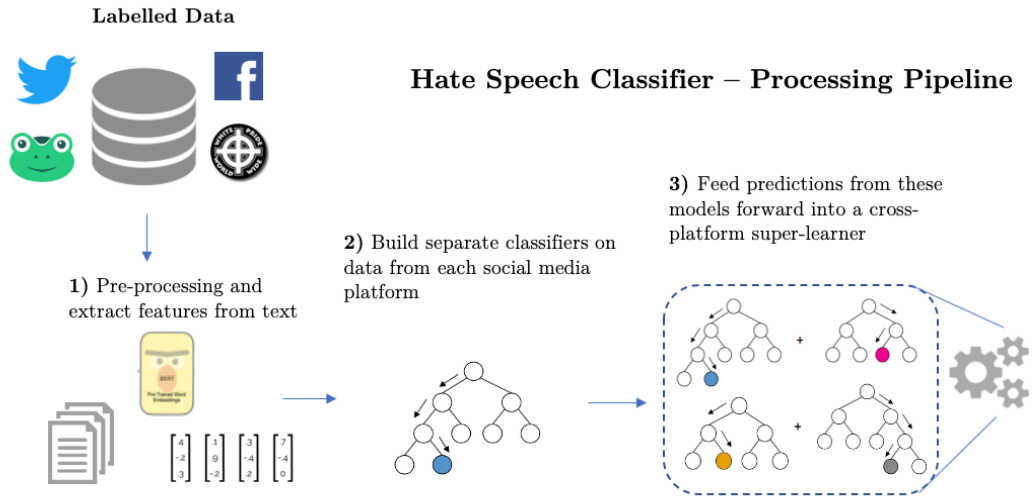


Figure 3 – The modelling pipeline for our hate speech detection and type of hate classification process

Syntactic features

Syntactic features include counts of the number of words, sentences, punctuation, and pronouns, plus the presence of negation. We also calculated the complexity of the messages measured as the lexical density (the ratio of content words to total words (Halliday, 1989) and their readability, measured with Flesch-Kincade reading ease scores (Kincaid, Fishburne, Rogers, & Chissom, 1975). These have both been shown to be predictive features of online hate speech in certain contexts (Zhang, Robinson, & Tepper, 2018), while more complex language is inversely related to extreme out-group derogation (Park & DeShon, 2018; Zmigrod, Rentfrow, & Robbins, 2019).

Machine learning approach - modelling

Our modelling takes a two-stage approach. We first built four separate platform-specific classifiers, one for each social media platform. Once these four models were trained and the hyper parameters tuned, we then combined them using a superlearner approach to improve performance while retaining platform-specific characteristics (Figure 3).

Step 1- Platform-specific classifiers

For each platform we built two binary classifiers, one trained on the labels 'hate speech' vs 'not hate speech' (classifier 1) and another trained on 'clean' vs 'not clean' (classifier 2). We then combined the outcomes of these two models using the formula below to create a single ordinal classification ranked

by increasing severity from 'clean' to 'offensive' to 'hate speech. For ordered classes this approach outperforms the naïve approach which treats the classes as an unordered set, and can be applied to any machine learning algorithm (Frank & Hall, 2001). Where *p(clean)* and *p(not clean)* are obtained from classifier 1 and *p(hate speech)* and *p(not hate speech)* are obtained from classifier 2.

Clean (lowest severity) : $p\ (clean) = 1 - p\ (not\ clean)$

Offensive (intermediate severity): $p\ (offensive) = p\ (not\ clean) - p\ (hate\ speech)$

Hate speech (highest severity): $p\ (hate\ speech) = 1 - p\ (not\ hate\ speech)$

We split the data into a training and a test set (80% and 20%) for each platform, stratified by categories to ensure balance to the 'real-world' (full) dataset. This test set was then held out for final model verification. The modelling validation was done using 10-fold cross validation (in addition to the 20% held out sample for final model verification), again with samples stratified on message category within each fold. In order to avoid information leakage between training and validation steps we included a number of measures within the cross-validation process. In each fold the data was down-sampled to a ratio of 2:1 from the majority class to the minority class using the original classes (Hate/Offensive/Clean) to truly reflect the underlying data. All predictors were then centred, scaled, and all the near zero variance predictors removed. The feature space of BERT word vectors was reduced using a supervised partial least squares (PLS) approach to maximise the separation of principle features between classes (Kuhn & Johnson, 2013), which has been shown to improve performance of word vector models (Gupta, Giesselbach, Rüping, & Bauckhage, 2019; Raunak, Gupta, & Metze, 2017). Using this PLS approach, we find that the vector space for the primary principle components populated by messages from across these three classes is different, but overlapping, suggesting that the semantic embeddings of the classes should offer some level of discrimination, especially in combination more components and other features (Figure 4).

We tested five types of models - C50 decision trees, general linear models, support vector machines, neural networks, and gradient boosted decision trees (xgboost). We found that the latter, xgboost (Chen & Guestrin, 2016), was the best performing model on all datasets and so retained this type of model for the remainder of the platform-specific analysis. For details on hyperparameter tuning, cross validation and equivalence zones see SI 1.5. Feature distinction for both semantic and syntactic features is shown in Figure SI 2.

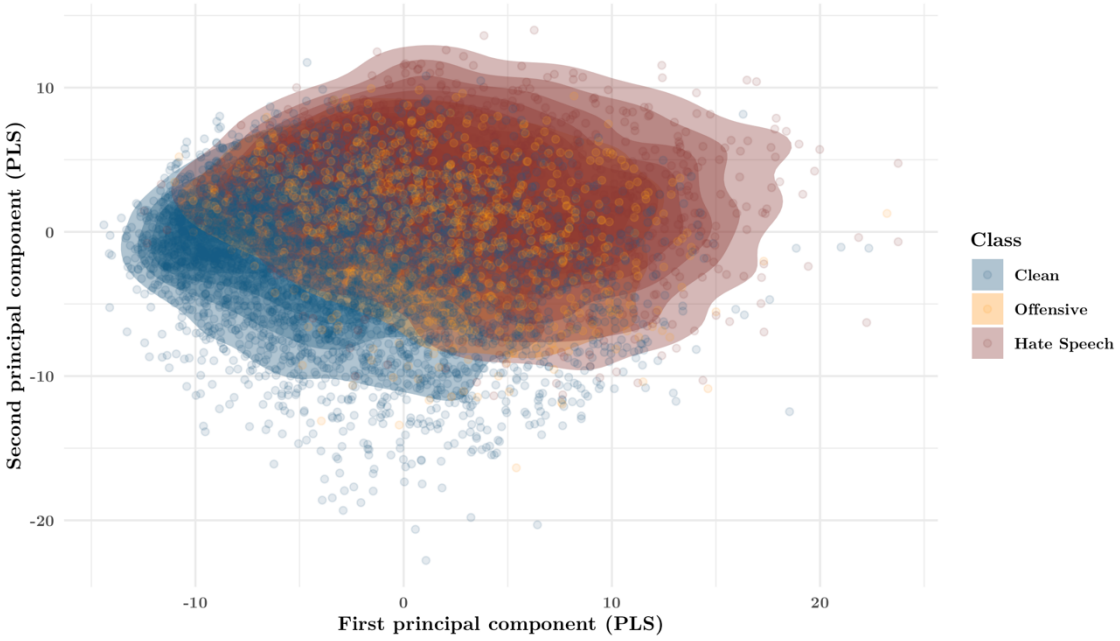


Figure 4 – Partial Least Squares (PLS) dimensionality reduction showing a separation in contextual word embeddings across the three message categories

Step 2 - Cross-platform superlearner

In training the superlearner we tested the same five machine learning algorithms again. This time we selected neural networks as they performed best in cross-fold validation. The superlearner takes platform-specific predictions from the models in step 1 as features and builds a new model by learning the optimum combination of these predictions to maximise performance. In addition to the predictions across the three classes for the four platform-specific models, each message's original platform was also included as a feature, giving a total of 13 features for the model to learn from.

To avoid the need to retain a further dataset for training the superlearner, whilst also avoiding information leakage between training and validation sets, we used the cross-validation predictions (rather than full predictions) from the platform-specific model corresponding to the original platform of the training data. We used predictions from the three other full platform-specific models. For example, for a Gab message in the training dataset of the superlearner, the predictions which serve as features were those from the full Facebook, Twitter and Stormfront models and the cross-validation Gab model. This means a model was never trained and tested on the same data points

Classifier validation on novel data

To compare our superlearner approach to platform-specific and existing 'state of the art' approaches, we made predictions on a new random sample of 1,000 annotated posts from a different platform, Reddit. We tested the prediction performance of four different models on this Reddit dataset.

Firstly, for comparison to the existing literature, we got predictions using 'HateSonar' (Nakayama, 2018), a Python implementation of a three-class model trained solely on Twitter data (Davidson et al., 2017). This provides a baseline performance for applying hate speech detection models to unseen data from a novel social media platform.

We then trained our own platform-specific 'Reddit' model, created using the same approach as for our four platform-specific models described above, giving a baseline for models trained on a small dataset of platform-specific data.

Following this we then got predictions for two superlearner models. Firstly, we used our original superlearner model described above (superlearner 1.0), i.e. trained on data from four social media platforms but not on Reddit data. We then created an updated superlearner (superlearner 2.0) trained on the outputs of all five platform-specific models, including the new Reddit model described above. This allowed us to verify the performance improvements provided by this cross-platform approach, as well as the opportunity it provides for model updating as new data becomes available. Both the Reddit-only model and the superlearner 2.0 models were created using a 60% training set with a 40% test set held out for performance measurements. We selected a larger percentage test set for this Reddit data, compared to prior datasets, in order to ensure valid results given the smaller dataset size.

Data analysis

All machine learning and analysis were done in R (version 3.6.1) using the tidyverse (Wickham et al., 2019) and tidymodels (Kuhn, Wickham, & RStudio, 2020) collections of packages. Where particular additional packages have been used, they have been referenced in the text.

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Ethics Reference: SSH_OII_CIA_19_062). All data collection was conducted using open-source methods and publicly available data, and hence, informed consent was not explicitly obtained.

Results

To assess the validity of our approach, we first compare the performance of the cross-platform superlearner to that of platform-specific models. Secondly, we compare these performance measures to those from the existing literature for identical or similar datasets. Finally, we compare the performance of our superlearner model on a completely new and unseen dataset from a different social media platform to an existing alternative state-of-the-art hate speech detection model.

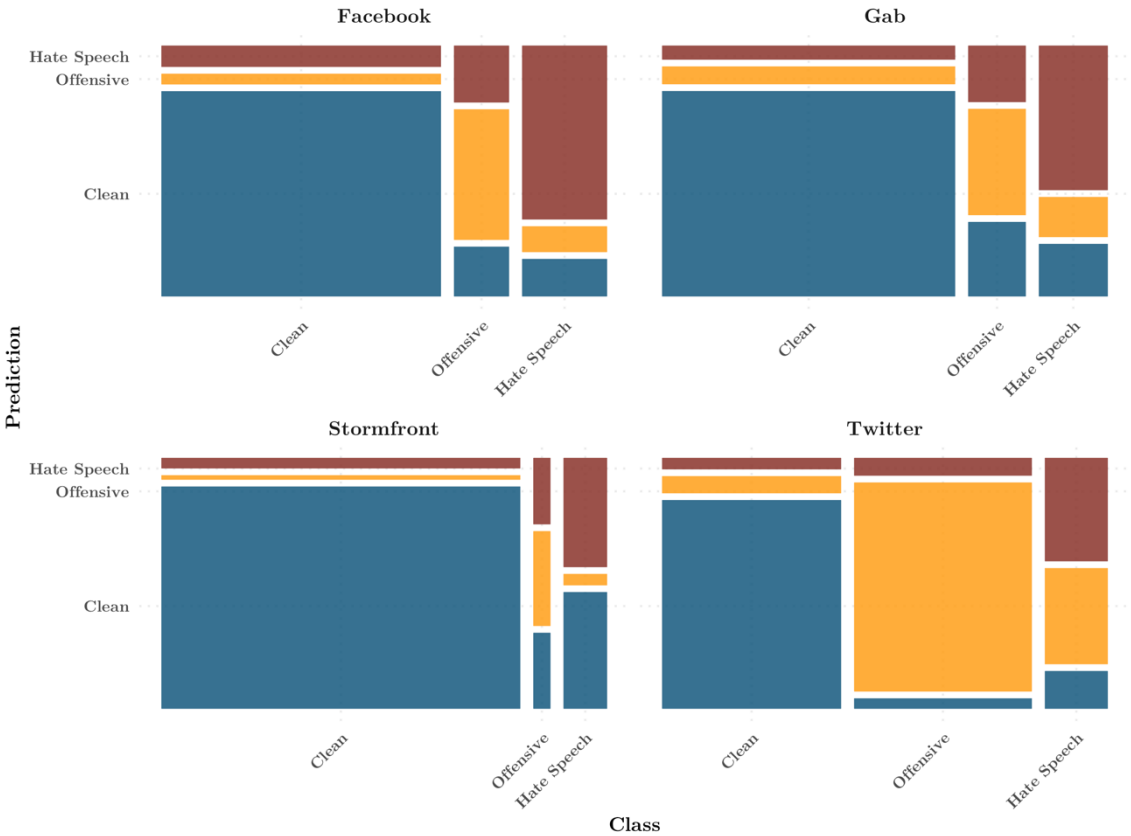


Figure 5 – Mosaic plots for performance on different classes for the superlearner across the four platforms. Colours related to data considered (x-axis) or classified (y-axis) as 'clean (blue), 'offensive' (yellow) or 'hate speech' (red).

Table 2 – Classifier performance for models trained on individual platforms and the superlearner

	Platform-Specific Model Performance			Superlearner Performance			Superlearner Change		
Platform	**Accuracy**	**HS Precision**	**HS Recall**	**Accuracy**	**HS Precision**	**HS Recall**	**Accuracy**	**HS Precision**	**HS Recall**
Overall	89.8	0.68	0.69	91.4	0.73	0.66	+1.6	+0.05	-0.03
Facebook	89.3	0.75	0.83	90.3	0.75	0.87	+ 1.0	0	+ 0.04
Gab	89.8	0.73	0.75	90.9	0.75	0.83	+ 1.2	+ 0.02	+ 0.08
Twitter	89.6	0.76	0.40	89.9	0.71	0.36	+ 0.3	- 0.05	- 0.04
Stormfront	90.5	0.50	0.77	93.9	0.67	0.42	+ 3.4	+ 0.18	- 0.35

Superlearner vs platform-specific models

The superlearner approach correctly labelled 91.4% of the messages across all four datasets, with an F1 score of 0.69 for hate speech detection (precision 0.73, recall 0.66), an F1 score of 0.96 for detection of clean messages (precision 0.95, recall 0.97), and an F1 score of 0.86 for offensive messages (precision 0.86, recall 0.86). The performance figures and confidence intervals from across all classes, platforms, and models are given in SI Table 3 (See SI 2.1 for confusion matrix and ROC curves).

Table 2 gives the classifier performance specifically for hate speech detection for models trained on individual platforms and the superlearner performance across all four platforms, along with the relative change in performance that the superlearner approach achieved. The superlearner implementation improved overall classifier accuracy by 1.6 percentage points compared to the individual platform-specific models, and this improvement was largest on Stormfront (3.4 percentage points) and Gab (1.2 percentage points). Specifically, this approach increased the precision of hate speech classification by 0.05 at the expense of recall which decreased by 0.03 (see SI 2.2 for exploration of minor/major classification errors). In other words, this suggests that the superlearner model performed better at detecting hate speech than platform-specific models.

When inspecting the performance of platform specific models (Figure 5, SI Table 3) it is apparent that all four models performed well in correctly identifying the largest class of clean messages (F1 range 0.94 to 0.95, shown in blue). Performance was more mixed however when presented with offensive messages (yellow), with a higher proportion of these being incorrectly classified as either clean or hate speech when compared to predictions for clean messages (F1 scores for offensive

classifications: Facebook 0.71, Gab 0.61, Stormfront 0.51). The exception to this is the Twitter model, which performed well when presented with offensive messages (Twitter offensive F1 score: 0.91), reflecting the higher proportion of messages from this class in the training set. With regards to hate speech (red) there were again difference in performance across the platforms, with Facebook and Gab models correctly recalling a higher proportion of hate messages (0.83 and 0.75 respectively), whereas the Twitter model failed to detect a much larger proportion of these messages (recall 0.40), and the Stormfront model achieved lower precision in hate speech classifications (precision 0.50). Interestingly, classification errors were not evenly distributed across incorrect classes, and instead followed the ordinal pattern, with hate speech messages more likely to be incorrectly classified as offensive than clean. The reverse is true for clean messages, which were more likely to be incorrectly classified as offensive than hate speech. This effect is shown clearly in the mosaic plots for Facebook and Gab in Figure 5.

For platform-specific models, the results also showed that when making predictions for data from a different social media platform to the training data (for example getting predictions for the Facebook data from the Twitter model), performance was reduced (SI Table 4). While performance was above chance for all combinations of models and platforms, it was significantly lower than the performance for congruent platform/data combinations, and also lower than the superlearner performance. The distributions of hate speech predictions for one platform against the others in shown in Figure SI 5, indicating how the overlap between platform-specific hate speech varies.

Comparison to existing approaches

Our hate speech detection model performed better in comparison to the published literature using the same or similar datasets.

For Twitter data, Davidson et al (2017), used a tf-idf bag of words approach with some additional features (lexical density, sentiment, and syntactic features) produced an overall model accuracy of 90% and precision and recall values for hate speech at 0.44 and 0.61 respectively[1]. Our approach reported similar overall accuracy performance for but improved the precision of hate speech detection

[1] *Note – Davidson et al appear to train the final model on the entire dataset and then get predictions for these same tweets i.e. no held out test set was used. This is likely to overestimate performance, particularly around recall and this should be taken into consideration when comparing performance.*

substantially. Our superlearner achieved 0.71 precision for hate speech messages on Twitter (Table 2) vs the 0.44 precision for this same data reported by Davidson et al, therefore reducing the presence of false positives.

For Gab data, a similar approach by Kennedy et al, (2020) using a model trained with BERT word embeddings reported an overall classifier accuracy of 87%, with precision and recall values for hate speech detection at 0.59 and 0.57 respectively. Our superlearner approach therefore improved both precision and recall substantially, reducing both false positives and false negatives for hate speech detection by approximately 20 percentage points.

For the Stormfront data, a recurrent neural network with short-term memory (LSTM) produces an overall accuracy of 78% under similar conditions (de Gibert et al., 2018), which is 16 percentage points lower than the accuracy our superlearner achieved. Precision and recall values are not available for comparison. There is no comparable model available for Facebook, however Facebook themselves reported in May 2020 that they detect 89% of hate speech messages posted to the platform using automated techniques (Rosen, 2020), which is broadly in line with our observed performance.

Performance on data from a new platform

In order to investigate how well this superlearner approach (superlearner 1.0) performed when tested on data from a novel social media platform not included in the original training sets, we measured its performance on a set of messages from Reddit, and compared this performance to an existing model (HateSonar) trained using data from just a single social media platform. For completeness, we also trained a new model using this Reddit data and included performance from this model in our comparisons, as well as from a superlearner updated with this new Reddit specific model (superlearner 2.0).

When testing performance on data from a novel social media platform not included in the original training set, we find that all three of our models (Reddit-only, original superlearner 1.0 and updated superlearner 2.0) performed substantially better than the HateSonar implementation of Davidson et al's approach (Table 3). This is particularly pronounced in the recall of hate speech messages, where the HateSonar model detected only 18% of the hate speech messages in the Reddit dataset while both of the superlearner approaches correctly detected 87% of these same messages – a substantial improvement.

The superlearner 1.0 model and newly trained platform-specific model achieved similar performance, with the platform-specific model scoring a higher overall accuracy but the superlearner 1.0 performing better on hate speech precision and recall, overall increasing the F1 score for hate speech detection by 0.02. The best performance however was obtained by combining the newly created Reddit-model with the pre-trained models from the other four platforms via the superlearner approach. This updated superlearner approach (superlearner 2.0) achieved an overall accuracy of 89.0% across the three classes and an F1 score of 0.92 for hate speech predictions (precision 0.96, recall 0.87). The performance figures for all classes and models on this Reddit dataset are given in SI Table 4.

Table 3 – Classifier performance on new and unseen Reddit data for the HateSonar model, platform-specific classifier, and the superlearner models

Model Performance on Reddit Data

Model	**Accuracy %**	**HS Precision**	**HS Recall**	**HS F1**
HateSonar	61.6	0.81	0.18	0.30
Reddit Platform-Specific	86.3	0.87	0.84	0.85
Superlearner 1.0	83.6	0.88	0.87	0.87
Superlearner 2.0	89.0	0.96	0.87	0.91

Discussion

Our cross-platform approach for hate speech detection outperformed existing models, both on data similar to that which these models are trained on, and for unseen data from a novel social media platform not included in the original training set. This highlights the opportunities of cross-platform approaches to improve automatic detection of online hate speech. In addition, we found that performance on novel platform data was easily improved by adding to the superlearner a new model trained on a small dataset from this novel platform, demonstrating the flexibility and ease of updating provided by this approach.

Benefits of the Superlearner

i) Leveraging Cross Platform Information

A key benefit from our approach is to combine predictions from multiple platform-specific datasets. This is found to improve performance on hate speech detection over using a data from a single social media platform in isolation. This improved performance is likely because new types of hate speech or specific formats of hate expression in the test data from a platform-specific model may be present in the training data from other platforms. Our cross-platform approach helps solve the challenge of data sparsity and lack of variability in hate speech detection training data (Schmidt & Wiegand, 2017) by leveraging datasets from across multiple social media platforms. Interestingly we find that simply applying models in an incongruent cross-platform approach without using this superlearner does not lead to good performance – platforms trained on just one social media platform do not extrapolate their predictions well to another platform without adjustment. This suggests that the nature of hate does vary significantly across platforms (which is the variability we wish to capture) but creating any cross-platform model that does take into account the platform from which a message originated will struggle to perform at a high level.

Interestingly, we find that overall this improvement in performance is driven by increased precision of predictions; in other words the superlearner is less likely to make a false positive hate speech judgement than the individual platform models, while the proportion of false negatives is increased slightly. The applicability of this approach may therefore be more useful in cases where this trade-off (fewer false positives) is preferable. This effect is platform-specific however. For Gab data for example, we observe that the superlearner improves both hate speech precision and recall; reducing both false positive and negatives.

Another benefit of our approach is that it will reduce the impact of biases or errors in training data, because the platform-specific model trained on this data will then be compared to other predictions from other datasets and models. Estimates of the Twitter data from Davidson et al (2017) suggest that up to 10% of the data is mislabelled, and this will reduce performance, especially on new out-of-sample datasets. In addition, the data has been shown to be skewed towards racism and sexism compared to other forms of hate speech (Vidgen, Tromble, et al., 2019). Similarly, the Stormfront dataset from de Gibert et al (2018) focuses primarily on white supremacy rather than other forms of hate. Using multiple disparate datasets to build models and combining predictions will reduce the impact of errors and biases in any one dataset.

ii) Future updating and expansion

The accuracy of hate speech detection models can degrade quickly as the nature of online language changes (Florio et al., 2020; Laaksonen et al., 2020). Keeping models up to date with new training data is therefore important to preserve accuracy. However, re-training the entire model is computationally expensive, and risks destructive interference if new data contains errors or differences in labelling from prior training data. Our superlearner approach is expandable without the need to retrain each single-platform model as only the new data needs to be used to train a new model. This reduces computational costs, and these predictions can easily be combined with existing predictions in a new superlearner, as we demonstrate with a new Reddit dataset. While the cross-platform superlearner (1.0) performed better than existing approaches on this dataset, training and combining a new small platform-specific model with existing pre-trained platform models into a new superlearner (2.0) noticeably improved performance. Our model can therefore be easily updated when new datasets becomes available by 'slotting' them into the existing superlearner.

iii) Ordinal distinction between offensive and hate speech

Another benefit of our approach is to include an ordinal approach to hate speech detection. Our model made fewer major category errors (misidentifying clean speech with hate speech or visa-versa) than minor category errors (SI section 2.2), which provides evidence for ordinal information being contained in features across the three classes. Training platform-specific models using this ordinal approach, with the addition of the theoretically informed features that we have included, improved performance within the Gab dataset compared to a comparable classifier containing just a binary 'hate' and 'not hate' distinction (benchmarking with Kennedy et al 2020). Future research should

investigate whether a more nuanced distinction between types of extreme digital speech and the creation of more than three classes can further improve this ordinal approach.

Limitations and future directions

Our superlearner approach shows potential for future automated hate speech detection, but limitations remain. First, variation in hate speech definition across studies and labelled datasets (Vidgen & Derczynski, 2020) makes comparing models or combining datasets challenging; in this study we had to manually re-label data for consistency. Combining predictions into a superlearner may smooth some variation but problems may still occur. Consistency in definitions and class labels will help the field move forwards and widen the applicability of this superlearner approach.

Bias in data labelling is another limitation which the field of hate speech detection should aim to identify and reduce. Prior models of hate speech detection found, for instance, that messages in African-American English were more susceptible to being misclassified as hate speech than those in 'standard American English' (Davidson, Bhattacharya, & Weber, 2019). These racial biases need to be addressed in any machine learning system prior to deployment to avoid discrimination, e.g. with balanced training sets from across a wide range of domains and users. Biases also occur less often with data labelled by expert annotators (Sap, Card, Gabriel, Choi, & Smith, 2019; Waseem, 2016), therefore datasets labelled by expert annotators should be used whenever possible and weighed higher than crowdsourced datasets in a superlearner approach. Here we used two annotators of different genders, nationalities, and backgrounds to reduce bias in data labelling. Bias can also be introduced at the stage of extracting word vectors however (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016). This is especially pernicious as interpretation of word vectors cannot be easily performed, and therefore the bias gets propagated forwards silently. The impact of bias in these pre-trained word vector models should be investigated further. Future work should also incorporate multiple languages as hate groups increasingly operate globally (Davey & Ebner, 2017; Sigurbergsson & Derczynski, 2019).

Automatic detection systems of extreme digital speech are also susceptible to adversarial attacks and perturbations of language (Gröndahl et al., 2018; Hosseini, Kannan, Zhang, & Poovendran, 2017), including spelling and grammar changes. This can lead to a lower judgement of hate by the classifier, but not decrease the impact on a human recipient of the message. Future research should take inspiration from the field of image recognition and include perturbations of data within the training set to help mitigate against this type of attack (Nandy, Hsu, & Lee, 2020).

Finally, while we used pre-trained general-purpose natural language models to get context-dependent word embeddings, recent improvements in training these language models on online-specific and topic-specific conversation data (Müller, Salathé, & Kummervold, 2020) may improve the context awareness and performance of the subsequent models. Pre-training BERT models on social media platforms containing hate speech is likely to improve classifier performance and is a promising avenue for future research.

Conclusion

Overall, our approach combining datasets from across multiple social media platforms shows promise in building better automatic tools to detect online hate speech and can help address the challenges of data scarcity and low variability, and improve applicability for novel social media platforms not included in the original training data. In the context of growing threats from extremism, intergroup conflict, and online hate, developing robust and adaptable methods for automatically detecting these threats is an important step in improving online safety.

Acknowledgements

We thank Davidson et al (2017), de Gibert et al (2018), Quian et al (2019), Zannettou et al (2018), and Pushift.io for making their data available and also the research assistants who helped annotate the datasets used in this analysis.

References

Alammar, J. (2018). *The Illustrated BERT, ELMo, and co: How NLP Cracked Transfer Learning*. Retrieved from http://jalammar.github.io/illustrated-bert/

Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). The enemy among us: Detecting hate speech with threats based "othering" language embeddings. *ACM Transactions on the Web*, *9*(4), 1–26. Retrieved from http://arxiv.org/abs/1801.07495

Anti-Defamation League. (2019). Hate symbols database. Retrieved July 21, 2020, from https://www.adl.org/hate-symbols

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *International World Wide Web Conference*, (2), 759–760. https://doi.org/10.1145/3041021.3054223

Bartlett, J., Reffin, J., Rumball, N., & Williamson, S. (2014). Anti-social media. *Demos*. Retrieved from https://demos.co.uk/project/anti-social-media/

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 4356–4364.

Burke, S. (2017). *Anti-Semitic and Islamophobic discourse of the British far-right on Facebook. Loughborough University*. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/27177/1/Thesis-2017-Burke.pdf

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0072-6

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Cihon, P., & Yasseri, T. (2016). A biased review of biases in Twitter studies on political Collective Action. *Frontiers in Physics*, *4*(August), 1–8. https://doi.org/10.3389/fphy.2016.00034

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2019). Cross-platform evaluation for Italian hate speech detection. *CEUR Workshop Proceedings*, *2481*. Retrieved from http://ceur-ws.org/Vol-2481/paper22.pdf

Davey, J., & Ebner, J. (2017). *The fringe insurgency: Connectivity, convergence and mainstreaming of the extreme right*. Retrieved from http://www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Retrieved from http://arxiv.org/abs/1905.12516

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 512–515. Retrieved from http://arxiv.org/abs/1703.04009

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 11–20. https://doi.org/10.18653/v1/w18-5102

Delgado, R. (1982). Words that wound: A tort action for racial insults, epithets, and name calling. *Harvard Civil Rights-Civil Liberties Law Review*, *17*, 89–110. https://doi.org/10.4324/9780429502941-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*. Retrieved from http://arxiv.org/abs/1810.04805

Elveljung, D. (2018). *The use of homophobic pejoratives among gamers: Discourse analysis of slurs within the gaming sphere*. Gothenburg University. Retrieved from https://gupea.ub.gu.se/bitstream/2077/57935/1/gupea_2077_57935_1.pdf

Facebook. (2020). Facebook community standards: Hate speech. Retrieved July 9, 2020, from https://www.facebook.com/communitystandards/hate_speech

Fernandez, M., Asif, M., & Alani, H. (2018). Understanding the roots of radicalisation on Twitter. *WebSci '18: 10th ACM Conference on Web Science*, 1–10. https://doi.org/10.1145/3201064.3201082

Florio, K., Basile, V., Polignano, M., Basile, P., & Patti, V. (2020). Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, *10*(12), 4180. https://doi.org/10.3390/app10124180

Fontgate Media. (2014). A list of 723 bad words to blacklist. Retrieved from https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4). https://doi.org/10.1145/3232676

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In *Lecture Notes in Computer Science* (pp. 145–156). https://doi.org/10.1007/3-540-44795-4_13

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2020). Online contact between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society*, 1–44.

Ganesh, B., & Bright, J. (2020). *Extreme digital speech: Contexts, responses and solutions*. Retrieved from https://www.voxpol.eu/new-vox-pol-report-extreme-digital-speech-contexts-responses-and-solutions/

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. Retrieved from https://www.semanticscholar.org/paper/SPSS-for-Windows-Step-by-Step%3A-A-Simple-Guide-and-George-Mallery/230e458b34cdcb463cfe4caa954253bd73456e2e

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215–230. https://doi.org/10.14257/ijmue.2015.10.4.21

Glen, S. (2014). Cohen's kappa statistic. Retrieved from https://www.statisticshowto.com/cohens-kappa-statistic/

Greevy, E., & Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 468–469. https://doi.org/10.1145/1008992.1009074

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love": Evading hate speech detection. *Proceedings of the ACM Conference on Computer and Communications Security*, 2–12. https://doi.org/10.1145/3270101.3270103

Gupta, V., Giesselbach, S., Rüping, S., & Bauckhage, C. (2019). Improving word embeddings using kernel PCA. *Proceedings Ofthe 4th Workshop on Representation Learning for NLP*, 200–208. https://doi.org/10.18653/v1/w19-4323

Halliday, M. A. K. (Michael A. K. (1989). *Spoken and written language*. Oxford University Press. Retrieved from https://books.google.co.uk/books?id=T9RpAAAACAAJ&redir_esc=y

Harel, T. O., Jameson, J. K., & Maoz, I. (2020). The normalization of hatred: Identity, affective polarization, and dehumanization on Facebook in the context of intractable political conflict. *Social Media + Society*. https://doi.org/10.1177/2056305120913983

Harmon, J., & Bratt, J. (2020). jonathanbratt/RBERT: Implementation of BERT in R. Retrieved July 20, 2020, from https://github.com/jonathanbratt/RBERT

Home Office. (2018). Hate crime, England and Wales, 2017/18. *Statistical Bulletin*, 1–40. Retrieved from

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/748598/hate-crime-1718-hosb2018.pdf

Hornik, K. (2016). Package "openNLP": Apache OpenNLP Tools Interface. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/openNLP/openNLP.pdf

Hosseini, H., Kannan, S., Zhang, B., & Poovendran, R. (2017). Deceiving Google's perspective API built for detecting toxic comments. *ArXiv*. Retrieved from http://arxiv.org/abs/1702.08138

Internet Live Stats. (2020). Twitter Usage Statistics. Retrieved July 20, 2020, from https://www.internetlivestats.com/twitter-statistics/

Jurgens, D., Chandrasekharan, E., & Hemphill, L. (2020). A just and comprehensive strategy for using NLP to address online abuse. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 3658–3666. Retrieved from https://www.aclweb.org/anthology/P19-1357.pdf

Karan, M., & Šnajder, J. (2018). Cross-Domain Detection of Abusive Language Online. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 132–137. https://doi.org/10.18653/v1/w18-5117

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., … Dehghani, M. (2020). The Gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv*, 1–47. Retrieved from https://osf.io/edua3/#!

Kennedy, C. (2017). Guide to SuperLearner. *CRAN*, 1–23. Retrieved from https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease Formula) for Navy enlisted personnel. *Institute for Simulation and Training*, *56*. https://doi.org/ERIC #:ED108134

King, P. (2018). Building resilience for terrorism researchers. *VoxPol*. https://doi.org/10.4324/9781315168272-17

Kleinberg, B., Vegt, I. Van Der, & Gill, P. (2020). The temporal evolution of a far-right forum. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-020-00064-x

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (1st ed.). Springer. Retrieved from https://link.springer.com/book/10.1007%2F978-1-4614-6849-3

Kuhn, M., Wickham, H., & RStudio. (2020). Package 'tidymodels': Easily Install and Load the "Tidymodels" Packages. *CRAN*, 1–5. Retrieved from https://cran.r-project.org/web/packages/tidymodels/index.html

Laaksonen, S.-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The datafication of hate: Expectations and challenges in automated hate speech monitoring. *Frontiers in Big Data*, *3*, 1–16. https://doi.org/10.3389/fdata.2020.00003

Law Commission. (2020). *Hate crime laws: A consultation paper*. Retrieved from https://www.lawcom.gov.uk/project/hate-crime/

Liu, H., Alorainy, W., Burnap, P., & Williams, M. L. (2019). Fuzzy multi-task learning for hate speech type identification. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. https://doi.org/10.1145/3308558.3313546

Liu, H., Burnap, P., Alorainy, W., & Williams, M. (2020). Scmhl5 at {TRAC}-2 shared task on aggression identification: Bert based ensemble learning approach. *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 62–68. Retrieved from https://www.aclweb.org/anthology/2020.trac-1.10

Liu, H., Burnap, P., Alorainy, W., & Williams, M. L. (2019). A fuzzy approach to text classification with two-stage training for ambiguous instances. *IEEE Transactions on Computational Social Systems*, *6*(2), 227–240. https://doi.org/10.1109/TCSS.2019.2892037

Ma, C. (2014). *What are you laughing at? A social semiotic analysis of ironic racial stereotypes in Chappelle's Show*. Retrieved from http://www.lse.ac.uk/media-and-communications/assets/documents/research/msc-dissertations/2014/Cindy-Ma-Reformatted-Dissertation-AF.pdf

Magu, R., & Luo, J. (2018). Determining code words in euphemistic hate speech using word embedding networks. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 93–100. Retrieved from http://www.aclweb.org/anthology/W18-5112

Marwick, A., & Miller, R. (2014). Online harassment, defamation, and hateful speech: A primer of the legal landscape. *Fordham Center on Law and Information Policy Report*, (2), 1–74.

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 173–182. https://doi.org/10.1145/3292522.3326034

Mishra, P., Yannakoudakis, H., & Shutova, E. (2019). Neural character-based composition models for abuse detection. *ArXiv*, 1–10. https://doi.org/10.18653/v1/w18-5101

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, *16*(4 SPEC. ISS.), 372–403. https://doi.org/10.1093/pan/mpn018

Müller, K., & Schwarz, C. (2020a). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *00*(0), 1–37. https://doi.org/10.1093/jeea/jvaa045

Müller, K., & Schwarz, C. (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *SSRN Electronic Journal*, 1–47. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

Müller, M., Salathé, M., & Kummervold, P. E. (2020). COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *ArXiv*. Retrieved from http://arxiv.org/abs/2005.07503

Nakayama, H. (2018). Hironsan/HateSonar: Hate speech detection library for python. Retrieved July 20, 2020, from https://github.com/Hironsan/HateSonar

Nandy, J., Hsu, W., & Lee, M. L. (2020). Approximate manifold defense against multiple adversarial perturbations. *ArXiv*. Retrieved from http://arxiv.org/abs/2004.02183

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *ArXiv*. Retrieved from http://arxiv.org/abs/1804.05704

Orlando, E., & Saab, A. (2020). Slurs, stereotypes and insults. *Acta Analytica*, *35*(4), 599–621. https://doi.org/10.1007/s12136-020-00424-2

Park, G., & DeShon, R. P. (2018). Effects of group-discussion integrative complexity on intergroup relations in a social dilemma. *Organizational Behavior and Human Decision Processes*, *146*(March), 62–75. https://doi.org/10.1016/j.obhdp.2018.04.001

Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Detecting offensive language in Tweets using deep learning. *ArXiv*, 1–17. https://doi.org/10.1007/s10489-018-1242-y

Pohjonen, M. (2018). *Horizons of hate: A comparative approach to social media hate speech*. Retrieved from https://www.voxpol.eu/download/vox-pol_publication/Horizons-of-Hate.pdf

Polley, E. C., & van der Laan, M. J. (2010). *Super learner in prediction. U.C. Berkeley Division of Biostatistics Working Paper*. Retrieved from http://biostats.bepress.com/ucbbiostat/paper266/

Qian, J., Bethke, A., Liu, Y., Belding, E., & Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 4755–4764. Retrieved from https://www.aclweb.org/anthology/D19-1482.pdf

Quinn, T., Tuckwood, C., & Boyd, D. (2019). Hatebase. Retrieved July 20, 2020, from https://hatebase.org/

Raunak, V., Gupta, V., & Metze, F. (2017). Effective dimensionality reduction for word embeddings. *31st Conference on Neural Information Processing Systems*, 235–243. https://doi.org/10.18653/v1/w19-4328

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W. (2018). Characterizing and detecting hateful users on Twitter. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 676–679. Retrieved from https://arxiv.org/pdf/1803.08977.pdf

Rinker, T. (2020). Package "qdap": Bridging the Gap Between Qualitative Data and Quantitative Analysis. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/qdap/qdap.pdf

Rizoiu, M.-A., Wang, T., Ferraro, G., & Suominen, H. (2019). Transfer learning for hate speech detection in social media. *ArXiv*. Retrieved from http://arxiv.org/abs/1906.03829

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press. https://doi.org/10.2307/j.ctvhrcz0v

Rosen, G. (2020). *Facebook community standards enforcement report, May 2020 Edition*. Retrieved from https://about.fb.com/news/2020/05/community-standards-enforcement-report-may-2020/

Rossini, P. (2019). Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*. https://doi.org/10.1177/0093650220921314

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2020). HateCheck: Functional Tests for Hate Speech Detection Models. Retrieved from http://arxiv.org/abs/2012.15606

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. gyo, Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1–34. https://doi.org/10.1186/s13673-019-0205-6

Samaratunge, S., & Hattotuwa, S. (2014). *Liking Violence: A study of hate speech on Facebook in Sri Lanka*.

Retrieved from https://www.cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings Ofthe 57th Annual Meeting Ofthe Association for Computational Linguistics*, 1668–1678. Retrieved from https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings Ofthe Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/w17-1101

Schnoebelen, T., Silge, J., & Hayes, A. (2020). Package 'tidylo': Weighted Tidy Log Odds Ratio. *CRAN*. https://doi.org/10.1093/pan/mpn018

Sellars, A. F. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *20*. https://doi.org/10.1093/jicj/mqaa023

Sigurbergsson, G. I., & Derczynski, L. (2019). Offensive language and hate speech detection for Danish. *ArXiv*, 1–13. Retrieved from http://arxiv.org/abs/1908.04531

Stecklow, S. (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*. Retrieved from https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Technau, B. (2018). Going beyond hate speech: The pragmatics of ethnic slur terms. *Lodz Papers in Pragmatics*, *14*(1), 25–43.

UK Safer Internet Centre. (2016). *Creating a Better Internet for All*. Retrieved from http://childnetsic.s3.amazonaws.com/ufiles/SID2016/Creating a Better Internet for All.pdf

van der Laan, M., Polley, E., & Hubbard, A. (2007). *Super Leaner. UC Berklely Division of Biostatistics Working Paper Series*. Retrieved from https://biostats.bepress.com/ucbbiostat/paper222/

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, 1–26. Retrieved from http://arxiv.org/abs/2004.01670

Vidgen, B., Margetts, H., & Harris, A. (2020). *How much online abuse is there? A systematic review of evidence for the UK*. Retrieved from https://www.turing.ac.uk/research/research-programmes/public-policy/online-hate-monitor

Vidgen, B., Tromble, R., Harris, A., Hale, S., Nguyen, D., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *3rd Workshop on Abusive Language Online*, 1–14. Retrieved from https://www.aclweb.org/anthology/W19-3509/

Vidgen, B., Yasseri, T., & Margetts, H. (2019). Trajectories of Islamophobic hate amongst far right actors on Twitter. *ArXiv*, 1–20. Retrieved from https://arxiv.org/pdf/1910.05794

Ward, K. D. (1998). Free speech and the development of liberal virtues: an examination of the controversies involving flag-burning and hate speech. *University of Miami Law Review*, *52*(3), 733–792.

Warner, W., & Hirschberg, J. (2012). Detecting hate speech on the world wide web. *Proceedings of the 2012 Workshop on Language in Social Media*, 19–26. Retrieved from http://dl.acm.org/citation.cfm?id=2390374.2390377

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*, 138–142. https://doi.org/10.18653/v1/w16-5618

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017). Understanding abuse: A typology of abusive language detection subtasks. *Proceedings Ofthe First Workshop on Abusive Language Online*, 78–84. https://doi.org/10.18653/v1/w17-3012

Wickham, H., Averick, M., Bryan, J., Chang, W., D', L., Mcgowan, A., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Williams, M., & Mishcon de Reya. (2019). *Hatred behind the screens A report on the rise of online hate speech*. Retrieved from https://www.mishcon.com/upload/files/Online Hate Final 25.11.pdf

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). *ArXiv*. Retrieved from http://arxiv.org/abs/1903.08983

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? *ArXiv*. Retrieved from

https://arxiv.org/abs/1802.05287

Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. *Proceedings of IMC '18*. https://doi.org/10.1007/BF02768242

Zhang, Z., & Luo, L. (2018). Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *Semantic Web*, *1*(0), 1–5. https://doi.org/arXiv:1803.03662v1

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. *Lecture Notes in Computer Science*, (June), 29–44. https://doi.org/10.1007/978-1-4020-4749-5_3

Zliobaite, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. *ArXiv*, 1–24. Retrieved from https://www.win.tue.nl/~mpechen/publications/pubs/CD_applications15.pdf

Zmigrod, L., Rentfrow, P., & Robbins, T. (2019). Cognitive inflexibility predicts extremist attitudes. *Frontiers in Psychology*, *10*(May), 1–13. https://doi.org/10.3389/fpsyg.2019.00989

Supplementary Information (SI) for Chapter 2:

Leveraging cross-platform data to improve automated hate speech detection

1 – Modelling approach

1.1 – Class balance and dataset size for the four platforms used in training

SI Table 1 gives the final size of the four training sets (Facebook, Gab, Twitter and Stormfront) used to build the platform-specific hate speech classifiers after pre-processing the data, along with the proportions across the three message classes.

SI Table 1- Size of datasets used for classification and balance of classes once the data were cleaned and pre-processed

		Class Balance		
	Number of Messages	**Prop. Clean**	**Prop. Offensive**	**Prop. Hate Speech**
Facebook	9,999	0.70	0.12	0.18
Gab	8,274	0.71	0.13	0.16
Twitter	9,730	0.43	0.43	0.14
Stormfront	9,715	0.87	0.04	0.10

1.2 – Extracting webpage title from URL

In the data pre-processing steps, where possible we extracted the web page title for URLs which were included in messages. This was done using regular expressions to extract text which occurred within a character string after either 'http://' or 'https://' and after the first occurring full stop and before the final forward slash. We then removed any punctuation that occurred between words in this URL. We were not able to obtain meaningful text every single URL in this way, but succeeded to identify a large majority. For example, URLs changed by link shorteners could not be processed.

As an example, for the URL '*https://dailystormer.ws/britain-to-give-houses-and-jobs-to-returning-isis-fighters/'* we would be able to use this approach to extract the words '*britain to give houses and jobs to returning isis fighters'*. This gives substantially more context for the remainder of the message that was posted alongside the URL and is likely to assist with hate speech detection.

1.3 Overview of Semantic and Syntactic features

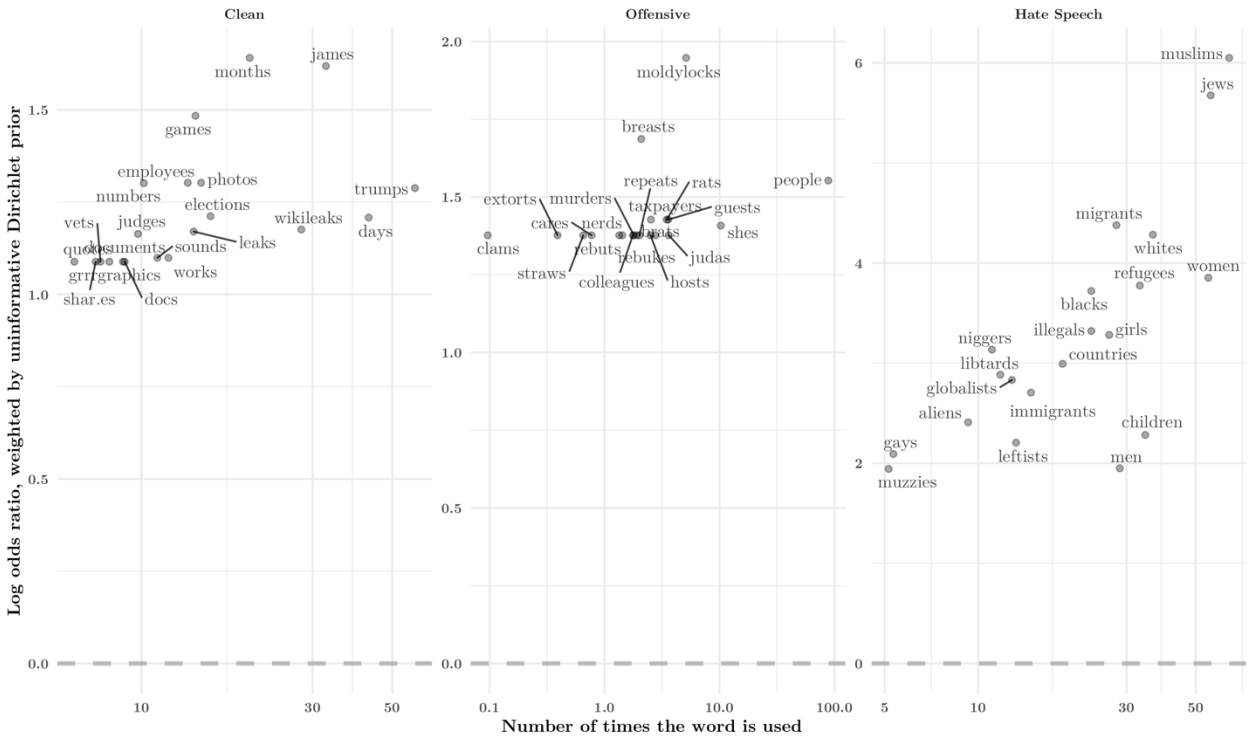
Table SI 2 gives an overview of the semantic and syntactic features used in building the platform-specific hate speech models. Implementation of each feature is given in the second column.

Table SI 2- Overview of semantic and syntactic features used in platform-specific models

	Feature	Implementation
Semantic	BERT Word Vectors (x 768)	rBERT (with Google Colab GPU)
	Plural Nouns / log odds	Apache OpenNLP / 'tidylo' R package
	Othering	Dictionary Approach
	Hate Terms	HateBase - Average scores
	Hate Severity	HateBase - Dictionary approach
	Hate Symbols	ADL - Dictionary approach
	Obscenity	Dictionary approach
	Sentiment	qdap implementation R
Syntactic	Lexical Density	Manual Calculation
	Flesch-Kincaid Reading Ease	Manual Calculation
	Document Length	Number of characters
	Punctuation	Count
	Negation	qdap implementation R
	Pronouns	Dictionary approach

1.4 – Log-odds for plural nouns

We identified plural nouns in the messages using the Apache OpenNLP Maxent Part of Speech tagger from the 'OpenNLP' package (Hornik, 2016). To identify the relevance of these plural nouns for each label category, we calculated the weighted-log-odds of these nouns and included them as additional features in the model, using the 'tidylo' package (Schnoebelen, Silge, & Hayes, 2020). Log-odds are an effective way to represent words that capture differences in political speech and to evaluate the relative importance of those words (Monroe, Colaresi, & Quinn, 2008). They offer an advantage over the more traditional term frequency / inverse term frequency (tf-idf) for calculating the importance of words within a corpus as they do not penalise common words across all classes. This is beneficial in our case, as some plural nouns are common across the whole dataset but also informative when combined with other features and so should not be penalised (e.g. 'Muslims' will be common in both hateful and non-hateful discussions on the topic of religion).



SI Figure 1 - Log-odds values for plural nouns from across the three text classes within the Gab training data (Clean, Offensive and Hate Speech).

SI Figure 1 shows the top plural nouns for each category within the Gab training data. Offensive and derogatory plural nouns, as well as standard terms for groups which often receive online abuse, receive high scores within the hate speech category, while the top terms in the clean category by contrast are not pejorative.

1.5 – Hyperparameter tuning and equivalence zones

In training all models, hyper-parameter tuning was done within the 10 fold cross validation, initially using a random grid search across standard values, and then tuned further during a second round focusing in on the values identified in the initial round (Kuhn & Johnson, 2013). For the xgboost platform-specific models, the final parameters for which a grid search was performed within the cross-validation are given below.

min_n – 10 : 30
mtry – 50 : 368 (Upper value equal to number of parameters after dimensionality reduction)
loss reduction – 10 : 30
learning rate - 0.05 : 0.1
number of trees – 1000 : 3000

Following model training, we defined an ordinal equivalence zone (Kuhn & Johnson, 2013) as when the difference in probability for the first (most likely) class and the second class was less than the inverse of the number of classes (in this case 0.33). Predictions in this zone were rejected as the models were not decisive enough.

1.6– Feature distinction for machine learning

SI Figure 2 shows the distribution of semantic and syntactic features across the three text categories (Clean, Offensive, Hate Speech) that were used in building the machine learning classifier. These values are normalised and so do not reflect the originally measured ranges. The high degree of overlap for all of these features suggests that in isolation they cannot be used and should be combined with broader semantic features.

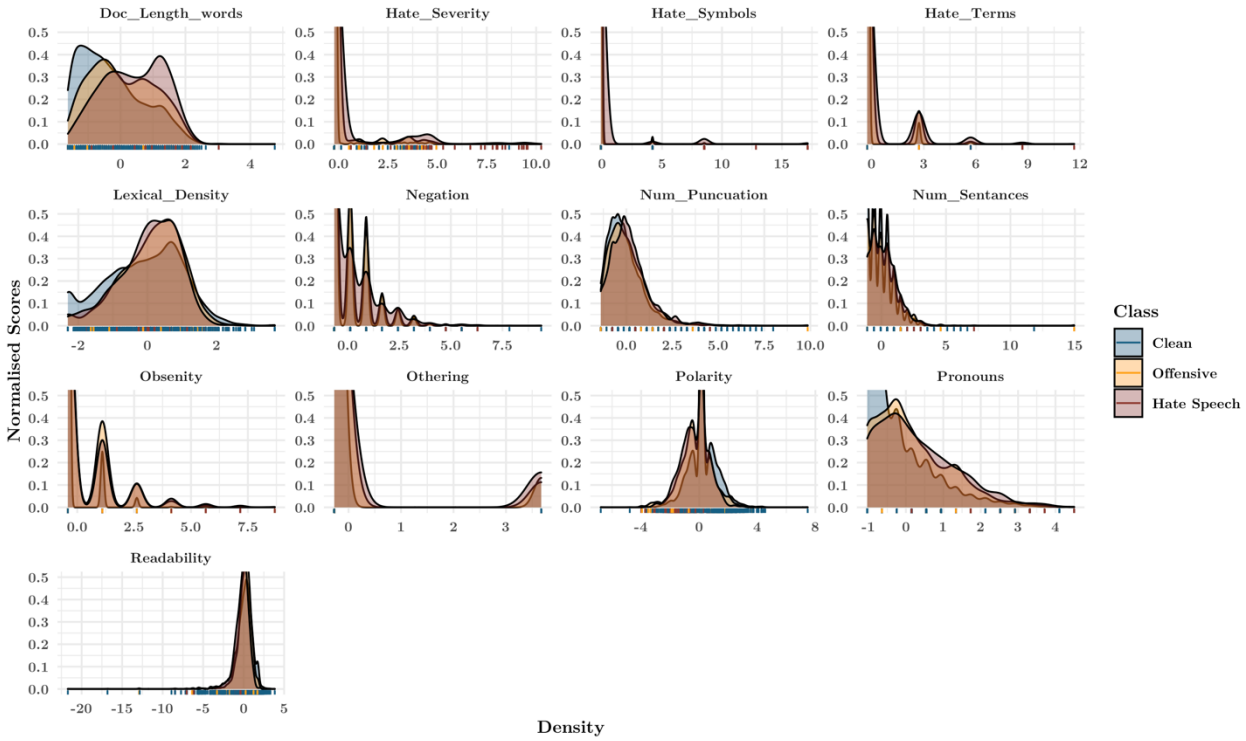


Figure SI 2 – Normalised distributions of features used to train the platform-specific classifiers for the three message classes (Clean, Offensive and Hate Speech)

2 – Machine Learning Models Performance

2.1 – Full platform-specific and superlearner model performance across three classes

Table SI 3 gives the performance metrics across all three classes for both the platform-specific models and the superlearner model. Accuracy, recall, precision and F1 scores are given for both the overall performance on the whole dataset, and specific performance on each social media platform.

2.2 ROC Curves for the Final Superlearner Model

Figure SI 3 shows the receiver operator characteristic curves (ROC curves) for the final superlearner model across the three classes of message.

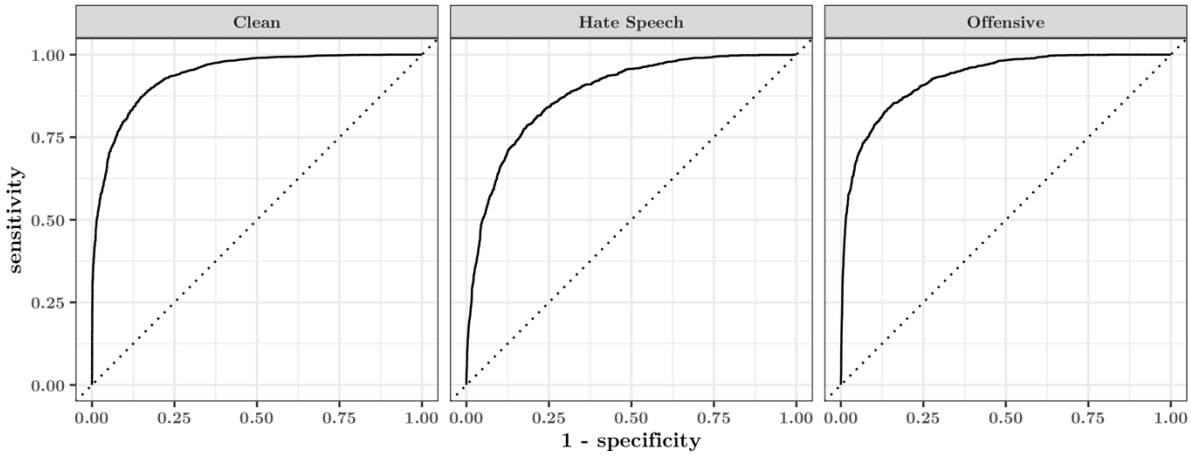


Figure SI 3 – ROC Curves for the superlearner model across the three classes (Clean, Offensive and Hate Speech)

Table SI 3 – Performance metrics (accuracy, precision recall and F1) for the three classes across platform specific and superlearner models

Platform	*Model*	*Overall Accuracy*		*Clean*			*Offensive*			*Hate Speech*		
		%	+/-	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Overall	Platform Specific	89.8	0.9	0.95	0.95	0.95	0.83	0.85	0.84	0.69	0.68	0.69
	Superlearner	91.4	0.8	0.95	0.97	0.96	0.86	0.86	0.86	0.73	0.66	0.69
Facebook	Platform Specific	89.3	1.8	0.94	0.95	0.94	0.79	0.64	0.71	0.75	0.83	0.78
	Superlearner	90.3	1.8	0.96	0.95	0.95	0.82	0.64	0.72	0.75	0.87	0.80
Gab	Platform Specific	89.8	2.1	0.95	0.96	0.95	0.66	0.58	0.61	0.73	0.75	0.74
	Superlearner	90.9	2.0	0.94	0.97	0.96	0.64	0.29	0.4	0.75	0.83	0.79
Twitter	Platform Specific	89.6	1.6	0.93	0.97	0.95	0.88	0.94	0.91	0.76	0.40	0.52
	Superlearner	89.9	1.6	0.94	0.95	0.95	0.88	0.96	0.92	0.71	0.36	0.48
Stormfront	Platform Specific	90.5	1.8	0.98	0.92	0.95	0.39	0.75	0.51	0.50	0.77	0.60
	Superlearner	93.9	1.3	0.96	0.98	0.97	0.67	0.58	0.62	0.67	0.42	0.52

2.3 Confusion Matrix for final Superlearner Model

Figure SI 4 shows the confusion matrix for the final super learner model across the three classes of message (Clean/Offensive/Hate Speech)

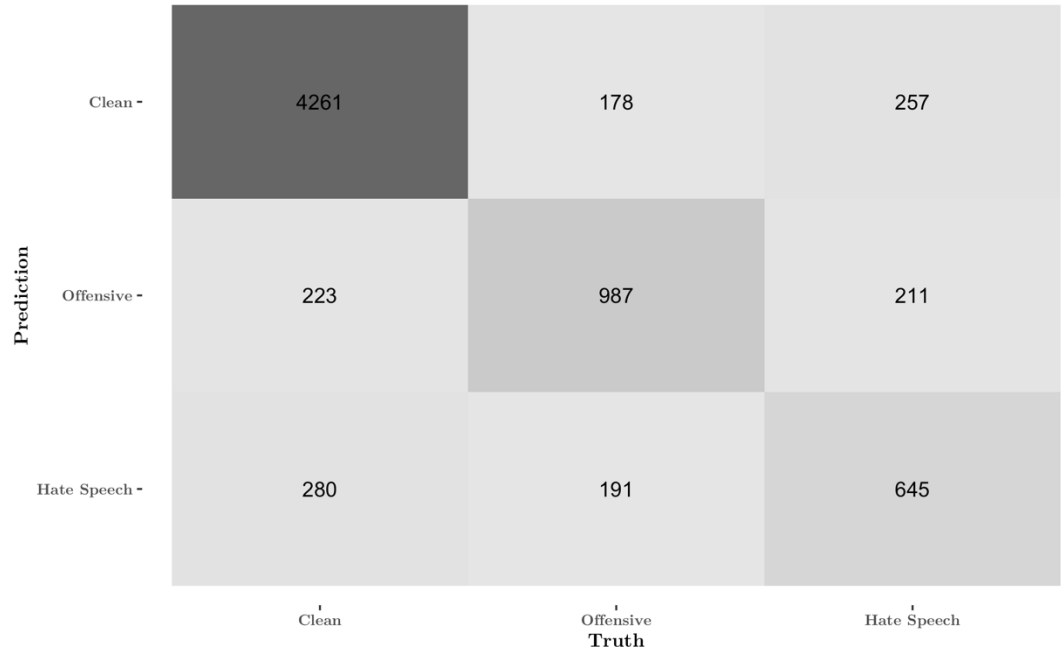


Figure SI 4 – Confusion matrix for the superlearner model across the three classes (Clean, Offensive and Hate Speech)

2.4 Error inspection

When assessing accuracy values for the hate speech detection models, it is important to consider that not all incorrect classifications are equal, and minor category errors (i.e. between consecutive classes) is a smaller mistake than a major category error (between non-adjacent categories). To take this into account, we calculated the percentage of major category errors in the overall dataset at the superlearner level. We found that only 5.88% of clean messages were mislabelled as hate speech and 23.10% of hate speech messages were mislabelled as clean.

Investigating this further using the superlearner predictions on the Gab dataset as an example, removing the 'minor errors' increases performance in hate speech detection to 0.86 precision and 0.86 recall, up from 0.75 and 0.83.

2.5 Performance on novel Reddit data

Table SI 4 gives the performance metrics for the new Reddit data. These scores are given for the four types of model tested on this dataset: the HateSonar trained on Twitter data alone (Nakayama, 2018), a platform-specific model trained on Reddit data, the original superlearner presented in this paper (superlearner 1.0) and an updated superlearner model which includes the addition of the Reddit platform-specific predictions at the superlearner stage (superlearner 2.0).

Table SI 4 – Performance metrics for Reddit data across four types of model; HateSonar, a platform-specific Reddit model, the original Superlearner and an updated Superlearner model with the inclusion of the Reddit platform specific model

Model	*Overall Accuracy*		*Clean*			*Offensive*			*Hate Speech*		
	%	+/-	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
HateSonar	61.6	3.1	0.87	0.84	0.85	0.89	0.37	0.52	0.19	0.81	0.30
Reddit Platform Specific	86.3	5.1	0.94	0.88	0.91	0.52	0.83	0.64	0.87	0.84	0.85
Superlearner 1.0	83.6	3.0	0.95	0.85	0.90	0.55	0.70	0.62	0.88	0.87	0.87
Superlearner 2.0	89.0	4.5	0.93	0.95	0.94	0.50	0.70	0.58	0.96	0.87	0.92

4 Cross platform (incongruent) performance on platform-specific models

Table SI 3 gives the performance for platform-specific models when giving prediction on incongruent platform data i.e. messages from the other three platforms which were not included in the training data. Predictions for the models on messages from that platform along with superlearner performance are also given in the main text in table 2.

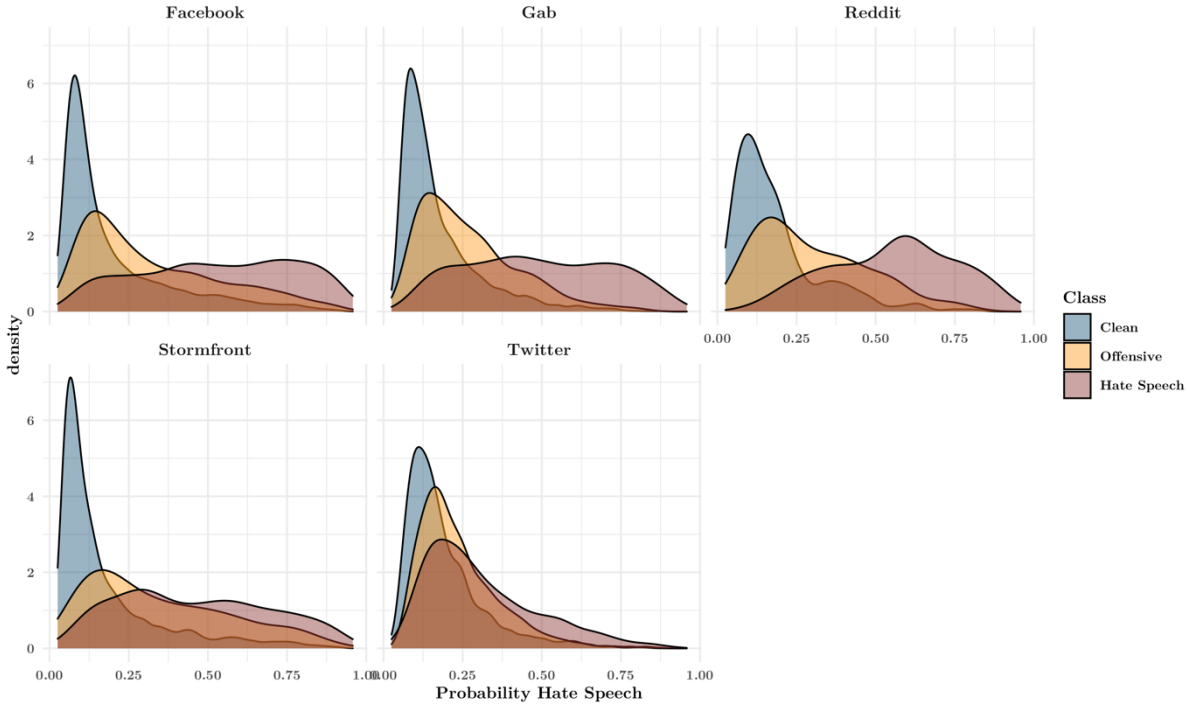
Overall, we see that performance is lower when making predictions for data from other social media platforms. This suggests that the nature of the hate speech varies across platforms, both in semantic and syntactic features. We did not observe a drop in performance for models on test sets compared to cross-validation so this is not due solely to over-fitting to the training data although we cannot rule this out completely or that the overfitting does not occur in the sampling approach (i.e. over sampling a certain type of hate speech from a specific platform).

Despite this overall reduction, we also observe substantial variation depending on which pair of platforms are being compared. We find for example that the performance is reasonably high when comparing Gab and Facebook models in both directions suggesting that these platforms share in the nature of the hate speech that they host, while predicting Twitter hate speech from Stormfront models performs poorly as does predicting Gab hate speech from the Twitter model. The latter of these gives particularly low levels of hate speech recall with nearly 9 out of every 10 hate speech messages missed. Indeed, we find that the first of these (Twitter models predicting Stormfront messages) performed at chance levels. All other models performed above chance overall however, showing that there is still useful information shared across platforms and why the superlearner approach allows us to improve performance overt single platform models.

Investigating this in more detail we find that the distribution of hate speech predictions for each classifier varies depending on the platform that hate speech messages come from. Figure SI 5 shows this for predictions across all 5 platforms (this time with the addition of the Reddit data) for the Gab specific model. This shows that platforms in the top row (Facebook, Reddit and Gab) have similar levels of distinction in predictions between classes, and therefore information is being used to make meaningful distinctions, while platforms in the bottom row (Stormfront and Twitter) have much less clear distinctions in the prediction distributions across classes.

Table SI 4- Comparison of model performance in detecting hate speech for datasets from different social media platforms

Model \ Data	Facebook	Gab	Twitter	Stormfront
Facebook	Accuracy – 82.3%	85.6	72.3	93.1
	HS Precision – 74.5	57.9	31.3	63.2
	HS Recall – 82.7	57.9	12.6	38.7
Gab	86.7	89.8	74.3	91.6
	67.6	73.0	34.0	52.8
	75.8	75.0	14.1	68.8
Twitter	81.3	78.5	89.6	90.1
	66.2	66.7	76.2	62.5
	28.7	11.0	39.8	26.9
Stormfront	88.1	84.4	73.5	90.5
	64.8	58.8	29.2	49.6
	76.7	62.2	14.7	76.7



SI Figure 5 – Distribution of hate speech predictions for the Gab specific classifier for the five platforms across the three messages classes (Clean, Offensive and Hate Speech)

SI References

Hornik, K. (2016). Apache OpenNLP Tools Interface. CRAN. Retrieved from https://cran.r-project.org/web/packages/openNLP/openNLP.pdf

Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling (1st ed.). Springer. Retrieved from https://link.springer.com/book/10.1007%2F978-1-4614-6849-3

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4 SPEC. ISS.), 372–403. https://doi.org/10.1093/pan/mpn018

Raunak, V., Gupta, V., & Metze, F. (2017). Effective Dimensionality Reduction for Word Embeddings. 31st Conference on Neural Information Processing Systems, 235–243. https://doi.org/10.18653/v1/w19-4328

Schnoebelen, T., Silge, J., & Hayes, A. (2020). Package 'tidylo': Weighted Tidy Log Odds Ratio. CRAN. https://doi.org/10.1093/pan/mpn018

Chapter 3

Hate Contagion: Measuring the spread and trajectory of hate on social media

An earlier version of this chapter was presented at the 2020 European Consortium for Political Research General (ECPR) Conference, and published in the conference proceedings.

Gallacher, J, D. (2020) The ontogeny of online hate speech: Do social media platforms drive increased hate or reflect existing prejudices? *Paper prepared for presentation at the ECPR General Conference, 25-28 September 2020*

https://gc.ecpr.eu/Filestore/paperproposal/988e78ed-b4e1-4378-a260-092fd83ae994.pdf

Abstract

Online hate speech is a growing concern, with minorities and vulnerable groups increasingly targeted with extreme denigration and hostility. The drivers of hate speech expression on social media are unclear, however. This study explores how hate speech develops on a fringe social media platform popular with the far-right, Gab. We investigate whether users seek out this platform in order to express hate, or whether instead they develop these opinions over time through a mechanism of socialisation, as they interact with other users on the platform. We find a positive association between the time users spend on the platform and their hate speech expression. We show that while some users do arrive on these platforms with pre-existing hate stances, others develop these expressions as they get exposed to the hateful opinions of others. Our analysis reveals how hate speech develops online, the important role of the group environment in accelerating its development, and gives valuable insight to inform the development of counter measures.

Introduction

Far-right extremist violence has been increasing across the western world in recent years, with a 320% increase since 2015 (Global Terrorism Index, 2019). This phenomenon has occurred in parallel with an increase in online hate speech (Williams, 2019). A growing number of studies show how those with prejudiced or negative views towards minority groups are using the social web to spread antagonistic, inflammatory and hateful messages targeting these groups (e.g. Alorainy, Burnap, Liu, & Williams, 2018; Zannettou, ElSherief, Belding, Nilizadeh, & Stringhini, 2020). This could lead to a reduction in group relations and an increase in intergroup conflict, with particular targeting of vulnerable and minority groups.

Social media has also begun to play a larger role in encouraging offline hate crimes (Williams & Burnap, 2019; Gaudette, Scrivens, & Venkatesh, 2020). Suspects in several recent hate-related terror attacks had an extensive history of social media activity within hate groups, including notably the Tree of life Synagogue shooting Pittsburgh and the Charlottesville 'Unite the Right' rally, both in the United States (Evans, 2019). Social media was also used to livestream the 2019 terror attack at two mosques in Christchurch, New Zealand. This latter event sparked a number of copy-cat attempts, demonstrating the power and influence of online hate (MacAvaney et al., 2019). However, the exact role that social media activity played in facilitating or influencing these hateful beliefs is unclear.

Online hate typically takes the form of verbal or written attacks against a specific group of people focused on, or motivated by, certain aspects of that group's identity (Chapter 2; Davidson, Warmsley, Macy, & Weber, 2017; de Gibert, Perez, García-Pablos, & Cuadros, 2018). Hate speech is particularly common in expressions shared by the far-right, both within traditional right-wing extremist groups and the more recent 'alt-right' movements (Mathew, Dutt, Goyal, & Mukherjee, 2019; Vidgen, Yasseri, & Margetts, 2019). These far-right groups are broadly defined by holding ideals of racial and ethnic nationalism, and typically framed in terms of white power and/or white identity; this forms the ingroup. Outgroups are framed in terms of perceived threats to this identity posed by a combination of non-whites, Jews, Muslims, immigrants, refugees, members of the LGBTQ community, and feminists, among others (Conway, Scrivens, & Macnair, 2019; Scrivens, 2020). The Internet is a key breeding ground for these hateful and extremists' narratives (All-Party Parliamentary Group (APPG) on Hate Crime, 2019).

Despite its prevalence, the influence that interacting with online hate speech has on processes of extremism and radicalisation are contested (Meleagrou-Hitchens & Kaderbhai, 2016), and the use of empirical evidence to investigate how online hate spreads is limited. Evidence suggests that exposure to extreme material on social media can exert influence on users, including the promotion of outgroup derogation, but how this relates to users' own opinions and long-term behaviours is unclear. Understanding how hate spreads on social media platforms is paramount to developing any mitigation strategies. In particular, how the hate speech expressions of individual users fluctuates over time, and this influences those they are connected to, is poorly understood (Kleinberg, Vegt, & Gill, 2020).

In addition, not all users active on fringe social media platforms express hate against outgroups (Kleinberg et al., 2020). One reason for this variation between users may be due to differences in exposure to hate speech on the platform, with this influencing their own hate expression (Ferrara, 2017; Ferrara, Wang, Varol, Flammini, & Galstyan, 2016), with some users experiencing greater social pressure to conform, i.e. to produce hate speech themselves. This has not been tested, however.

In the current study we investigate the nature of hate speech expression on the fringe social media platform Gab and examine the role that social influence plays in propagating hate. We first test whether users express hate speech and display pre-existing prejudices from the moment they join the platform, or instead whether hate speech expression grows with time spent on the platform. We

subsequently look at the social contagion effects of online hate speech, and whether greater exposure to online hate increases a user's likelihood of adopting this behaviour and producing hate speech themselves. Finally, we examine whether hate exposure can lead to transitive effects across target groups – i.e. whether exposure to hate against a particular group can lead to wider hate against other groups. By measuring the role of this social influence, we capture the way in which social media users affect each other's beliefs, emotions, attitudes, and behaviours, and as such give insight into the complex ways that hateful behaviours develop. Studying these effects is key to understanding the spread of hate speech online and its consequences, which is vital to inform countermeasures.

Literature Review

In order to understand what drives users to express hate speech on social media, it is first essential to understand the social psychological underpinnings of outgroup denigration and how collective hate can develop, and the existing evidence for the role of social media in this process.

Hate as a group process

Recent evidence suggests that there is no pathological 'terrorist' or 'extremist' personality, but rather terrorism and hateful extremist violence occur as group phenomena (Doosje et al., 2016). Ingroup and outgroup distinctions can be a strong driver within the radicalisation process, especially when the outgroup is felt to present a direct threat (Sageman, 2004). Therefore, it is important to study hate speech in the context of intergroup relations and to understand the collective group dynamics to which individuals belong and take their social identity from. This social identity gives individuals a shared and collective representation of who they are, how they should behave, and what values they hold, including which outgroups they denigrate (Hogg, Abrams, & Brewer, 2017; Tajfel, 1974). The importance of this group identity is also the exhibited by the victims of hate, who are typically targeted because of the groups they belong to rather than their individual actions or characteristics (Reicher, Haslam, & Rath, 2008).

When viewing offline hate as a group process, evidence suggests that counter to the traditional literature (Haney, Banks, & Zimbardo, 1973; Milgram, 1963), groups do not commit acts of extreme inhumanity due to a lack of awareness or control, but rather because they believe what they are doing is right, morally correct, and to be celebrated (Reicher et al., 2008). In this process, the ingroup is championed as (uniquely) good and the eradication of the outgroup is embraced as necessary to

defend ingroup virtue. Extremist positions are adopted when there is the belief that the survival of the ingroup is inseparable from direct or offensive action against the outgroup (Berger, 2018), and radicalisation is therefore the increased preparation and commitment to intergroup conflict (McCauley & Moskalenko, 2008). Group socialisation is thought to be especially important in the shift towards extremism, with new group members learning the norms of the group and developing shared expectations and belief system about how to behave (Levine & Moreland, 1994; Moreland & Levine, 1982). If the norms of the ingroup include derogation towards outgroups, then these can also be learned through socialisation (Bracegirdle, 2020). Smith, Blackwood & Thomas (2019) propose a group socialisation process of radicalisation where group members gradually adopt shared identification with a set of radical norms, including outgroup hate, through social interactions with the group that leverage their shared perceptions, experiences, and grievances, whilst also changing the nature of the group as a whole towards an extreme position.

Hate speech has two distinct roles in this group radicalisation process. Firstly, it signals ingroup membership to other ingroup members through a shared outgroup 'enemy'. Secondly, it dehumanizes and diminishes members of the targeted outgroup, expresses the threat posed by this group, highlights the shared grievance with other members of the ingroup, and calls for direct action against the perceived threat (Gagliardone, Gal, Alves, & Martinez, 2015). Hate speech therefore does not only represent a reduction in intergroup relations but also leads such relations to reduce further. Frequent and repetitive exposure to hate speech also increases outgroup prejudice by leading to desensitization to violence and subsequently to lower evaluations of the victims of this violence (Soral, Bilewicz, & Winiewski, 2018). Because social interactions focused around shared grievances are particularly powerful in driving extreme forms of social action (Thomas, Mcgarty, & Louis, 2014), understanding the spread of hate speech on social media and how it influences those exposed to it is therefore crucial.

The adoption of outgroup hate through social contagion

Offline, this spread of outgroup hate from one ingroup member to another can be thought of as social contagion, a form of social influence whereby individuals adopt novel behaviours as a result of interacting with others who are already behaving in this way (Rogers, 2003; Valente, 1995). In other contexts, the influence of this exposure on adoption of the behaviour increases with social pressure to conform (Asch, 1955). In this way, it is not the individuals themselves, but rather the relationships between individuals and their interpersonal interactions, which determine whether someone adopts the behaviour or not. Historically, this effect has been shown to occur in the adoption of novel farming

practices (Ryan & Gross, 1943), medical innovations (Coleman, Katz, & Menzel, 1966) and the use of communication networks (Rogers & Kincaid, 1984). More recently, this social contagion effect has been shown offline in the case of outgroup denigration, whereby social consensus plays a large role in shaping local social norms and acceptability of certain prejudices (Crandall & Stangor, 2008). Social contagion via networks of relationships can occur through direct ties and immediate contact with others, but also through indirect ties – connections of those we ourselves are connected with (Granovetter, 1973). This latter form of contact can expose individuals to wider information and greater novelty than direct connections.

More recently, the concept of social contagion has been applied to the study of online behaviours. In particular, it has been used to study the spread of information through online social networks (Lazer et al., 2009), with applications ranging from politics (Bond et al., 2012), technology (Onnela & Reed-Tsochas, 2010), public health (Centola, 2010), and moralised discourse (Brady, Crockett, & Van Bavel, 2020). Individuals who experience higher levels of information exposure are more likely to pass it on (Bakshy, Rosenn, Marlow, & Adamic, 2012), and weak ties are responsible for greater propagation of novel information through the network than strong ties (Bakshy et al, 2012). There is a strong social component to this process: users are more likely to share a piece of information if they believe a greater number of their close contacts have already shared it (Hodas & Lerman, 2014). In addition to informational influence, emotional states can also be transferred via contact with others. Those exposed to greater positive/negative emotional content have been shown to express this emotion more themselves (Kramer, Guillory, & Hancock, 2014), while aggressive states on Twitter have been shown to propagate from one user to another (Terizi, Chatzakou, Pitoura, Tsaparas, & Kourtellis, 2020).

In the context of extremist content, these online social contagion effects have been shown both for Islamic extremist and far-right material (Ferrara, 2017; Mathew et al., 2019). By mapping the spread of propaganda, influential Islamic extremist supporters on Twitter were shown to influence other previously non-supportive users, with the average Islamic extremist "infecting" 2.13 other users before being suspended (Ferrara, 2017). On far-right platforms, users expressing hate speech were shown to instigate larger information cascades than non-hateful users, indicating that these users are more likely to 'go viral' on the platform (Mathew et al., 2019). However, in these cases, contagion was measured by the successful propagation of extreme content, which may be biased by a number of factors such as user popularity, relative activity of more extreme users, and prevalence of this content.

As such, whether online exposure to hate speech from other users increases the likelihood that users will adopt hate speech themselves remains unknown. In controlled environments, exposure to online content implicating outgroups as perpetrators of violence is shown to increase support for violent policies against these outgroups, as well as increasing the likelihood users will make violent comments themselves (Javed & Miller, 2019), but whether this holds for 'real-world' online environments is unclear.

The spread of online hate is also shown to be influenced by offline 'trigger' events, such as terror attacks (Burnap et al., 2014), which can lead to the rapid spread of hate against the perceived perpetrator groups (Williams & Burnap, 2016). This spike in online hate quickly diminishes however, and on the whole, following these trigger events, messages containing explicit hate terms are less likely to be further shared by others than messages without hateful language (Burnap & Williams, 2015). This therefore suggests that online contagion effects may be driven by other factors, such as network position or social status.

The social dynamics of hate on social media

Given the prevalence of hate speech on social media, understanding the dynamics of how it spreads is vital for informing counter measures or mitigation strategies, however there is much still to be understood. A key question is whether hate speech expression on social media is a normative / socialisation effect or a fixed phenomenon, i.e. whether a user's propensity to express hate speech increases with time spent interacting with others within a social media platform, or whether users instead arrive with fixed views on outgroup denigration which they retain and express. This is an important question because it can provide insight into whether social media platforms are playing a driving role in the observed increases in far-right extremism by allowing hate filled conversations to take place on their platforms. Addressing this question is a key objective of this study.

There is already emerging evidence for effects of socialisation on intergroup conflict and hate speech expression on social media. For example, frequent users of Facebook have an increased perception of social distance and negative attitudes towards outgroups (Settle, 2018). On Twitter, approximately half of users who express Islamophobia change their rate of posting hate over time (Vidgen, Yasseri, & Margetts, 2019), suggesting they may be adapting to the norms of the environment. Similarly, while hateful expressions such as Islamophobia, homophobia, or racial slurs are considered norm violations

on most parts of mainstream platforms, within certain sub-communities on these platforms these expressions are considered acceptable and within 'normal' discourse, leading to much higher prevalence (Chandrasekharan et al., 2018). This high prevalence is maintained even when these groups are isolated from the wider online community (Chandrasekharan, Jhaver, Bruckman, & Gilbert, 2020). Similarly, qualitative analysis of online forums popular with the far right shows that they are used as a space for informal social learning of political ideology and community norms, as well as more formal recruitment or operational planning (Lee & Knott, 2020).

The nature of the language used in these more extreme environments also supports this idea, with marginal and extreme topics becoming mainstream over time (Braun et al., 2020). This shift supports ideas of contextualisation of conversations within the group, while the numerous ingroup jokes, coded language, and nebulous 'othering' present in fringe far-right platforms suggests that socialisation is occurring to allow for collective understanding (Tuters & Hagen, 2018, 2019). Evidence for socialisation effects also comes through investigating the impact of censorship on hate speech expression. Suspensions of hateful forums on Reddit (subreddits) leads to an influx of users from the banned spaces on other subreddits, but does not cause significant changes in hate speech within these spaces (Chandrasekharan et al., 2017) suggesting that hate expression is driven by the community or group level characteristics of the previous group--with the caveat that many users from the banned subreddits may not migrate to other subreddits and instead take their hate expression elsewhere on the web. Similarly, users banned by Twitter for spreading hate are more popular than random users (Chatzakou et al., 2017), suggesting that they had exerted social influence on other users prior to suspension. Despite these findings, it is unclear whether social media is instantiating hateful opinions in previously non-hateful users.

Alternative evidence however suggests that users seek out online hate communities to express their pre-existing opinions, and therefore arrive into an online space with already established prejudices. This is supported by findings that users migrate from less hateful to more hateful online spaces as their own hate expression increases (Ribeiro et al., 2020), suggesting they seek out areas to express these views. Similarly, hateful users on Twitter have newer accounts than non-hateful users, supporting the idea that this hate is not driven by social interactions over time (Ribeiro, Calais, Santos, Almeida, & Meira, 2018). Corroborating this, the relationship between exposure to extremist content on social media and self-reported political violence is strongest when individual actively seek

out extremist content compared to passive encounters (Pauwels & Schils, 2016). Overall, the role of socialisation in driving hate expressions remains therefore unclear.

These ideas need not be mutually exclusive however, and it is possible that users could arrive into a social media platform with existing prejudices and also develop these further through interactions with other users. This duality is demonstrated in studies of Reddit. Hateful subreddits have been shown to be fairly stable in their levels of hate over time, suggesting that the norms of the groups are fixed (Grover & Mark, 2019). Additionally, users on Reddit are shown to engage in a process of pre-entry learning prior to posting in a new subreddit, and to moderate the toxicity of their language to match that of the community that they are joining (Rajadesingan, Resnick, & Budak, 2020). This suggests that socialisation effects may occur prior to posting messages. However, the shift in language observed in those users does not affect their posts within other communities on the platform. This suggests that these users may also be seeking out these hateful communities to express hate which they would not share in more moderate spaces. More research is therefore needed to understand the role of hateful social media environments in driving user individual expressions of hate speech, whilst accounting for users' initial positions upon joining the platforms.

Secondary transfer effect of hate

If exposure to hate speech increases the likelihood that individuals will express hate themselves, an important question arising is whether this newly expressed hate will remain specific to the targeted group in the original hateful message(s) which the new expresser was exposed to, or whether this hate can spread to wider targets. Offline, increased contact between opposing groups is shown to reduce prejudice and conflict (Allport, 1954; Brown & Hewstone, 2005; Pettigrew & Tropp, 2008). This effect is not limited to the groups involved directly in the contact, but rather that there is a secondary transfer effect by which contact with a primary outgroup reduces prejudice toward other (secondary) outgroups (Pettigrew, 2009; Schmid, Hewstone, & Tausch, 2014; Tausch et al., 2010).

This secondary transfer effect occurs in part due to attitude generalisation – a process whereby attitudes towards one group generalise to another linked but less familiar group (Fazio, Eiser, & Shook, 2004) – and also due to ingroup reappraisal, whereby individuals realise that the specific ingroup norms they are familiar with are not the only acceptable set of social rules (Pettigrew, 1997). Together these effects give rise to the transfer of positive contact effects across groups (Tausch et al.,

2010). For example, contact with foreigners has been shown to lead to more positive attitudes towards this outgroup, but this positive effect also generalises towards other much wider outgroups that were not involved in the original contact situation, such as the homeless or LGBTQ+ groups (Pettigrew, 2009).

While much of the research in this field has focused on the constructive outcomes of positive intergroup contact, there is increasing evidence for a similar effect in reverse, with negative intergroup contact leading to more generalized negative outgroup attitudes (Jasinskaja-Lahti et al., 2020; Meleady & Forder, 2019). In the context of online hate speech, it is therefore possible that increased exposure to the negative attitudes of ingroup members towards one outgroup will lead to worsening attitudes against both this target outgroup and more widely against other outgroups. This may occur through an effect of generalised outgroup prejudice, or through a reduction/reversal of the ingroup reappraisal process, whereby the social norms of the ingroup are viewed as the only 'correct' or acceptable situation. To our knowledge this has yet to be tested.

The current research

This study aims to test whether participation in online conversations containing hate speech increases the likelihood that users will adopt the hateful norms of the group, through a process of social contagion, and whether users express greater levels of hate speech themselves as they spend longer on the platform.

We address this question using a fringe social media platform popular with the far-right and known to have high levels of hate speech, Gab. We construct our analysis in three steps. First, we identify hate speech in online conversations using a supervised machine learning approach, and determine the targeted group of this hate. Second, we build a picture of the overall trajectory of hate speech expressions for users on the platform by calculating the likelihood of a user expressing hate speech as they post successive messages. Finally, we use network contagion models to look at the structure of the network (i.e. which users are interacting with one another) and measure the network-wide patterns of user exposure to, and expression of, hate speech.

Gab – An ideal platform to study hate speech contagion

In this study we analyse messages from Gab, is a micro-blogging platform designed as a broad copy of Twitter (Benson, 2016; Weissman, 2016). Users post short messages which can be reposted, replied to, or 'liked' by other users of the platform. Gab is known for its lax hate speech moderation policies and the high proportion of messages containing explicit hate terms (Lima et al., 2018; Zannettou et al., 2018), and is seen as '*rolling out the welcome mat'* to users banned from social media sites (Anti-Defamation League, 2019). The platform is heavily focused on political content and topics closely follow current affairs, particularly around political ideology, race, and terrorism (Zhou, Dredze, Broniatowski, & Adler, 2019), including Nazi imagery, antisemitic material, holocaust denial, islamophobia, along with much wider forms of hate against other target groups (Weich, 2019). In addition, the platform has been used as a recruitment tool by several neo-Nazi and alt-right groups (Katz, 2018), and the rise in popularity of extremist far-right groups on Gab has been suggested to be "remarkably similar" to the rise of ISIS on mainstream social media in prior years (Makuch, 2019). In October 2018 Gab was taken offline after the shooter at the Tree of Life Synagogue attack, Pittsburgh, USA, was found to have posted anti-Semitic messages on Gab immediately prior to the attack and received substantial support from platform members (Mathew et al., 2019).

As the platform has hosted conversations which became increasingly problematic up to a point of extreme offline violence, it is a valuable case study in online hate. The high prevalence of hate speech additionally means that exposure to this type of content will be much higher than on mainstream platforms, with larger samples of hate messages allowing for more reliable estimates and more statistical power. In addition, Gab has been shown to have low levels of algorithmic filtering and artificial promotion of content compared to other social media platforms (Reed, Whittaker, Votta, & Looney, 2019). This makes Gab a valuable platform to study the effects of social exposure to content on online behaviours because users' connections on the platform are responsible for the patterns of content they are exposed to, rather than artificial promotion of content or algorithmic filtering (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015). Finally, the networked structure of the platform, akin to Twitter networks, means that content can theoretically propagate across the entire network and is not restricted to specific pages or groups as would be the case on a more compartmentalised platform such as Facebook, allowing us to investigate network effects more easily.

Research aims

This study aims to provide new insight into how hate develops on social media platforms.
More specifically, our first objective is to investigate whether users seek out social media platforms to express pre-established hate, or whether hate speech expression arises as users spend more time within these platforms. Based on theories of group extremism we hypothesise that users will increase their hate speech expressions over time spent on the platform.

Our second objective is to test whether online hate speech spreads through social contagion. Based on prior work on online informational and emotional influence we hypothesise that social contagion will occur and therefore expect to find that users will be more likely to express hate speech after having been exposed to it from other users.

Finally, we will aim to explore these effects of hate speech contagion in more detail, by testing (i) whether its effects are specific to certain hate types or whether they transfer prejudice across hate types, and (ii) at what temporal scale this effect occurs.

Methods

This study uses a combination of natural language processing, multi-level general additive modelling, and network analysis to investigate how the hate speech expressions of users on fringe social media platforms vary with time spent on the platform.

Data collection and pre-processing

We analyse all the messages from Gab posted between its formation on 10th August 2016 and the 29th October 2018 when the platform was taken offline following a terror attack committed by one of its users. This dataset contains 33,089,208 messages posted from 259,598 different accounts. This period consists of an amalgamation of data shared by (Zannettou et al., 2018) which covers August 2016 – January 2018 and data from the online repository Pushshift which covers the remainder of this period from January 2018 – October 29th 2018.

To prevent spam or automated 'bot' accounts from influencing of the results, we identified and removed accounts which made >10,000 posts over the entire period (450 accounts). This coarse method may not have removed all automated accounts, but fully detecting automated activity was beyond the scope of this study. Importantly, we included these accounts when making platform-wide calculations on the level of hate present on the platform each day, as this can be affected by inauthentic activity (Gallacher & Heerdink, 2019), but excluded them from all other user-level calculations.

Automatic detection of hate speech in social media posts

Hate speech is defined as messages which express hatred towards a targeted group with the intention to be derogatory, to humiliate, or to insult members of that group. Specifically, a message requires three features to constitute hate speech: the message must be (1) a deliberate attack, (2) directed towards, or about, a specific group of people, and (3) motivated by, or focused on, aspects of the group's identity. (Chapter 2; Davidson, Warmsley, Macy, & Weber, 2017; de Gibert, Perez, García-Pablos, & Cuadros, 2018). This definition is broad enough to contain any targeted group and is not restricted only to specifically protected groups and characteristics (See Chapter 2, for more details).

We identified hate speech in the dataset using a supervised machine learning approach with extrapolates human annotations on a random sample of messages to the entire dataset. The model we used leverages cross-platform datasets of hate speech along with pre-trained contextualised word embedding models (BERT) to encode semantic information along a number of manually engineered syntactic features commonly associated with hate expressions (see details in Chapter 2). These features include the message length, the presence of explicit hate terms, hate symbols, and obscenity, the lexical density, the presence of 'othering' and the sentiment (positive vs negative) of the message. This model is trained on a total of 40,000 messages from Facebook, Twitter, Stormfront, and Gab. The data from these four individual platforms is used to train four platform-specific models, which are then combined into one model using a 'superlearner' approach (the superlearner 1.0 outlined in Chapter 2). These platforms were chosen as they represent a broad range of presentations of hate speech, and range from the mainstream to more fringe areas. By combining data from multiple platforms, this model outperforms a model trained on any one single platform in isolation. This approach retains platform specific hate characteristics however, and therefore aspects of hate expression which are specific to Gab should be retained and detected by the model.

We used this approach to classify all messages in our datasets into 'clean', 'offensive' and 'hate speech' (as defined above) categories and obtained a confidence score for each message. Offensive messages are broadly defined as messages likely or with the intention to cause offence (Davidson et al., 2017). This includes uncivil, rude, inappropriate, or overly disrespectful content. Often this content contains obscenity and profanity, although the presence of these words does not necessarily signal malicious intent or group directed abuse (see Chapter 2 for further details). To minimise the impact of false positives in the subsequent analysis, we re-classified any message with a low confidence score as 'clean', as this is the majority class in the dataset. This approach performed well for Gab data, with an overall accuracy of 90.9% when measured on a representative sample of the Gab messages used in this study. The precision and recall scores for hate speech in particular were 0.75 and 0.83 (see Chapter 2 Supplementary Information (SI) Table 3 for full performance figures and Chapter 2 Figure 5 for the confusion matrix). The lower values on these measures when compared to overall accuracy reflect the fact that hate speech is a minority class on these social media platforms, and the majority of messages do not contain hate. Given this, a precision score of 0.75 (one of four messages flagged as hate speech will be done so incorrectly), represents a reasonable level of performance, and favourable in relation to comparable models (e.g. Davidson et al., 2017; Kennedy et al., 2020).

In addition to the classification for each message, we also extracted and stored a number of semantic and syntactic linguistic features which were used for this classification, so as to compare how the nature of hate speech changes over time. These linguistic features are given in Table 1. Further details of how these features are calculated and their relation to hate speech are given in Chapter 2.

Table 1: The additional semantic and syntactic features extracted for each message to investigate how the nature and presentation of hate messages changed over time.

Feature	***Description and Implementation***
Othering	Detection of the use of two-sided pronouns that contain a distinction between the in-group and out-group in a single message (e.g. your/our, them/us, they/we), calculated using dictionaries of popular personal pronouns for both the in-group and out-group. Binary measure (0/1)
Hate Severity	The average severity score for the hate terms in the message (range 0-100). Calculated using a dictionary approach from a database maintained the Anti-Defamation League (ADL)
Hate Terms	The number of hate terms present, calculated using a dictionary approach from the ADL
Hate Symbols	The number of hate symbols present in the message, calculated using the ADL hate database
Obscenity	The number of swear words, calculated using a dictionary approach
Sentiment	The sentiment (positive / negative polarity) of the message, calculated using the qdap implementation in R. (range -1 to +1)
Lexical Density	Measure of syntactic complexity, calculated as the ratio of content words to total words (range 0-100)
Flesch-Kincaid Reading Ease	A measure of message readability in relation to the projected education level required to understand the message (range 0-100)
Post Length	Count of the number of words in the message
Punctuation	Count of the punctuation symbols in the message
Negation	The number of negation words present in a message, calculated with qdap implementation in R
Pronouns	The number of pronouns included in the message, calculated using a dictionary approach

Target of hate detection

To identify the type of hate speech expressed in the messages, i.e., the group targeted, we used guided topic modelling (Anoop, Asharaf, & Deepak, 2016; Li, Chen, Xing, Sun, & Ma, 2019). This combination of supervised machine learning to detect abuse and unsupervised machine learning to identify the nature of the abuse has proven successful before (Agarwal & Sureka, 2015; Theocharis, Barberá, Fazekas, & Popa, 2019), and can lead to superior performance over training specific supervised classifiers for each individual hate type (Park & Fung, 2017).

Topic modelling

To perform this topic modelling we used a form of semi-supervised machine learning called Guided Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003). This approach treats each message as a mixture of topics, and each topic as a mixture of words from within all of the messages in the dataset. This allows messages to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language (Silge & Robinson, 2017). Different types of hate speech (e.g., anti-Semitism, Islamophobia, homophobia) will form different topics within the hate speech messages and we can use this distinction to identify one type of hate from another. We trained the LDA models on the dataset of hate speech messages from the same four social media platforms (Gab, Facebook, Twitter, and Stormfront) used to train the earlier hate speech detection model to ensure consistency. The steps in this process are shown in SI Figure 1.

Prior to performing the LDA we pre-processed the text to remove stop words, extract unigrams and bigrams, and then used an information retrieval approach to increase the co-occurrence of words within the messages and improve consistency and performance by identifying the most similar messages in the dataset and combining them for model training (see details in SI 1.2). The optimum number of topics for the LDA model was calculated dynamically using a range of model performance criteria (Nikita, 2016) (see SI 1.2 & SI Figure 2). This process identified 11 distinct hate speech topics.

To improve the consistency and distinctiveness of these topics we introduced a subsequent guided step (Anoop et al., 2016; Li et al., 2019), whereby we manually identified the most salient words per topic, and input these as seed words to retrain the model to help anchor the topics towards these areas. These words were then weighted higher in topic creation, to improve both topic-word distributions (by biasing topics to coalesce around appropriate seed words) and to improve document-topic

distributions (by guiding documents towards topics related to the seed words they contain). This has been shown to improve the performance of unsupervised topic modelling by adding subject matter knowledge into the process and by guiding the topic models to converge in meaningful directions (Jagarlamudi, Iii, & Udupa, 2012). For example, when racial slurs are used in place of actual group names, the slur and the group name will not co-occur naturally, but by guiding them manually to the same topic they can be placed together. Equally, words which often co-occur but form different topics may be placed into the same category incorrectly in classic topic modelling. Adding these two words as seeds into different topics can help prevent this from occurring. At this stage we re-ran the LDA for the same number of topics, this time guided towards the identified topics (SI Figure 1 stage 2.3). The final topics that the model produced are shown in in table 2.

Topic model performance

We measured performance of this hate topic classifier by taking a random sample of 100 messages from each topic and manually identifying whether these topic judgements were correct. These accuracy measurements for each topic are shown in table 2. These judgements were independently checked by a second coder. Intercoder reliability scores gave a percentage agreement of 89.4%, a Krippendorff's Alpha of 0.717, and a Cohen's Kappa score of 0.720. The former is at the level of 'good' agreement (George & Mallery, 2003) while the latter is well above the 0.61 threshold for reliable coding and substantial agreement (Glen, 2014), and so we retained our labelling as accurate.

The accuracy in detecting the type of hate speech varied greatly between topics. Anti-Semitism, anti-Black racism, and anti-immigration sentiment were the most accurate topics, with accuracy scores ranging from 90.0% to 81.0% (Table 2). These are high accuracy scores compared to the no-information rate of this classifier of 9.1% (assuming random distribution of these 11 topics). Accuracy in detecting other forms of hate speech, such as hate directed towards Mexicans and Liberals, was lower, although still higher than chance. Based on these results we retained the most accurate 6 classes of topic models: anti-Semitism, anti-Black racism, anti-immigration, Islamophobia, homophobia and misogyny. All other classes were re-classified into the 'other/unknown' category. Figure 1 shows the top 20 terms for each of these six retained topics. These six hate types have been previously highlighted as the most prominent online by a UN report on online hate (Ischinger, 2020), and as the primary groups targeted by the online far-right (Conway, 2018).

DISCLAIMER – Hateful terms included in the figure below

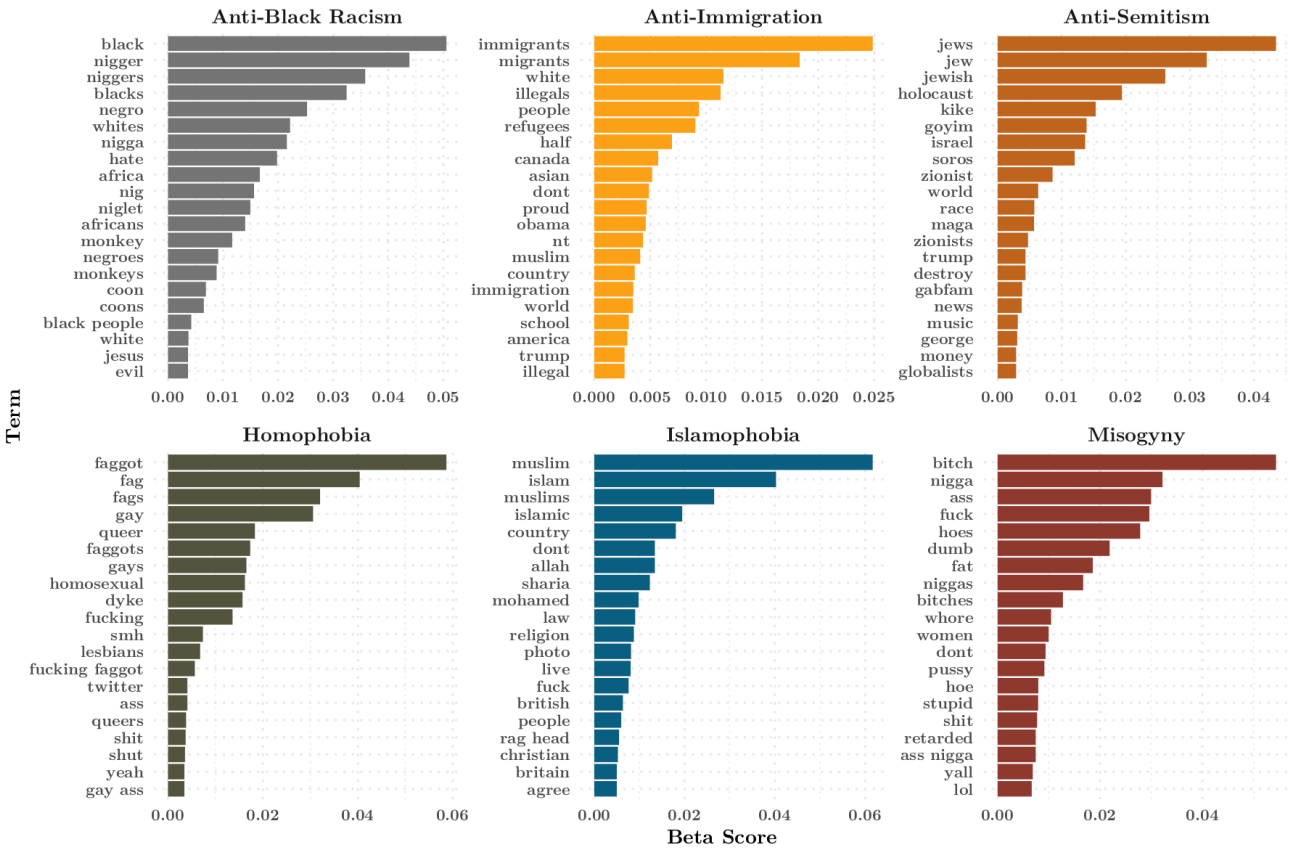


Figure 1 – The top 20 terms across the 6 retained hate topics (anti-Black racism, anti-immigration, anti-Semitism, homophobia, Islamophobia and misogyny) from the guided LDA model. Beta scores are given for each for the terms

Table 2 – Hate topics and accuracy of classification

Topic	***Accuracy (%)***	
Anti-Semitism	**90.0**	↑ Retained
Anti-Black racism	**89.6**	
Anti-Immigration	**81.0**	
Islamophobia	**73.7**	
Homophobia	**72.2**	
Misogyny	**71.8**	
Anti-left hate	62.3	↓ Re-coded to 'other'
White Supremacy	53.0	
Anti-Mexican racism	31.0	
White Trash hate	23.0	
Unknown / Residual Error	NA	

Analysing the dynamics of hate speech expression

To investigate the relationship between time spent on the platform by a user and their hate expressions (Objective 1), we calculated the correlation between users' total number of posts and the proportion of their messages which contained hate speech. To account for periods of inactivity or 'lurking', we also performed this analysis for the overall time they spent active on the platform (number of days between their first and last post). We used Spearman rank correlation as variables were not normally distributed.

Measuring hate trajectories

We further investigated the nature and trajectory of user's hate speech expressions (Objective 1) over time with multi-level general additive models (GAMMs) implemented in the 'mgcv' package (Wood, 2019). GAMMs are a class of statistical models where the relationship between the independent and dependent variables is represented by non-linear smoothing functions to capture non-linearities in relationships between variables. We modelled how the probability that a user expressed a hate message changed over their posting period within the platform (their 'hate trajectory'). We accounted for inter-user differences by including user ID as a random effect. As the overall proportion of hate messages varied greatly between days (SI Figure 3), we controlled for the overall level of hate expression on each day (the proportion of all messages on that specific day which were classified as hate speech) to observe how individual behaviour changed over time rather than the general level of hate speech on the platform.

Only 5.9% of the users posted beyond 250 posts, so to conserve good sample sizes we truncated users posts at the 250th post. Users who posted hate to the platform but made fewer than 250 contributions overall had their full post history included. Users who never posted hate on the platform were excluded from this analysis as it would not be possible to assign them a hate trajectory. Following this filtering we retained 24.7% of posts from the dataset (8,173,034 posts). For robustness, we also ran the same analysis over a longer post history, including either the first 2150 posts from each user or their entre post history, whichever was longest. This 2150 threshold was chosen as 99% of users did not post past this point). The posts included represented 59.7% of overall posts.

Changes in the nature of hate expression

In order to test whether the style of hate expressed by users changed over time as they spent time on the platform, we firstly compared the linguistic properties (metrics detailed in table 1) of their first hate post to the average of the inter quartile range (IQR) of their subsequent hate posts. Comparing the first message to the average of the IQR controls for the different number of hate speech posts each user makes, and for each user we are therefore comparing their first message against their latter messages, regardless of how long or how many messages they contributed to the platform overall. For example, for a user who posted 15 messages in total, we compared their first message to the average of their 4^{th} to 12^{th} messages. Conversely, for a user who posted 100 messages, this would be comparing their first to the average of their 25^{th} to 75^{th} messages.

To get a further measurement of how users' hate expressions changed over time, we also counted the number of different groups targeted by hate speech that users expressed in their early (first 100) and compared this to their subsequent (101^{st} – 200^{th}) messages.

The effects of hate message engagement

To measure the role social feedback plays in these hate speech trajectories, we tested how the number of reactions (engagement) received by a hate speech post affected the likelihood of the user who posted the message posting subsequent hate messages. The number of reactions was measured as the sum of the 'Likes' and 'Dislikes'. We did not differentiate between the nature of these responses as it is difficult to distinguish between endorsements and opposition in these options, and likely it depends on the nature of the message. For completeness, we also compared the average number of responses that hate speech messages received compared to offensive and clean messages. These values were normalised due to highly right-skewed distributions (i.e. many messages receiving few reactions, and a few receiving a very high level of engagement).

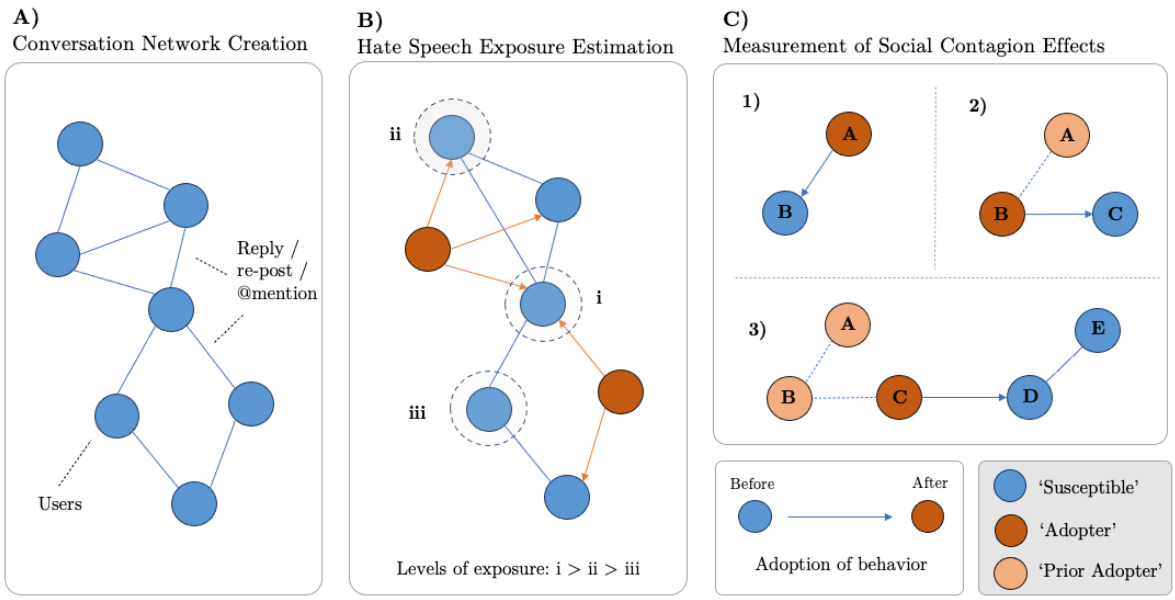


Figure 2 - Creation of the conversation networks and measurement of exposure and contagion effects. Here, adopting the behaviour represents expressing hate speech (dark red), 'susceptible' individuals have not yet expressed hate speech (blue), and 'prior adopter' individuals have but cannot contaminate nor become contaminated again (orange).

Measuring network contagion and exposure to hate speech

In order to measure network contagion effects of hate speech spread (Objective 2), we first pre-processed the data to create daily conversation networks containing dynamic information about the spread of hateful messages between users on that day. For each day in the dataset we identified all users taking part in the conversation, and the interactions between them. These interactions can be replies to others' messages, new messages directed at other users through mentions (@mentions), or reposts of other users' prior messages. We did not differentiate between these types of interaction. We recorded the time each interaction took place. For each user, we then identified the time (hour) within that day when they first posted hate speech, which represents their time of 'adoption' of hateful behaviour for that day. We found that the Gab conversation followed a strong daily rhythm with a peak of activity between 1600hrs and 2000hrs (Eastern time zone, UTC−5) and a minimum at 0700hrs. This supports the separation of the conversation into daily networks, however to minimise conversation split across days, we therefore used 0700hrs as the boundary between two days to separate daily conversations, rather than 0000hrs which is near the peak activity (see details in SI Figure 4).

With this information, we constructed the dynamic conversation network between users, with weights included for multiple connections but without self-loops (Figure 2a). To filter out peripheral activity

and isolated nodes, and ensure consistency and computational efficiency, we performed a k-core reduction with a threshold of k=2. This means that all remaining users have connections to at least two other users on that day.

This network was then made dynamic using the NetdiffuseR package (Valente, Dyal, Chu, Wipfli, & Fujimoto, 2015; Valente & Vega Yon, 2020; Vega Yon, Valente, Dyal, & Hayes, 2020), with each day treated as a different conversation network and with each hour a 'slice' of that daily conversation network. This allowed us to calculate the 'exposure' to hate speech that each user received at each hour. Higher exposure occurs if the users are closer to users in the network when these users express hate speech (Figure 2b). Specifically, the exposure to hate speech was calculated as the proportion of each users' connections who adopted the behaviour (i.e. expressed hate speech) in that time period, taking the weights of the connections into consideration (Valente, 1996). This value is then normalised between 0 and 1, with higher scores reflecting higher levels of exposure. Network exposure is measured on direct contact (rather than structural equivalence) and therefore captures social influence conveyed through overt transmission of information, persuasion or direct pressure (Valente, 1996). Social contagion effects will be observed in these networks if users are more likely to express hate speech after having previously been exposed to it by other users (Figure 2c).

Testing for the effect of network structure – short-term vs longer-term hate contagion

We tested the social contagion effects for hate speech (Figure 2c) at two levels: in the short-term and in the longer term (Objective 3).

At the short-term (hourly) level we looked at hourly network slices (over all days) and tested whether network exposure to hate speech in hour h was associated with a greater likelihood of subsequent hate speech expression during the next hour $h+1$. For this we used a logistic regression model (generalised linear mixed model, GLMM) with user ID and date as random effects to control for differences between users and changes within the platform as a whole over time. We also added hour of the day (1-24) as a fixed effect to control for any effects of certain times of day containing greater hate speech expressions. A caveat of using one-hour slices is that if users are exposed to hate and then post it themselves in the same hour, we will miss this behaviour, however looking at sub 1-hour slices is not feasible due to the computational demands caused by long-term nature of the dataset. On the other

hand, our method means that we are unlikely to detect users simply echoing others' sentiments in the direct replies, as these would likely occur within the same one-hour network slice.

At the longer-term level we tested for relationships between network exposure to hate speech and expressions of hate speech that existed across multiple days. For each user, on each day, we calculated their exposure to hate speech on that day up to the point of posting hate speech themselves (if they didn't post any hate, then their exposure would be calculated over the whole 24h period). Additionally, we calculated the average daily exposure to hate speech that they received over the preceding 7 days. For each day, the response variable was a binary indicator of whether they posted hate that day or not. This longer-term approach also allows us to account for another caveat of the short-term approach. Due to the nature of the diffusion network models within NetdiffuseR, it is not possible to include multiple times of adoption for a behaviour by the same user within the same network, and so we are only measuring exposure effects for the first time a user expresses hate speech on each day. It is therefore possible that we are missing cumulative effects of users expressing hate speech repeatedly within a day, leading to greater exposure of their contacts. Looking at longer-term effects over multiple days allows us to address this issue.

For these long-term models we again used GLMMs and tested simultaneously for the effects of both daily and weekly exposures with user ID and date as random effects, but additionally including the cumulative number of days that users were active up until the day of posting as random effect, to account for any effects of platform use on increasing likelihood of posting hate speech. Model performance for short and longer-term effects was compared using Akaike Information Criterion (AIC). By controlling for users' exposure to hate speech at the short-term level within these longer-term models, we are separating any immediate effects of conversation topic clustering from longer-term exposure. This is important, as clustering (aka homophily; the tendency for similar users to group together e.g. Barberá, 2015; McPherson, Smith-Lovin, & Cook, 2001) and contagion are easily confounded in observational studies of contagion effects (Shalizi & Thomas, 2011). These clustering effects can be assumed to be equal across days, so can be controlled for by controlling for daily conversation exposure.

Secondary transfer effect

To test for transfer effects of hate speech exposure across hate types, we performed a similar analysis at both the short and longer-term level separately for each specific hate type identified by the topic modelling. We calculated each user's exposure to specific types of hate at each network slice, along with their adoptions of these specific hate types. We tested for the predictive power of exposure to each hate type on the adoption of each hate type with similar GLMMs as above, in each model including exposure to all other hate types to control for these effects. To control for the effects of multiple comparisons, Holm's corrections were applied to the obtained p-values.

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Ethics Reference: SSH_OII_CIA_19_062). All data collection was conducted using open source methods and publicly available data, and hence, informed consent was not explicitly obtained. No private groups were joined, and no accounts 'befriended' in order to access data that are not publicly available. In order to preserve anonymity, we took a cryptographic hash of all usernames prior to analysis and real account usernames have not been used in the analysis at any point.

Data analysis

All analysis was done in R (version 3.6.1) using the tidyverse (Wickham et al., 2019) and tidymodels (Kuhn, Wickham, & RStudio, 2020) collections of packages. Where particular additional packages have been used, they have been referenced in the text. Where applicable, summary statistics are presented as mean +/- standard error.

Results

Overall, 8.9% of messages on Gab during the period August 2016 – October 2018 were classified as hate speech, with 28.9% of users expressing hate speech at least once when active on the platform. For users who expressed hate, Islamophobia was the most common type of hate speech (39.4%), followed by anti-immigration sentiment (36.0%) and anti-Semitism (15.5%) (Table 3, Figure 3). Many users shared multiple types of hate, with 18.2% of hate expressing users sharing a combination of Islamophobia and anti-immigration sentiment, while 7.9% of these users shared a combination of Islamophobia, anti-Semitism, and anti-immigration sentiment and 10.3% shared all six types of hate.

Together these six hate types formed 61.4% of the overall hate speech within Gab over the study period, with the other hate types (plus unidentifiable hate messages) making up the remaining 38.6% of hate speech messages.

Users' hate speech expression over time

We found a positive relationship between users' level of activity on Gab and the proportion of hate speech in their posts. Both users' total number of posts and the duration of their active period on the

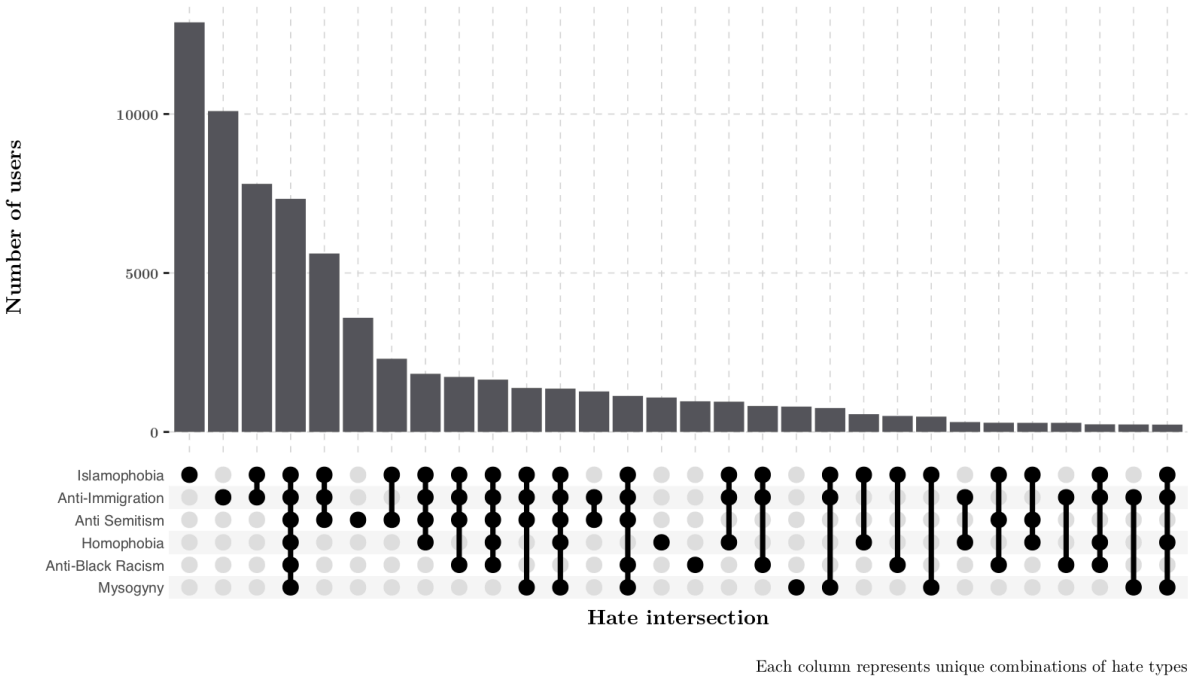


Figure 3 – The frequency of common intersections of types of hate speech expression across hate types. Columns show the number of users who uniquely shared this combination of hate types, with the intersections given below.

platform were positively correlated with the proportion of their posts expressing hate (Spearman's Rho, number of posts: $r_s = 0.562$, $p < 0.001$, active period: $r_s = 0.864$, $p < 0.001$)

When investigating users' trajectories of hate speech expression, we found a high likelihood of users posting hate early, with 10.8% of users expressing hate in their first post upon joining the platform. This is higher than expected by chance if all hate posts were evenly distributed (8.9%). The number of subsequent posts had a large impact on the likelihood of hate expression, and followed a non-linear relationship over the subsequent 250 posts (GAMM, n = 8,173,034, df = 7.259, F = 5.061, $p < 0.001$). Initially, the likelihood of expressing hate speech decreased as they started to take part in the conversations, but this trend later reversed, and the likelihood of hate expression increased with every subsequent posted message. The turning point of this trajectory occurred around the 25th post (Figure 4). Over a longer period of up to 2150 posts on the likelihood of a user posting hate speech increased until around the 1000th post before levelling off (SI Figure 5a, GAMM, n = 18,159,111, df = 1.533, F = 153.3, $p < 0.001$).

To control for the possible artefact of more moderate users leaving the platform early, and driving the latter stages of these user hate trajectories, we additionally (i) calculated 'hate trajectories' over the log of user post numbers to minimise the impact of fewer users being active over longer periods and

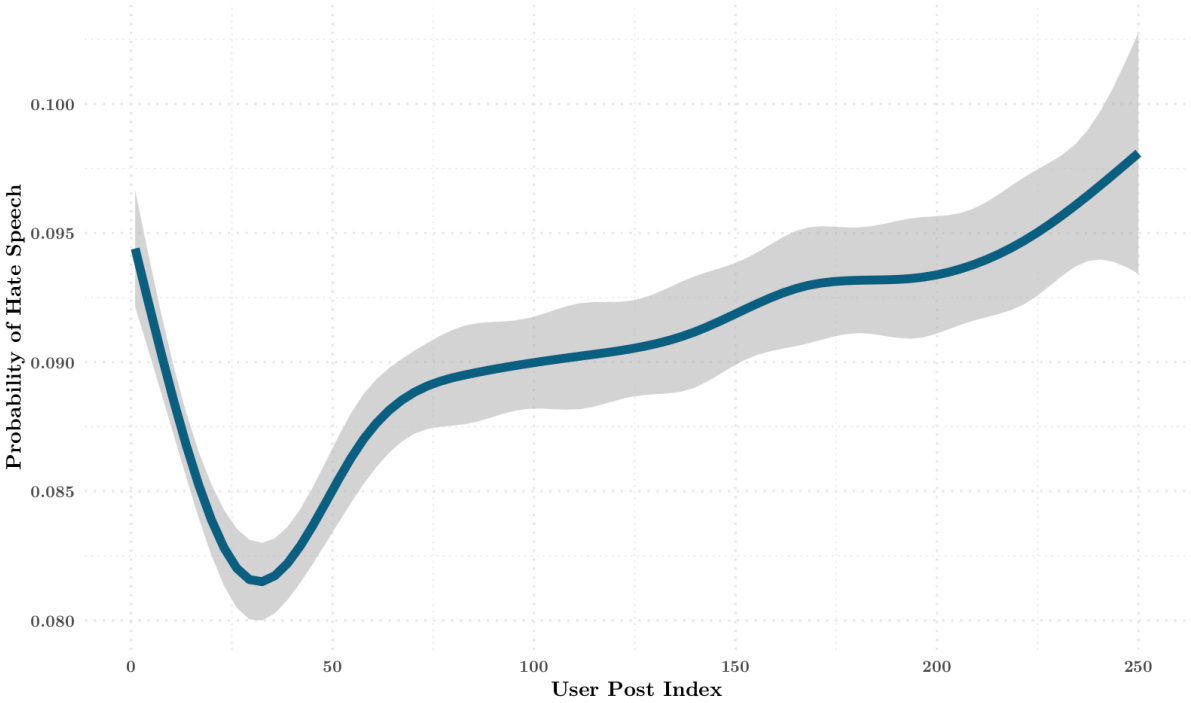


Figure 4 - Overall user hate trajectories (probability of posting hate for each consecutive message to the platform) for Gab for the first 250 posts users made to the platforms.

Table 2 – Differences between initial hate speech messages and later (average of Interquartile Range, IQR) hate speech messages for users within Gab (means +/- se)

Metric	First hate post	se (+/-)	Later hate posts (IQR)	se (+/-)	Difference
Daily Hate Density (0-1)	0.084	0.000	0.091	0.001	-0.007
Post Length (words)	34.530	0.064	34.57	0.401	-0.042
Hate Severity (0-100)	9.530	0.110	8.041	0.585	+1.489
No of Hate Symbols	0.023	0.001	0.028	0.005	+0.005
No of Hate Terms	0.171	0.002	0.154	0.012	+0.017
Lexical Density (0-100)	40.160	0.056	40.36	0.416	-0.203
No of Negation terms	1.112	0.005	1.109	0.035	+0.003
No of Reactions	33.163	0.364	32.55	2.305	+0.614
No of Obscenity terms	0.467	0.004	0.390	0.022	+0.076
Othering (0-1)	0.134	0.001	0.121	0.007	+0.013
Polarity (-1 : 1)	-0.193	0.002	-0.160	0.004	-0.034
No of Pronouns	3.553	0.012	3.565	0.075	-0.012
Readability (0-100)	61.933	0.085	63.61	0.491	-1.680

(ii) looked at the hate trajectories for the most active users (with ≥ 250 posts) only. The results from these robustness checks showed that trajectory was not caused by moderate users leaving the conversation; looking at the hate trajectories over the log of user post numbers gave a similar u-shaped trajectory (SI Figure 5b). The hate trajectories for just those users who posted >250 messages followed an upwards trajectory over time (SI Figure 6a&b).

Difference between early and late posts

On Gab, users' initial hate messages differed both in style and substance from later messages (Table 2). Early messages had higher levels of obscenity, more explicit and severe hate terms, lower lexical density, a higher level of 'othering', and held a more negative sentiment. These messages also contained fewer hate symbols. In other words, early hate posts seemed more explicit and less nuanced. They also received more reactions from other users than later hate posts. Initial and later hate posts were posted on days with similar overall hate level on the platform.

Users also increased the breadth of their hate targets over time from early to later hate messages. Within users' initial 100 messages, on average they expressed hate towards 3.2 +/- 0.01 target groups, and this increased to 5.2 +/- 0.02 for the subsequent 100 messages. By comparison, an average user's initial ten messages contained hate messages targeted towards 1.6 +/- 0.001 different groups.

Engagement with hate messages

The level of engagement a post received varied depending on the level of abuse it contained. 'Hate speech' messages received the greatest engagement, followed by 'offensive' then 'clean' messages (Mean normalised reactions: clean = 0.88 +/- 0.0002, offensive = 0.90 +/- 0.0005, hate speech = 1.1 +/- 0.0005, Figure 5a). The level of engagement a hate post received also marginally increased the likelihood of the user posting subsequent hate messages. Users who received no reactions to their hate messages had a 90.8% likelihood of posting another hate message, and this increased to 95.2% when a hate post received 10 reactions (Figure 5b).

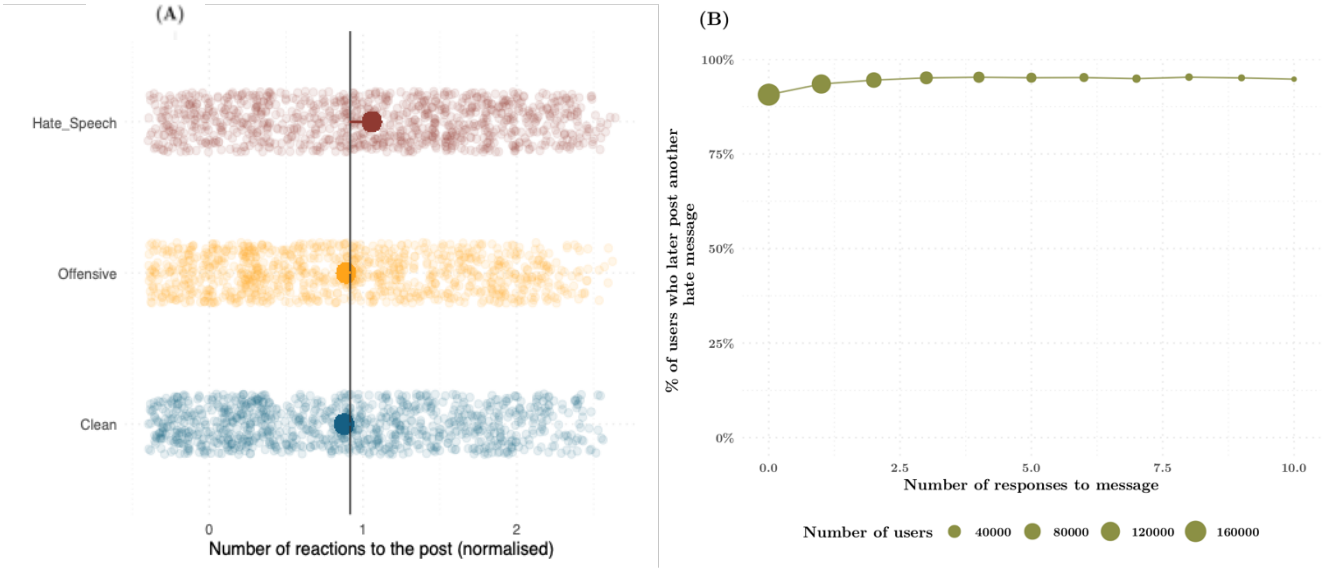


Figure 5 – A) Number of reactions to clean, offensive and hate speech messages on Gab. Circles show mean number of reactions per type, the black line shows mean reactions across all types. B) The probability that a hate speech message will be followed by another for varying levels of reactions (averaged across all users). Dots sized by the number of users who posted that many hate speech messages.

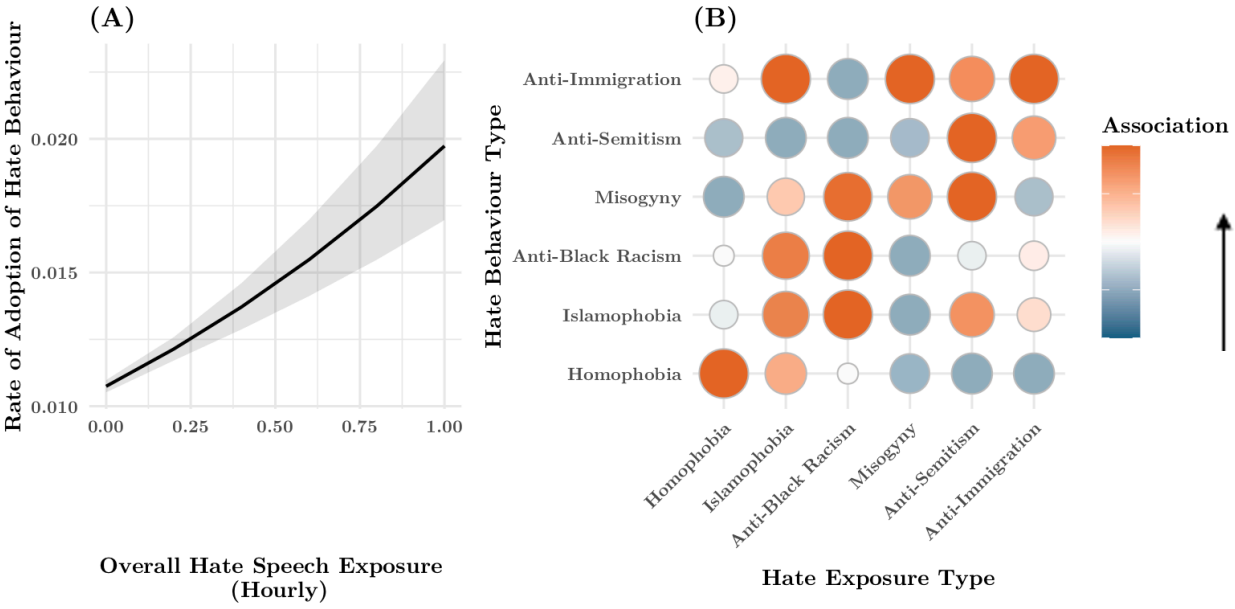


Figure 6 – A) Overall effect of exposure to hate speech on adoption of this behaviour (the grey area represents the 95% CI, the large CI towards high hate speech exposure is caused by low sample sizes of users exposed to such levels of hate). B) Effect sizes for exposure to hate types on adoption of these hate behaviours. Colour and point size indicate strength of association

Social contagion of online hate speech

We observed a strong overall effect of social contagion of hate speech at the short-term (hourly) and longer-term (daily/weekly) levels.

At the short-term (hourly) level, exposure to hate speech in a given hour increased the likelihood that a user would express hate speech themselves in the following hour (GLMM, n = 11,034,480, Estimate = 0.61 ± 0.08, Wald's Z = 7.73, $p < 0.001$, Figure 6a, SI Figure 8&9). Each 10% increase in exposure to hate speech increased the odds of a user expressing hate themselves in the subsequent hour by 8.5%, while a user with 100% of their network connections expressing hate speech was 85% more likely to adopt this behaviour than a user with zero connections expressing hate speech.

To test for longer-term social contagion effects of exposure to hate speech we measured a user's total daily exposure to hate and their average exposure over the prior seven days. Results showed that both of these types of exposure influenced future hate expression (Figure 7a&b, GLMM, n = 479,760, Daily exposure: Estimate = 1.42 ± 0.48, Wald's Z = 2.96, $p = 0.003$. Weekly exposure: Estimate = 0.47 ± 0.05, Wald's Z = 9.41, $p < 0.001$). Comparing these two models, results indicate that longer-term exposure to hate speech over the previous seven days had a larger impact on whether a user

Table 3 - Statistical effects of exposure to hate types on the likelihood to express hate speech of this or another type (short-term model). The blue bars show the relative proportion of each hate type. Significant relationships are highlighted in yellow.

Hate Speech Exposure Type

OBSERVED BEHAVIOUR	Anti-semitism			Islamophobia			Misogyny			Anti Black Racism			Anti Immigration			Homophobia			RELATIVE PROPROTION
	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	
Anti-Semitism	**1.9**	**11.4**	**0.000**	**1.2**	**6.9**	**0.000**	1.1	2.1	0.168	1.0	1.9	0.056	**1.7**	**8.0**	**0.000**	1.1	2.133	0.165	
Islamophobia	**1.3**	**7.5**	**0.000**	**1.2**	**9.5**	**0.000**	0.1	0.2	1.000	**1.6**	**4.4**	**0.000**	**0.8**	**3.5**	**0.002**	0.5	1.039	1.000	
Misogyny	**2.2**	**6.5**	**0.000**	**1.3**	**3.3**	**0.001**	1.8	1.7	0.354	**2.2**	**2.5**	**0.042**	0.4	0.4	0.675	0.1	−0.002	1.000	
Anti-Black Racism	0.7	1.0	0.309	**1.7**	**5.4**	**0.000**	0.1	−0.0	1.000	**2.2**	**2.6**	**0.034**	1.0	1.5	0.396	0.8	0.560	1.000	
Anti-immigration	**1.4**	**7.5**	**0.000**	**1.5**	**9.4**	**0.000**	**1.9**	**6.0**	**0.000**	**−1.4**	**−3.4**	**0.004**	**1.9**	**8.6**	**0.000**	−0.0	−0.134	1.000	
Homophobia	0.9	1.6	0.233	**2.0**	**6.6**	**0.000**	0.9	1.3	0.625	1.5	2.2	0.054	1.0	1.4	0.396	**2.7**	**4.657**	**0.000**	

expressed hate speech themselves compared to daily exposure (ΔAIC between the two GLMMs = -85.77).

Interestingly, within these models, the number of days a user had been active prior to the focal day was also a significant predictor of hate speech expression, even when accounting for prior levels of exposure to hate speech in both the daily and weekly models. (GLMM, Weekly Model; Days active: Estimate = 0.002 ± 0.001, Wald's Z = 2.26, p = 0.024. Daily Model; Days active: Estimate = 0.002 ± 0.001, Wald's Z = 2.06, p = 0.039)

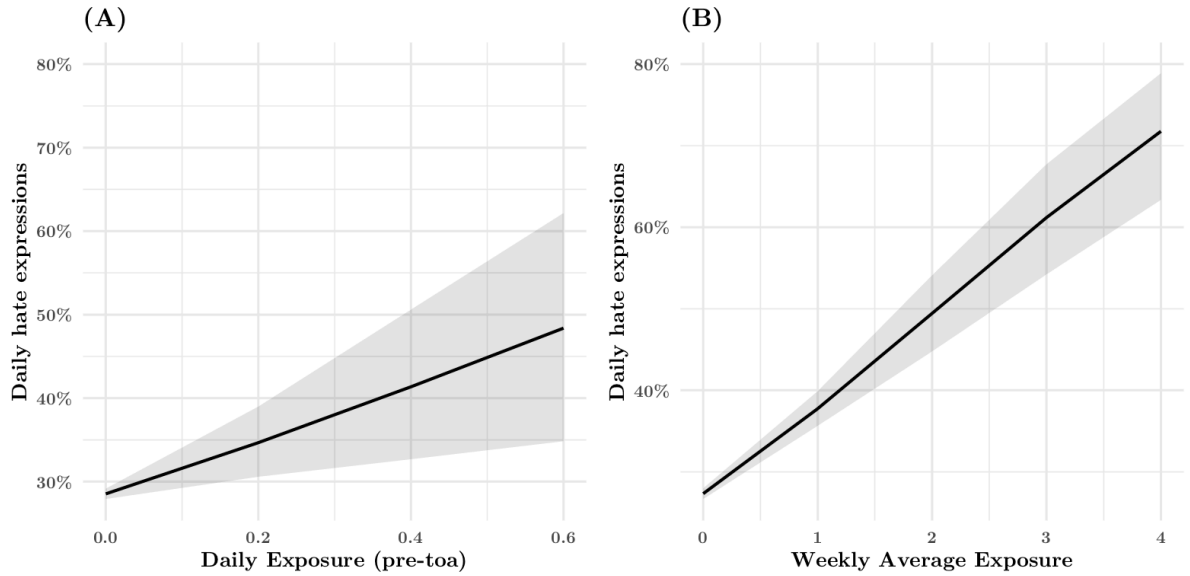


Figure 7 - A) Effect of short-term exposure to hate types on adoption of these hate behaviours while controlling for long-term exposure. B) Long-term effect of exposure to hate speech on adoption of this behaviour controlling for short-term exposure.

Table 4 - Statistical effects of 7-Day exposure to hate types on the likelihood to express hate speech of this or another type (longer-term model). Significant relationships are highlighted in yellow.

Hate Speech Exposure Type

BEHAVIOUR (HATE TYPE)	Anti-Semitism			Islamophobia			Misogyny			Anti Black Racism			Anti Immigration			Homophobia		
	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P
7 Day Average																		
Anti-Semitism	**1.30**	**8.00**	**0.00**	**0.71**	**5.86**	**0.00**	0.26	0.50	1.00	−0.09	−0.06	1.00	**0.55**	**2.91**	**0.03**	0.75	0.41	1.00
Islamophobia	**0.71**	**5.63**	**0.00**	**0.86**	**9.38**	**0.00**	0.45	1.28	0.40	0.60	1.95	0.20	**0.76**	**5.30**	**0.00**	**1.35**	**4.20**	**0.00**
Misogyny	0.40	1.08	1.00	0.41	1.42	0.94	1.17	1.14	1.00	−1.75	−1.18	1.00	0.91	1.69	0.83	1.31	0.96	1.00
Anti-Black Racism	0.14	0.33	1.00	0.10	0.36	1.00	0.59	0.62	1.00	**3.23**	**5.65**	**0.00**	0.38	0.80	1.00	−2.82	−1.75	0.72
Anti-Immigration	**0.61**	**3.36**	**0.01**	**0.56**	**4.38**	**0.00**	0.28	0.52	1.00	−0.10	−0.16	1.00	0.50	2.39	0.13	0.65	1.22	1.00
Homophobia	1.07	2.42	0.15	0.18	0.58	1.00	−0.24	−0.20	1.00	−3.56	−1.95	0.41	0.62	1.26	1.00	−0.05	−0.05	1.00

Transfer effects of exposure to hate speech

We tested whether this social contagion effect was specific to the type of hate speech exposed to, or whether there were secondary transfer effects across hate types. At the short-term (hourly) level, results indicated that while the strongest effects were for matching hate targets, there were also significant transitive effects across hate types, i.e., users exposed to one type of hate speech then went on to express another type of hate, at a rate above chance (Figure 6b, Table 3). This was particularly pronounced for users exposed to anti-Semitism, Islamophobia, and anti-Black racism. The effects of this transfer of hate appear strongest between targeted groups which have overlapping traits, such as Islamophobia and anti-immigration. Many combinations of hate targets did not transfer however (Table 3), suggesting that there is also a degree of specificity to these effects.

We found similar patterns at the daily/weekly (Table 4), with anti-Semitism and Islamophobia having the greatest transitive effects, followed by anti-immigration sentiment. The effect of exposure to anti-Black racism, misogyny and homophobia appeared less transferable and therefore more specialised. (See SI Table 2 for short-term transfer effects when controlling for these long-term trends).

Discussion

This study investigates how hate develops on fringe social media platforms and how user's hate speech expression relates to their participation in the social media platform and to the exposure to hate from they received from other users. We found that hate speech expressions increased with time spent on the platform, although a number of users started posting hate from their first post, and that higher exposure to hate speech increased the likelihood that users expressed hate speech themselves, over both short and longer timeframes. This effect was largest for the specific type of hate that users have been exposed to, but also transitive across hate types. Overall, our results (i) provide evidence socialisation effects on hate speech expression on social media platforms, without excluding that some users join platforms with pre-existing prejudice, (ii) reveal that hate speech on social media platform spreads through social contagion, both in the short and longer term, and (iii) highlight the existence of transfer effects across hate types following exposure to hate speech from other ingroup members.

We found a substantial quantity of hate speech on Gab, in line with previous research (Mathew et al., 2019; Zannettou et al., 2018), which highlights the damage these spaces may do to group relations and the promotion of violence (Tajfel, 1974; Williams, Burnap, Javed, Liu, & Ozalp, 2019). In agreement with previous research on other fringe platforms (Finkelstein, Zannettou, Bradlyn, & Blackburn, 2020; Kleinberg et al., 2020; Ribeiro et al., 2020), we also find that the overall density of hate speech on Gab has increased over time, stressing the importance on understanding the impact that this hate has on users and the wider community.

Socialisation effects on hate speech expression

At the platform level, we found evidence that socialisation affects hate expression through interactions with other group members. The proportion of a user's posts containing hate speech increased with the number of posts they contributed to the platform. This relationship was also true for the number of days users spent on the platform regardless of their posting activity, suggesting an effect of observation (or 'lurking') on driving hate speech expression even without active participation in the conversations. This supports suggestions that fringe platforms such as Gab play a role in connecting more moderate right-leaning users to extreme far-right ideologies (McSwiney & Jasser, 2020).

Investigating hate trajectories at the user level allowed us to better understand how platform usage affects hate speech expression, and to test whether users are drawn to these platforms to express pre-

existing hate, or whether this hate develops with presence on the platforms. We find evidence for both effects playing a role. There was a consistent 'U-shaped' pattern in users' hate trajectories, indicating that the likelihood of hate expression was high for the first few messages, before decreasing then rising again as users spent longer on the platform. Like our platform-wide results, this latter increase suggests that users become more extreme in their expressions as they spend longer on the platform, taking on the group norms through a process of socialisation (Smith et al., 2019). This supports evidence that social media use is associated with outgroup prejudice and affective polarisation between mainstream groups (Settle, 2018), and goes a step further by showing that this may also hold for extreme negative and hostile outgroup attitudes. The upwards trajectory in hate speech expression occurred even when taking into account the overall density of hate messages posted on the days the users were active, suggesting that this effect is not simply driven by immediate group conversations or topics, but by a deeper internalisation of the hostile sentiment. Overall, this supports our hypothesis that users who spend time on hateful fringe social media platforms increase the hateful nature of their own messages, likely through socialisation effects.

To further investigate these apparent socialisation effects on hate speech expression, we looked at the number of responses, and therefore engagement / attention, that different types of message received. Hate speech messages received the greatest engagement, compared to offensive or clean message. This aligns with previous research showing that participating in online hate can increase a user's online visibility and subsequently increases their influence on the platform as a whole (ElSherief, Nilizadeh, Nguyen, Vigna, & Belding, 2018). This result also supports the idea that hate speech occurs through group socialisation, because over time influential users will start to shape not just their own expressions but also drive the extremity of the group as a whole, which leads to a reciprocal relationship between the individual's and the group's extremity (Levine & Moreland, 1994). We did not find a large effect of engagement to hate speech encouraging the users to post hate again, and the likelihood of posting hate messages remained fairly constant regardless of reactions, but this may have been due to a ceiling effect caused by the long 2-year period of observation with most hate messages followed by another regardless of social feedback over this period.

Evidence for pre-existing prejudice

The high starting position and initial drop in hate speech expression we found in users' hate speech trajectories points towards another effect for early hate postings however, and suggests that

socialisation with other users may not be the sole cause of hate expression. As such, there may be differential explanations and causes for early vs later hate speech expressions. This is supported by the fact that the first hate message on Gab was posted after just 46 other posts and 1.5 days after the platform was created. These early hate posts are very unlikely to result from group socialisation and therefore indicate pre-established hate must have pre-dated the platform.

There are three possible causes for these high early hate levels. Firstly, users may deliberately join these platforms to express hate speech and seek out similar extreme content for themselves (Pauwels & Schils, 2016). Alternatively, users may over-express hate speech in their initial posts to 'prove' group membership, loyalty, and to gain acceptance by other in-group members (Doosje et al., 2016; Harel, Jameson, & Maoz, 2020). Offline evidence suggests that peripheral members of a desirable ingroup derogate an outgroup in public to enhance status, even if they do not hold this attitude in private (Noel, Wann, & Branscombe, 1995). Finally, there may be a period of 'observation' or 'lurking' before users post a first message, during which they may build up an affective state in relation to the outgroup through the influence of the emotional expressions of others, and only post a message when this reaches a certain threshold (Brady, Wills, Jost, Tucker, & Van Bavel, 2017a; Van Kleef, 2009; Van Kleef, Van Den Berg, & Heerdink, 2015).

In the current study it is not possible to definitively test for which of these explanations is likely to be correct. However, several of our findings can help further indicate that users' initial hate posts may be driven by pre-existing opinions which they have come to the platform to present, but over time the nature of their hate expressions both increases and evolves to adopt the norms of the online community as a whole. Users' initial hate posts were more explicit and less nuanced compared to later posts; they contained a higher proportion of explicit hate terms, high obscenity, lower sentiment, and fewer obfuscated hate symbols. This lack of subtlety, together with the fact that ingroup specific language only occurred in later hate posts, suggest that the latter develops by spending time on the platform, supporting the idea that differential causes underpin early and late hate postings (Tuters & Hagen, 2019). Similarly, the number of outgroups that users targeted with hate increased over time spent in the platform, while users initial hate targets were more focused on a specific group. This supports ideas that prejudice or 'fixation' against a certain group may drive users onto these platforms (Cohen, Johansson, Kaati, & Mork, 2014; Grover & Mark, 2019), but this prejudice may expand to other groups through interactions on the platform. Finally, the hate trajectories of the most active users did not show this u-shaped trajectory, and instead we observed a more consistent overall

upwards pattern in hate speech expressions. This indicates that the initial drop may be driven by users who do not remain active in the platform over the longer term, suggesting that this drop might be caused by 'venting behaviour' after which the users are satiated and leave the conversation. Similar patterns, with many users quickly leaving the platform, have been observed on far-right extremist forums (Scrivens, Wojciechowski, & Frank, 2020). The overall hate speech on the platform was not higher for users' initial hate post, suggesting that affective triggers or 'outrage' (Brady et al., 2020; Brady, Wills, Jost, Tucker, & Van Bavel, 2017) may not be solely driving initial hate posts–users are less likely to have seen something which makes them respond with hostility if the overall platform is no more hate-filled.

Social contagion effects on hate speech expression

Another key finding of this study is that social contagion appears to play a role in hate speech expression. Indeed, users who experienced greater exposure to hate speech from other users they interacted with on the platform were more likely to subsequently express hate speech themselves. Our results support prior work which shows that emotional states can spread between users on social media (Bond et al., 2012; Brady et al., 2017b; Kramer et al., 2014; Terizi et al., 2020), and demonstrate that such effects are also applicable to more extreme prejudicial behaviours. This raises concerns about the negative impacts of the online environment on intergroup relations and conflict.

This effect of exposure on hate expressions occurred in both the short- (hourly) and longer-term (daily/weekly) but was larger following longer-term exposure, suggesting that these effects may accumulate over time and reflect longer term changes in attitude rather than reactionary responses to the immediate conversations that users are a part of. Interestingly, in these longer-term models we observed that the total duration of user activity on the platform increased the likelihood of expressing hate speech, regardless of the exposure patterns they experienced. This is in accordance with our earlier results that suggest time spend on the platform (in days) was correlated with proportion of hate speech messages, and suggests an effect of passive consumption of hate, in addition to the measured activity of interactions with other users. This supports prior evidence that within social media users are influenced as much by the content that they passively consume as that which they interact with (Settle, 2018), and that these even users who never contribute to conversations consider themselves a member of the online community (Nonnecke, 2000). It is not possible for us to measure this passive consumption from publicly available data, but in future exploring this may give valuable insight into how hate speech spreads online even amongst those active but not participatory.

The social contagion effect of hate speech we observe suggests that conformity effects are influencing users' propensity to express hate speech themselves. In other words, if a greater proportion of their direct connections are themselves expressing hate, then they themselves are also more likely to do so. This supports traditional ideas about the role of conformity in prejudice (Allport, 1954). The Internet, and social media in particular, has been implicated in skewing conformity pressures by giving the impression of false consensus on extreme positions (Pariser, 2011). Those with the most extreme ideological positions are the most active (Barberá & Rivero, 2015; Preoţiuc-Pietro, Liu, Hopkins, & Ungar, 2017), receive the most engagement (as discussed above), and this may be accentuated on fringe platforms due to self-selection bias (Robbinson, Schlegel, Janin, & Deverell, 2020). Together, this means that conformity along the lines of the majority behaviour of social media connections may be driving users to more extreme online social norms than exist in reality. This is likely to play a role in exacerbating intergroup conflict as groups adopt increasingly extreme positions (Reicher et al., 2008; Smith et al., 2019). Conformity effects are especially pronounced in far-right groups and in those scoring higher in right-wing authoritarianism (Altemeyer, 1981), which can result in the acceptance of political ideologies that demean outgroups. In addition, influential 'extremist' leaders who deviate from the average group position have a disproportionate impact on the overall group position (Sherif, 1936). Given that in online environments these leaders will form central nodes in the networks, they will also be responsible for much of the hate speech exposure patterns (Barberá et al., 2015; Susarla, Oh, & Tan, 2012), and therefore adoption patterns, of other users (Haslam, Reicher, & Bavel, 2019).

Transfer effects for hate contagion

Another important result of our study relates to the transitive effects of hate speech contagion. Users were most likely to express the same type of hate to which they had been exposed, which supports evidence that group members learn their negative outgroup attitudes from other ingroup members (Bracegirdle, 2020). However, we also found substantial transfer effects, at both the short-term hourly conversation level, and the longer-term daily/weekly level. This indicates that group members' negative outgroup attitudes may propagate across outgroups in a similar fashion to the transfer of positive attitudes in cases of positive intergroup contact (Pettigrew, 2009; Schmid et al., 2014; Tausch et al., 2010). This has important implications for group extremism and intergroup conflict, because as the number of outgroups targeted with hate increases, the negative contact opportunities increase and the likelihood of resolution decreases. There is evidence that exposure to hate speech online can lead

to an effect of 'nebulous othering', whereby groups construct an increasingly vaguely defined 'other' to whom prejudice and intolerance is targeted (Tuters & Hagen, 2019). Ultimately, this transfer effect of hate speech across groups could expand to encompass all those not part of the ingroup.

Transfer effects were largest for types of hate which could be viewed as similar (Islamophobia and anti-immigration sentiment), but also crossed what might be seen as wider group boundaries (homophobia and Islamophobia / misogyny and anti-immigration for example). This may be in part due to the way that certain hate types are presented. For instance, much anti-immigration sentiment is presented by the far-right as being in defence of women and preventing foreign 'defilement' of Western women by 'invaders' (Davey & Ebner, 2017; Ekman, 2019; Olteanu, Castillo, Boy, & Varshney, 2018). Misogyny could then easily derive from this stance (Sarrasin, Fasel, Green, & Helbling, 2015). This combination and conflation of hate targets may partly explain the transfer of hate from one target group to another. However, our approach precludes us from exploring this because we assigns one type of hate to each hate message. Future research should explore this phenomenon in greater detail by including more detailed and nuanced hate type distinctions and investigate whether this is a true transfer across distinct outgroups or instead a 'step up' in abstraction where the outgroup encompasses wider distinctions (all foreigners for example).

Limitations and future research

When discussing the propensity to for users to express hate on social media is important to acknowledge that the majority of users do not exhibit this behaviour, even on fringe platforms. For example, in this study 71.1% of users never expressed hate speech. The reasons for this are unclear, and we did not focus on these users, however further investigating these non-hateful users may be vital for discovering and developing mitigation strategies for the spread of hate speech online. Similarly, in this work we rely on the measurement of interactions with the platform, which leaves a digital trace, rather than consumption of content which does not. Users are influenced as much by the content that they passively consume (Settle, 2018), and while recent work has made progress in building estimates of users' consumption behaviours on Twitter (Dunn et al., 2020), this needs validating and expanding to other social media platforms. Future work should investigate how much of user's outgroup attitudes can be affected by the content which they passively consume from other ingroup members, compared to active interactions. This is especially important given that the production of content is highly unequal, with a few users creating most of the content (Nielsen, 2006).

This is also true for hate speech on these fringe platforms (SI Figure 7), and future work should therefore look to explore the impact on this majority of users who consume social media content rather than create it.

In this study we have also made a number of assumptions in the creation of the contagion networks which should be tested further in future work. Notably, while we have endeavoured to measure 'exposure' to online hate speech through network connections, we cannot be sure that this network measure relates directly to the content that users actually see on screen or attend to if it is presented. The users may simply not be online when their network connections post hate, or they may not read it even if they are online, however it is not possible to measure this from social media trace data alone. In aggregate we can assume a positive relationship between the rate at which connections post hate and the amount of hate users sees when active on the platform, but for individual users this may not hold. Future work should investigate this relationship between network exposure and the level of hate speech exposure experienced. Additionally, in order to implement the diffusion network models we made the assumption that once an individual has adopted the behaviour they remain in this state, and do not reverse or indeed get 'infected' again. This is an oversimplification of real-world behaviours, and in reality, users can express hate speech multiple times, and indeed change between hateful and tolerant states quickly and easily (Rajadesingan et al., 2020). We address this, in part, by looking at multiple days and allowing users to adopt different positions on different days– but differential effects may still occur for multiple 'adoptions' of hate behaviour within a single day. These assumptions, and the effect of user's multiple adoptions of hate or indeed retractions of hate on the behaviour of those around them, should be tested. In a similar way, the relationship between exposure and likelihood of adoption of behaviour within online social networks is likely to follow a complex rather than simple relationship; users adoption patterns will depend on the numbers of adopters within a user's social vicinity, with the adoption probability increasing slowly for low numbers of unique exposure sources, and then quickly when the number of exposure sources reaches a threshold level (Mønsted, Sapieżyński, Ferrara, & Lehmann, 2017). This means that non-linear adoption patterns may occur. Future work should investigate this threshold effect and identify if thresholds are constant across users. If so, then mitigation strategies which can ensure that exposure remains below this threshold might be a useful 'soft' mitigation strategy.

Future research should also make attempts to analyse multiple social media platforms and consider the interplay between platforms and in particular the changes caused by moderation efforts and bans

from mainstream platforms. Hate speech is not restricted to a single social media platform or online group and the hate ecosystem is multi-dimensional, and these wider effects should be taken into account (Johnson et al., 2019). This is especially important as mainstream social media platforms enforce more stringent content moderation policies and users migrate to smaller platforms where exposure patterns to hate speech may be more concentrated (Urman & Katz, 2020). Additionally, in this work we also do not distinguish group membership of those posting hate speech, and therefore there is a chance that some of the hate is directed from out-group members at the platform or ingroup itself. In this way some initial hate posts could be a 'trolling effect' where users turn up to post hate 'about' the group and then leave (Graham, 2019; Honeycutt, 2006). Future work should consider both intergroup and intragroup communication effects on hate speech expression.

Finally, an important aspect to consider in future is the relationship between online hate speech to the offline world. It is unlikely that the Internet alone is the cause of extremism, intergroup conflict, and hate (Meleagrou-Hitchens & Kaderbhai, 2016), and offline relationships and events are key in the development of extreme opinions and behaviours (Koehler, 2014). There is evidence that online hate speech can spike in the days and weeks following offline violence such as terror attacks and mass shootings (Rowe & Saif, 2016; Theocharis et al., 2019; Zannettou et al., 2020; Burnap & Williams 2014; Williams & Burnap, 2016). How hate speech changes following such events, whether it is specific to the type of event (e.g. Islamophobia following a Islamic terror attack) or more generalised, and whether the relationship also holds in reverse where hate speech against a certain group increases prior to offline violence against this group, should be investigated. This is particularly important given evidence that interactions between ingroup members can invoke radical intentions if external prompts for legitimising beliefs are present (Thomas et al., 2014). In this way, offline violence such as terror attacks may act as the legitimising external prompt, leading to a shared group belief in the perceived grievance and an increase in hate speech towards a target group. This should be tested in future work.

Conclusion

This work provides new insight into the drivers of hate speech expression online, highlighting the role of group socialisation and social contagion. By participating in the conversations on fringe social media platforms popular with the far-right, individuals become increasingly likely to express hate themselves. This effect of exposure to hate speech is not limited to one type of hate and appears to transfer across hate targets. These results are likely widely applicable to other platforms and types of

extremism, and have important implications for our understanding of hostile intergroup relations and the role of the Internet in intergroup conflict and hate. As fringe social media platforms become more popular with far-right extremist groups, these patterns of hate exposure will increase, potentially driving dangerous trends in hate towards outgroups.

References

Agarwal, S., & Sureka, A. (2015). Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. *ArXiv*, 1–18. Retrieved from http://arxiv.org/abs/1511.06858

All-Party Parliamentary Group (APPG) on Hate Crime. (2019). *How do we build community cohesion when hate crime is on the rise?* Retrieved from www.appghatecrime.org

Allport, G. (1954). The nature of prejudice. Addison-Wesley Publishing Company.

Alorainy, W., Burnap, P., Liu, H., & Williams, M. (2018). The enemy among us: Detecting hate speech with threats based "othering" language embeddings. *ACM Transactions on the Web*, *9*(4), 1–26. Retrieved from http://arxiv.org/abs/1801.07495

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba Press.

Anoop, V. S., Asharaf, S., & Deepak, P. (2016). Unsupervised concept hierarchy learning: A topic modeling guided approach. *Procedia Computer Science*, *89*, 386–394. https://doi.org/10.1016/j.procs.2016.06.086

Anti-Defamation League. (2019). When Twitter bans extremists, GAB puts out the welcome mat. Retrieved March 20, 2019, from https://www.adl.org/blog/when-twitter-bans-extremists-gab-puts-out-the-welcome-mat

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. *Advances in Knowledge Discovery and Data Mining*, 1–138. Retrieved from http://hdl.handle.net/10722/26088

Asch, S. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35. https://doi.org/10.1038/scientificamerican1155-31

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, 519–528. https://doi.org/10.1145/2187836.2187907

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, *23*(1), 76–91. https://doi.org/10.1093/pan/mpu011

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The critical periphery in the growth of social protests. *PLoS ONE*, *10*(11), 1–15. https://doi.org/10.1371/journal.pone.0143611

Benson, T. (2016). Inside the " Twitter for racists": Gab — the site where Milo Yiannopoulos goes to troll now. *Salon*. Retrieved from https://www.salon.com/2016/11/05/inside-the-twitter-for-racists-gab-the-site-where-milo-yiannopoulos-goes-to-troll-now/

Berger, J. M. (2018). *Extremism*. Cambridge, Massachusetts: The MIT Press.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *12*(4), 421. https://doi.org/10.2307/3050792

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Bracegirdle, C. (2020). *Exploring the limits of intergroup contact: Ingroup contact and group status*. University of Oxford.

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017a). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017b). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Braun, A. F., Dickey, B., Heide, J. Van Der, Maragkou, E., Peeters, S., & Steffen, B. (2020). *Normiefication of extreme speech and the widening of the Overton window*. Retrieved from https://oilab.eu/normiefication-of-

extreme-speech-and-the-widening-of-the-overton-window/

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet*, *7*(2), 223–242. https://doi.org/10.1002/poi3.85

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., … Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 1–14. https://doi.org/10.1007/s13278-014-0206-4

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–1197. https://doi.org/10.1126/science.1185231

Chandrasekharan, E., Jhaver, S., Bruckman, A., & Gilbert, E. (2020). Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ArXiv*. Retrieved from http://arxiv.org/abs/2009.11483

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–22. https://doi.org/10.1145/3134666

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., … Gilbert, E. (2018). The Internet's hidden rules. *Proceedings of the ACM on Human-Computer Interaction*, 1–25. https://doi.org/10.1145/3274301

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Hate is not binary: Studying abusive behavior of #GamerGate on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 65–74). https://doi.org/10.1145/3078714.3078721

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. https://doi.org/10.1080/09546553.2014.849948

Coleman, J., Katz, E., & Menzel, H. (1966). The diffusion of an innovation among physicians. *Sociometry*, *20*, 253–270. https://doi.org/10.1145/1031607.1031658

Conway, M. (2018). *Violent Extremism and Terrorism Online in 2018*. Retrieved from https://www.voxpol.eu/year-in-review-2018/

Conway, M., Scrivens, R., & Macnair, L. (2019). Right-wing extremists' persistent online presence: History and contemporary trends. *ICCT Policy Brief*. https://doi.org/10.19165/2019.3.12

Crandall, C. S., & Stangor, C. (2008). Conformity and prejudice. In *On the pature of Prejudice; Fifty years after Allport*. Blackwell Publishing Ltd.

Davey, J., & Ebner, J. (2017). *The fringe insurgency: Connectivity, convergence and mainstreaming of the extreme right*. Retrieved from http://www.isdglobal.org/wp-content/uploads/2017/10/The-Fringe-Insurgency-221017.pdf

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 512–515. Retrieved from http://arxiv.org/abs/1703.04009

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 11–20. https://doi.org/10.18653/v1/w18-5102

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective Latent Concept Modeling for ad hoc information retrieval. *Document Numerique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

Doosje, B., Moghaddam, F. M., Kruglanski, A. W., de Wolf, A., Mann, L., & Feddes, A. R. (2016). Terrorism, radicalization and de-radicalization. *Current Opinion in Psychology*, *11*, 79–84. https://doi.org/10.1016/j.copsyc.2016.06.008

Dunn, A. G., Surian, D., Dalmazzo, J., Rezazadegan, D., Steffens, M., Dyda, A., … Mandl, K. D. (2020). Limited role of bots in spreading vaccine-critical information among active Twitter users in the United States: 2017-2019. *AJPH Open-Themed Research*, *110*(3).

Ekman, M. (2019). Anti-immigration and racist discourse in social media. *European Journal of Communication*,

34(6), 606–618. https://doi.org/10.1177/0267323119886151

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. *ICWSM*. Retrieved from http://arxiv.org/abs/1804.04649

Evans, R. (2019). *Shitposting, inspirational terrorism, and the Christchurch mosque massacre*. *Bellingcat*. Retrieved from https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology*, *87*(3), 293–311. https://doi.org/10.1037/0022-3514.87.3.293

Ferrara, E. (2017). Contagion dynamics of extremist propaganda in social networks. *Information Sciences*, *418–419*, 1–12. https://doi.org/10.1016/j.ins.2017.07.030

Ferrara, E., Wang, W. Q., Varol, O., Flammini, A., & Galstyan, A. (2016). Predicting online extremism, content adopters, and interaction reciprocity. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10047 LNCS*, 22–39. https://doi.org/10.1007/978-3-319-47874-6_3

Finkelstein, J., Zannettou, S., Bradlyn, B., & Blackburn, J. (2020). A quantitative approach to understanding online Antisemitism. *ICWSM*. Retrieved from http://arxiv.org/abs/1809.01644

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. United Nations Educational, Scientific and Cultural Organization. Retrieved from http://unesdoc.unesco.org/images/0023/002332/233231e.pdf

Gallacher, J. D., & Heerdink, M. W. (2019). Measuring the effect of Russian Internet Research agency information operations in online conversations. *Defence Strategic Communications*, *6*, 155:198. https://doi.org/10.30966/2018.RIGA.6

Gallacher, John D, & Fredheim, R. E. (2019). Division abroad, cohesion at home: How the Russian troll factory works to divide societies overseas but spread pro-regime messages at home. In *Responding to Cognitive Security Challenges* (p. 60:79). Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Ganesh, B., & Bright, J. (2020). *Extreme digital speech: Contexts, responses and solutions*. Retrieved from https://www.voxpol.eu/new-vox-pol-report-extreme-digital-speech-contexts-responses-and-solutions/

Gaudette, T., Scrivens, R., & Venkatesh, V. (2020). The role of the Internet in facilitating violent extremism: Insights from former right-wing extremists. *Terrorism and Political Violence*, 1–18. https://doi.org/10.1080/09546553.2020.1784147

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. Retrieved from https://www.semanticscholar.org/paper/SPSS-for-Windows-Step-by-Step%3A-A-Simple-Guide-and-George-Mallery/230e458b34cdcb463cfe4caa954253bd73456e2e

Glen, S. (2014). Cohen's kappa statistic. Retrieved from https://www.statisticshowto.com/cohens-kappa-statistic/

Global Terrorism Index. (2019). *Global terrorism index 2019: Measuring the impact of terrorism*. Retrieved from https://www.visionofhumanity.org/wp-content/uploads/2020/11/GTI-2019-web.pdf

Graham, E. (2019). Boundary maintenance and the origins of trolling. *New Media and Society*, *21*(9), 2029 –2047. https://doi.org/10.1177/1461444819837561

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(SUPPL. 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grover, T., & Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, 193–204.

Hajjem, M., & Latiri, C. (2017). Combining IR and and LDA topic modelling for filtering microblogs. *Procedia Computer Science*, *112*, 761–770. https://doi.org/10.1016/j.procs.2017.08.166

Haney, C., Banks, C., & Zimbardo, P. (1973). Interpersonal dynamics in a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–97. https://doi.org/10.1037/h0076835

Harel, T. O., Jameson, J. K., & Maoz, I. (2020). The normalization of hatred: Identity, affective polarization, and dehumanization on Facebook in the context of intractable political conflict. *Social Media + Society*. https://doi.org/10.1177/2056305120913983

Haslam, S. A., Reicher, S. D., & Bavel, J. J. Van. (2019). Rethinking the nature of cruelty: The role of identity

leadership in the Stanford prison experiment. *American Psychologist*, *74*(7), 809–822. https://doi.org/10.1037/amp0000443.supp

Hodas, N. O., & Lerman, K. (2014). The simple rules of social contagion. *Scientific Reports*, *4*. https://doi.org/10.1038/srep04343

Hogg, M. A., Abrams, D., & Brewer, M. B. (2017). Social identity: The role of self in group processes and intergroup relations. *Group Processes and Intergroup Relations*, *20*(5), 570–581. https://doi.org/10.1177/1368430217690909

Honeycutt, C. (2006). Hazing as a process of boundary maintenance in an online community. *Journal of Computer-Mediated Communication*, *10*(2), 00–00. https://doi.org/10.1111/j.1083-6101.2005.tb00240.x

Ischinger, W. (2020). *Munich Security Report 2020*. Retrieved from https://securityconference.org/en/

Jagarlamudi, J., Iii, H. D., & Udupa, R. (2012). Incorporating lexical priors into topic models. *EACL 2012 - 13th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, 204–213.

Jasinskaja-Lahti, I., Vezzali, L., Ranta, M., Pacilli, M. G., Giacomantonio, M., & Pagliaro, S. (2020). Conditional secondary transfer effect: The moderating role of moral credentials and prejudice. *Group Processes and Intergroup Relations*. https://doi.org/10.1177/1368430220940401

Javed, J., & Miller, B. (2019). When content promotes hate: Moral-emotional content, outgroup cues, and attitudes toward violence and anti-muslim policies. *ArXiv*, 1–29. Retrieved from http://www.jeffreyjaved.com/uploads/8/6/1/2/86128854/jm_when_content_promotes_hate_041919.pdf

Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., … Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, *573*(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7

Katz, R. (2018). Inside the online cesspool of anti-semitism that housed Robert Bowers. *Politico*. Retrieved from https://www.politico.com/magazine/story/2018/10/29/inside-the-online-cesspool-of-anti-semitism-that-housed-robert-bowers-221949

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., … Dehghani, M. (2020). The Gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv*, 1–47. Retrieved from https://osf.io/edua3/#!

Kleinberg, B., Vegt, I. Van Der, & Gill, P. (2020). The temporal evolution of a far-right forum. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-020-00064-x

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the Internet. *Journal for Deradicalization*, *1*(2014), 116–134.

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kuhn, M., Wickham, H., & RStudio. (2020). Package 'tidymodels': Easily Install and Load the "Tidymodels" Packages. *CRAN*, 1–5. Retrieved from https://cran.r-project.org/web/packages/tidymodels/index.html

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., … Alstyne, M. Van. (2009). Computational Social Science. *Science*, *323*, 721–724.

Lee, B., & Knott, K. (2020). Fascist aspirants: Fascist forge and ideological learning in the extreme right online milieu. *Behavioral Sciences of Terrorism and Political Aggression*, 1–25. https://doi.org/10.1080/19434472.2020.1850842

Levine, J. M., & Moreland, R. L. (1994). Group socialization: Theory and research. *European Review of Social Psychology*, *5*(1), 305–336. https://doi.org/10.1080/14792779543000093

Li, C., Chen, S., Xing, J., Sun, A., & Ma, Z. (2019). Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems*, *37*(1). https://doi.org/10.1145/3238250

Lima, L., Reis, J. C. S., Melo, P., Murai, F., Araujo, L., Vikatos, P., & Benevenuto, F. (2018). Inside the right-leaning echo chambers: Characterizing Gab, an unmoderated social system. *Asonam'18*.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *Plos One*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Makuch, B. (2019). The Nazi-free alternative to Twitter is now home to the biggest far right social network. *Vice Motherboard*. Retrieved from https://www.vice.com/en/article/mb8y3x/the-nazi-free-alternative-to-twitter-is-now-home-to-the-biggest-far-right-social-network

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media.

Proceedings of the 10th ACM Conference on Web Science - WebSci '19, 173–182. https://doi.org/10.1145/3292522.3326034

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*, 415–444.

McSwiney, J., & Jasser, G. (2020). *Gab.com: The pro-Trump alternive social media. VoxPol*. Retrieved from https://www.voxpol.eu/gab-com-the-pro-trump-alternative-to-social-media/

Meleady, R., & Forder, L. (2019). When contact goes wrong: Negative intergroup contact promotes generalized outgroup avoidance. *Group Processes and Intergroup Relations*, *22*(5), 688–707. https://doi.org/10.1177/1368430218761568

Meleagrou-Hitchens, A., & Kaderbhai, N. (2016). *Research perspectives on online radicalization*. VOX-Pol Network of Excellence. Retrieved from https://icsr.info/wp-content/uploads/2017/05/ICSR-Paper_Research-Perspectives-on-Online-Radicalisation-A-Literature-Review-2006-2016.pdf

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS ONE*, *12*(9), 1–12. https://doi.org/10.1371/journal.pone.0184148

Moreland, R. L., & Levine, J. M. (1982). Socialization in small groups: Temporal changes in individual-group relations. *Advances in Experimental Social Psychology*, *15*(C), 137–192. https://doi.org/10.1016/S0065-2601(08)60297-X

Nielsen, J. (2006). The 90-9-1 rule for participation inequality in social media and online communities. *Nielsen Norman Group*, 5–9. Retrieved from https://www.nngroup.com/articles/participation-inequality/

Nikita, M. (2016). Select number of topics for LDA model. *RPubs*. Retrieved from https://rpubs.com/siri/ldatuning

Noel, J. G., Wann, D. L., & Branscombe, N. R. (1995). Peripheral ingroup membership status and public negativity toward outgroups. *Journal of Personality and Social Psychology*, *68*(1), 127–137. https://doi.org/10.1037/0022-3514.68.1.127

Nonnecke, B. (2000). *Lurking in email-based discussion lists*. South Bank University.

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *ArXiv*. Retrieved from http://arxiv.org/abs/1804.05704

Onnela, J. P., & Reed-Tsochas, F. (2010). Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(43), 18375–18380. https://doi.org/10.1073/pnas.0914572107

Pariser, E. (2011). *The Filtter Bubble: What the internet is hiding from you*. New York, New York, USA: The Penguin Press.

Park, J. H., & Fung, P. (2017). One-step and two-step classification for abusive language detection on Twitter. *ArXiv*. https://doi.org/10.18653/v1/w17-3006

Pauwels, L., & Schils, N. (2016). Differential online exposure to extremist content and political Violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence*, *28*(1), 1–29. https://doi.org/10.1080/09546553.2013.876414

Pettigrew, T. F. (1997). Generalized intergroup contact effects on prejudice. *Personality and Social Psychology Bulletin*, *23*(2), 173–185. https://doi.org/10.1177/0146167297232006

Pettigrew, T. F. (2009). Secondary transfer effect of contact: Do intergroup contact effects spread to noncontacted outgroups? *Social Psychology*, *40*(2), 55–65. https://doi.org/10.1027/1864-9335.40.2.55

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*, 922–934. https://doi.org/10.1002/ejsp

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. *Proceedings Ofthe 55th Annual Meeting Ofthe Association for Computational Linguistics*, 729–740. https://doi.org/10.18653/v1/p17-1068

Rajadesingan, A., Resnick, P., & Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*. Retrieved from https://www.aaai.org/ojs/index.php/ICWSM/article/view/7323

Reed, A., Whittaker, J., Votta, F., & Looney, S. (2019). Radical filter bubbles and extremist content. *Global Research Network on Terrorism and Technology*, (8). Retrieved from https://rusi.org/sites/default/files/20190726_grntt_paper_08_0.pdf

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, *2*(3), 1313–1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., … Zannettou, S. (2020). The evolution of the manosphere across the web. *ArXiv*. Retrieved from http://arxiv.org/abs/2001.07600

Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A. F., & Meira, W. (2018). Characterizing and detecting hateful users on Twitter. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 676–679. Retrieved from https://arxiv.org/pdf/1803.08977.pdf

Robbinson, N., Schlegel, L., Janin, M., & Deverell, F. (2020). How are terrorists and violent extremists using gamification? Retrieved July 17, 2020, from https://www.techagainstterrorism.fm/how-are-terrorists-and-violent-extremists-using-gamification/

Rogers, E. M. (2003). *Diffusion of Innovations*. The Free Press.

Rogers, E. M., & Kincaid, L. (1984). Communication networks: Toward a new paradigm for research. *American Journal of Sociology*, *89*(4), 986–988. https://doi.org/10.1086/227967

Rowe, M., & Saif, H. (2016). Mining pro-ISIS radicalisation signals from social media users. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 329–338. Retrieved from http://oro.open.ac.uk/48477/1/13023-57822-1-PB.pdf

Ryan, B., & Gross, N. C. (1943). Acceptance and diffusion of hybrid corn seed in two Iowa communities. *Agricultural Experiment Station - Iowa State College of Agriculture and Mechanic Arts*, *372*(372), 663–705. Retrieved from http://ezproxy.cul.columbia.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=ehh&AN=13074695&site=ehost-live&scope=site

Sageman, M. (2004). *Understanding terror networks*. University of Pennsylvania Press.

Sarrasin, O., Fasel, N., Green, E. G. T., & Helbling, M. (2015). When sexual threat cues shape attitudes toward immigrants: the role of insecurity and benevolent sexism. *Frontiers in Psychology*, *6*(July), 1–13. https://doi.org/10.3389/fpsyg.2015.01033

Schmid, K., Hewstone, M., & Tausch, N. (2014). Secondary transfer effects of intergroup contact via social identity complexity. *British Journal of Social Psychology*, *53*(3), 443–462. https://doi.org/10.1111/bjso.12045

Schofield, A., Magnusson, M., Thompson, L., & Mimno, D. (2017). Understanding text pre-processing for Latent Dirichlet Allocation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, *2*, 432–436. https://doi.org/10.1145/1378773.1378800

Scrivens, R. (2020). Exploring radical right-wing posting behaviors online. *Deviant Behavior*, 1–15. https://doi.org/10.1080/01639625.2020.1756391

Scrivens, R., Wojciechowski, T. W., & Frank, R. (2020). Examining the developmental pathways of online posting behavior in violent right-wing extremist forums. *Terrorism and Political Violence*, 1–18. https://doi.org/10.1080/09546553.2020.1833862

Settle, J. E. (2018). *Frenemies*. Cambridge University Press. https://doi.org/10.1017/9781108560573

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, *40*(2), 211–239. https://doi.org/10.1177/0049124111404820

Sherif, M. (1936). *The psychology of social norms*. Harper.

Silge, J., & Robinson, D. (2017). *Text mining with R: a tidy approach*. Sebastopol: O'Reilly Media.

Singhal, A. (2001). Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*. https://doi.org/10.1350/ijps.2005.7.3.197

Smith, L. G. E., Blackwood, L., & Thomas, E. F. (2019). The need to refocus on the group as the site of radicalization. *Perspectives on Psychological Science*, *15*(2), 327–352. https://doi.org/10.1177/1745691619885870

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive Behavior*, *44*(2), 136–146. https://doi.org/10.1002/ab.21737

Susarla, A., Oh, J. H., & Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from youtube. *Information Systems Research*, *23*(1), 23–41. https://doi.org/10.1287/isre.1100.0339

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tausch, N., Hewstone, M., Kenworthy, J. B., Psaltis, C., Schmid, K., Popan, J. R., … Hughes, J. (2010). Secondary transfer effects of intergroup contact: Alternative accounts and underlying processes. *Journal of Personality and Social Psychology*, *99*(2), 282–302. https://doi.org/10.1037/a0018553

Terizi, C., Chatzakou, D., Pitoura, E., Tsaparas, P., & Kourtellis, N. (2020). Angry birds flock together: Aggression propagation on social media. *ArXiv*. Retrieved from http://arxiv.org/abs/2002.10131

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. (2019). The dynamics of political incivility on Twitter. *SAGE Open*, 1–15. https://doi.org/10.1177/ToBeAssigned

Thomas, E. F., Mcgarty, C., & Louis, W. (2014). Social interaction and psychological pathways to political engagement and extremism. *European Journal of Social Psychology*, *44*(1), 15–22. https://doi.org/10.1002/ejsp.1988

Tuters, M., & Hagen, S. (2018). *Who are (((they)))?: On online hate, tasteless transgression, and memetic versatility*. Retrieved from https://oilab.eu/who-are-they-on-online-hate-tasteless-transgression-and-memetic-versatility/

Tuters, M., & Hagen, S. (2019). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media and Society*, 1–20. https://doi.org/10.1177/1461444819888746

Urman, A., & Katz, S. (2020). What they do in the shadows: examining the far-right networks on Telegram. *Information Communication and Society*, *0*(0), 1–20. https://doi.org/10.1080/1369118X.2020.1803946

Valente, T. W. (1995). *Network models of the diffusion of innovations*. Hampton Press.

Valente, T. W. (1996). Network models of the diffusion of innovations. *Computational and Mathematical Organization Theory*, *2*.

Valente, T. W., Dyal, S. R., Chu, K. H., Wipfli, H., & Fujimoto, K. (2015). Diffusion of innovations theory applied to global tobacco control treaty ratification. *Social Science and Medicine*, *145*, 89–97. https://doi.org/10.1016/j.socscimed.2015.10.001

Valente, T. W., & Vega Yon, G. G. (2020). Diffusion/contagion processes on social networks. *Health Education and Behavior*, *47*(2), 235–248. https://doi.org/10.1177/1090198120901497

Van Kleef, G. A. (2009). How emotions regulate social life. *Current Directions in Psychological Science*, *18*(3), 184–188. https://doi.org/10.1111/j.1467-8721.2009.01633.x

Van Kleef, G. A., Van Den Berg, H., & Heerdink, M. W. (2015). The persuasive power of emotions: Effects of emotional expressions on attitude formation and change. *Journal of Applied Psychology*, *100*(4), 1124–1142. https://doi.org/10.1037/apl0000003

Vega Yon, G., Valente, T., Dyal, S., & Hayes, T. (2020). Package 'netdiffuseR': Analysis of diffusion and contagion processes on networks. *CRAN*. https://doi.org/10.1086/303110

Vidgen, B., Yasseri, T., & Margetts, H. (2019). Trajectories of Islamophobic hate amongst far right actors on Twitter. *ArXiv*, 1–20. Retrieved from https://arxiv.org/pdf/1910.05794

Weich, B. (2019). Inside Gab, the alt-right's social media network that is awash with antisemitism. *The Jewish Chronicle*. Retrieved from https://www.thejc.com/news/world/inside-the-social-media-network-of-the-alt-right-1.477090

Weissman, C. G. (2016). Inside Gab: The new Twitter alternative championed by the alt-right. *Fast Company*. Retrieved from https://www.fastcompany.com/3065777/inside-gab-the-new-twitter-alternative-championed-by-the-alt-right

Wickham, H., Averick, M., Bryan, J., Chang, W., D', L., Mcgowan, A., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, M. L., & Burnap, P. (2016). Cyberhate on Social Media in the aftermath of Woolwich: A Case Study in Computational Criminology and Big Data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Williams, M., & Mishcon de Reya. (2019). *Hatred behind the screens A report on the rise of online hate speech*. Retrieved from https://www.mishcon.com/upload/files/Online Hate Final 25.11.pdf

Wood, S. (2019). Package "mgcv": Mixed GAM computation vehicle with automatic smoothness estimation. *CRAN*. https://doi.org/10.1201/9781315370279>

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? *ArXiv*. https://doi.org/10.1145/3184558.3191531

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. *ArXiv*. Retrieved from http://arxiv.org/abs/2005.07926

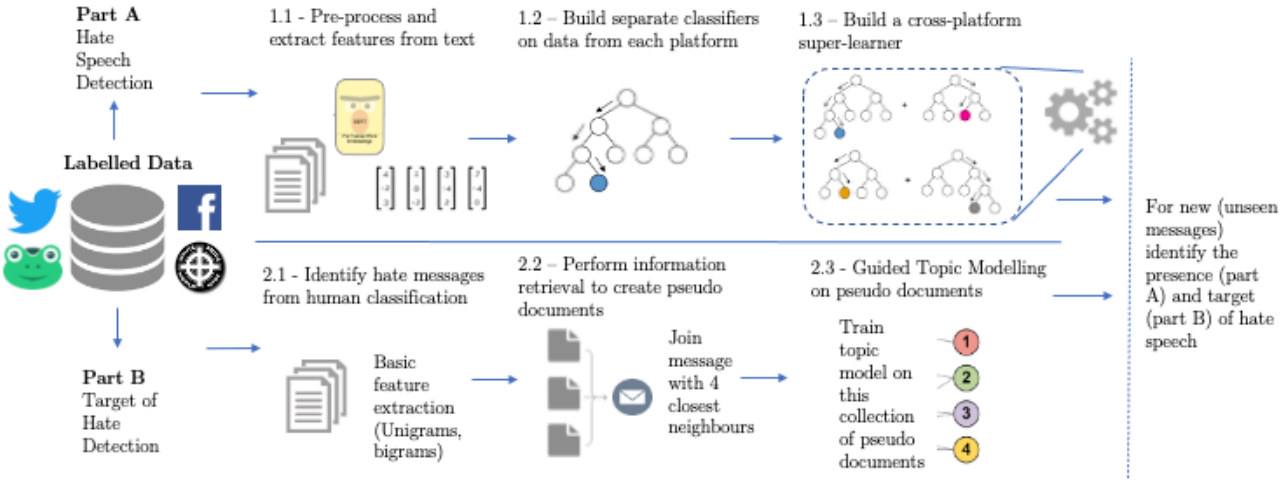
Zhou, Y., Dredze, M., Broniatowski, D. A., & Adler, W. D. (2019). Elites and foreign actors among the alt-right: The Gab social media platform. *First Monday*. https://doi.org/10.5210/fm.v24i9.10062

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., & Xiong, H. (2016). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2105–2114). https://doi.org/10.1145/2939672.2939880

Supplementary Information (SI) for Chapter 3: Hate Contagion: Measuring the spread and trajectory of hate on social media

1 – Hate speech classification

1.1 - Hate classification processing pipeline



SI Figure 1 - Modelling pipeline for hate speech detection (A) and type of hate classification (B)

SI Figure 1 shows the processing pipeline for hate speech detection. Part A gives an overview of how hate speech is detected in social media messages. See Chapter 2 for more details on this step.

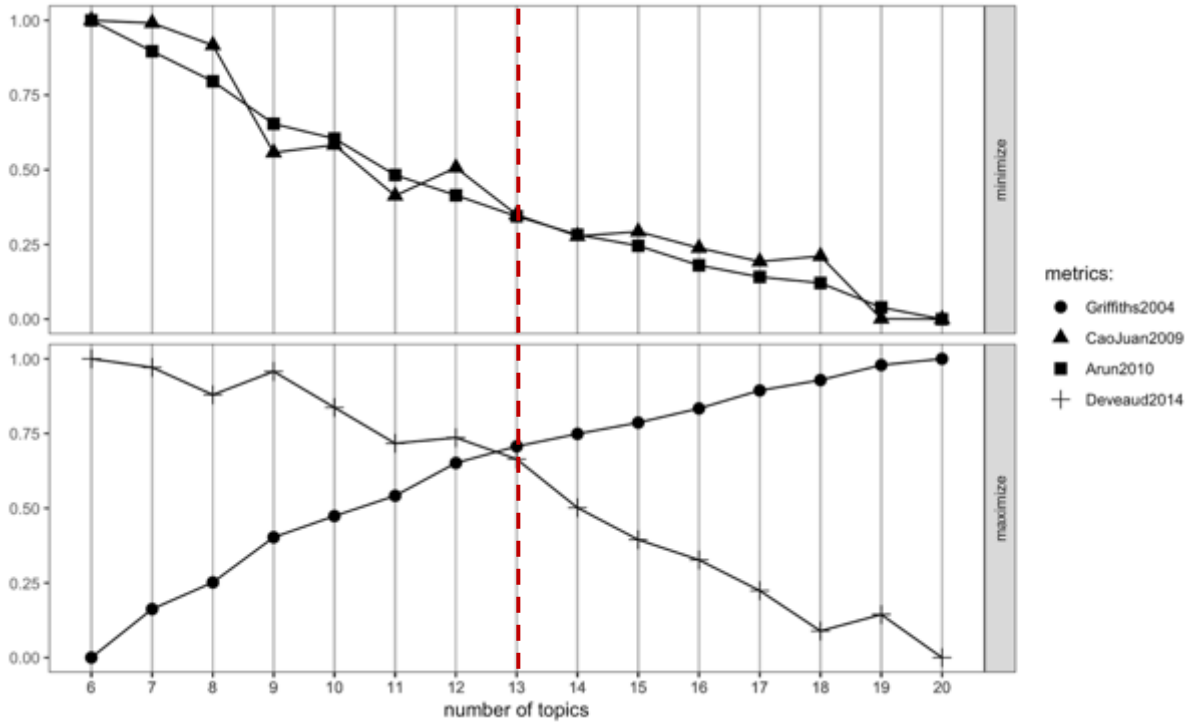
1.2 - Hate speech topic modelling / target of hate detection

In order to identify the group which is targeted within a hate speech message we use a form of semi-supervised machine learning called guided topic modelling (Anoop et al., 2016; Li et al., 2019). The steps in this process are shown in SI Figure 1 part B, and detailed below.

Data pre-processing

Prior to training any topic models we filter out stop words from the messages (linking words without any semantic value, e.g., "and"), make all words lowercase, but avoid stemming the words as is traditional, as this has been found more recently to be detrimental to topic modelling using LDA (Schofield, Magnusson, Thompson, & Mimno, 2017). We then extract both unigrams (single words) and bigrams (word pairs) from the messages and cast the dataset to a document term matrix (SI Figure 1 2.1).

Topic modelling on short texts such as social media messages is notoriously difficult because of the low co-occurrence of words within the messages (Zuo et al., 2016). To overcome this, we use an



SI Figure 2 - LDA tuning to select the appropriate number of topics. Red dotted line gives the number of topics (k) which maximally satisfies all four metrics for model performance.

information retrieval approach with a pseudo-message model. When training the model, for each message across the corpus, we first gather the top four most similar messages to it from across the corpus (see below), and join these together to form a longer 'pseudo-message'. This approach increases the word co-occurrence within each topic and builds greater stability of the topics (Hajjem & Latiri, 2017; Zuo et al., 2016), and performs well on short messages such as those on Twitter or Gab (Gallacher & Fredheim, 2019). The similarity of messages is identified by creating a matrix of term-frequency/inverse document frequency (tf-idf) values for all unigrams & bigrams in the messages, and calculating the cosine similarity of messages (Singhal, 2001). These pseudo-messages are then treated as the documents for topic modelling (SI Figure 1, 2.2).

Selecting the optimum number of topics

Latent Dirichlet Allocation (LDA) approaches for topic modelling demand that the number of topics (here, the number of types of hate speech) is identified ahead of the modelling process. This creates a challenge when we do not know how many topics exist within the dataset. In order to identify this in a more structured way, we use the 'ldatuning' approach (Nikita, 2016), whereby multiple LDA models are created with topics numbering from 2-20 and four different measures of model performance are computed (Arun, Suresh, Madhavan, & Murthy, 2010; Cao, Xia, Li, Zhang, & Tang, 2009; Deveaud, SanJuan, & Bellot, 2014; Griffiths & Steyvers, 2004). The optimum number of topics is then defined

as the value which maximally satisfies all 4 of the measures. For these 19 models we select alpha levels (higher alpha indicates that messages are assumed to be made up of greater mixture of topics) equal to 20 / the number of topics (k) as per the literature (Blei et al., 2003). This method identifies an optimal number of 13 topics as shown in SI Figure 2.

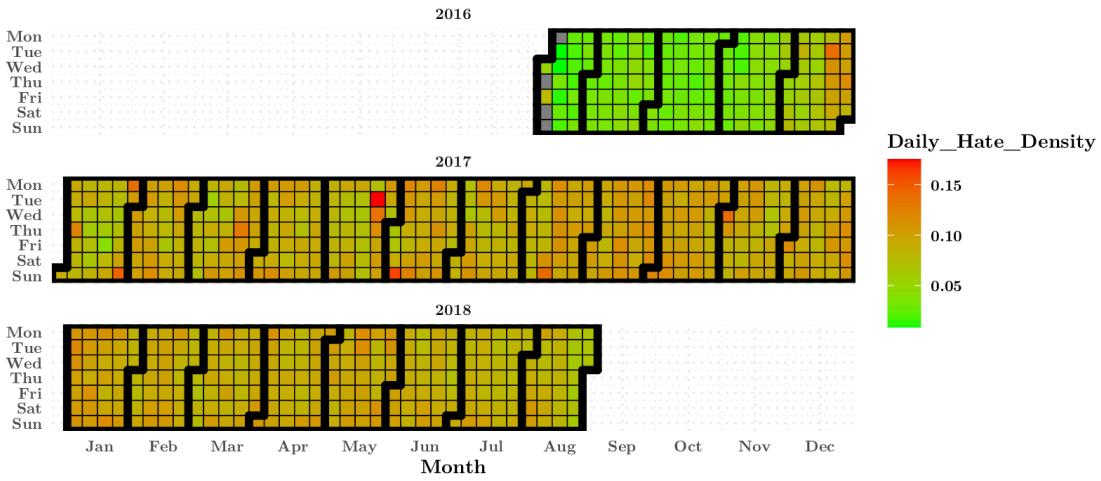
We then manually classify these topics into the type of hate that they represent using the top words (Figure 1) and the most confident judgements (messages with the highest gamma score) per topic. Where multiple of these topics covered the same type of hate, often with subtly different angles, they were combined, leaving 11 distinct topics (Table 2).

2 - Conversation dynamics on Gab

2.1 - Daily hate density over time

SI Figure SI 3 shows the overall proportion of messages which were classified as hate speech (daily hate density) across Gab over the entire collection periods. The results show that there was considerable variation in the amount of hate posted on different days, with some days having much higher densities of hate in the platform conversations than others.

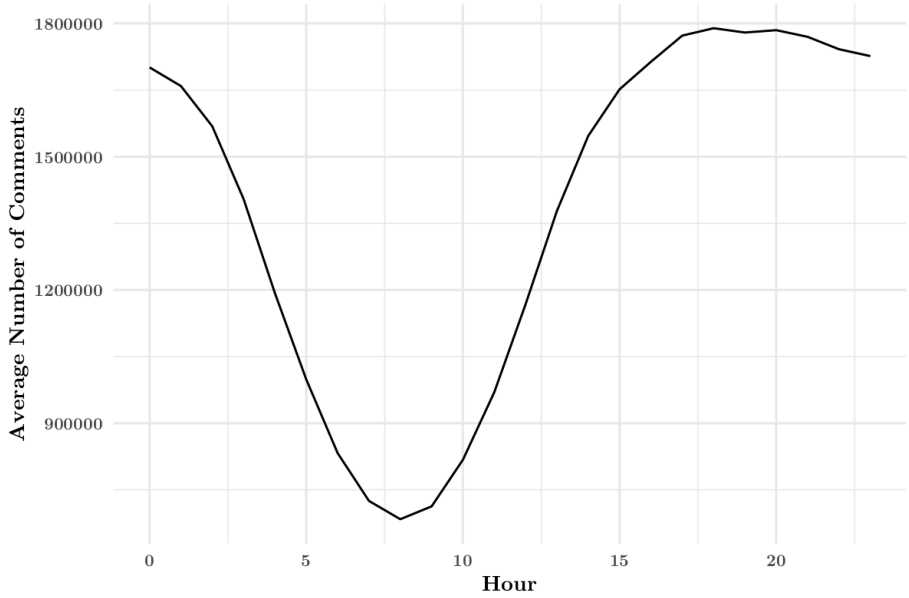
On Gab we find that the density of hate speech quickly rises from the formation of the platform August 2016 before reaching a plateau in January 2017 and then climbing only slightly from here onwards. These trends highlight the importance of accounting for the overall level of hate speech on the platform when looking at individual user trajectories.



Si Figure 3 – Daily proportions of hate speech across (A) Facebook and (B) Gab. Green = lower hate density, red = higher hate density

2.2 - Daily conversation rhythm

Daily Gab conversations followed a consistent pattern with a peak of activity between 1600hrs and 2000hrs and a minimum at 0700hrs (SI Figure 4). Because of this we created breaks in the network at 0700hrs, treating each intervening period as a single 'daily conversation'. This was chosen as breaking the networks at 0000hrs (midnight) would have been near the peak activity, so using 0700hrs instead would lead to minimal conversation broken across days.



SI Figure 4 – Average daily Gab conversation rhythm over the entire period. Distribution of messages shows a clear minimum of activity at 0700hrs

3 – Additional hate speech trajectory analysis

3.1 – Robustness checks for longer term trajectory trends on Gab & log adjusting the user post index for the trajectory models

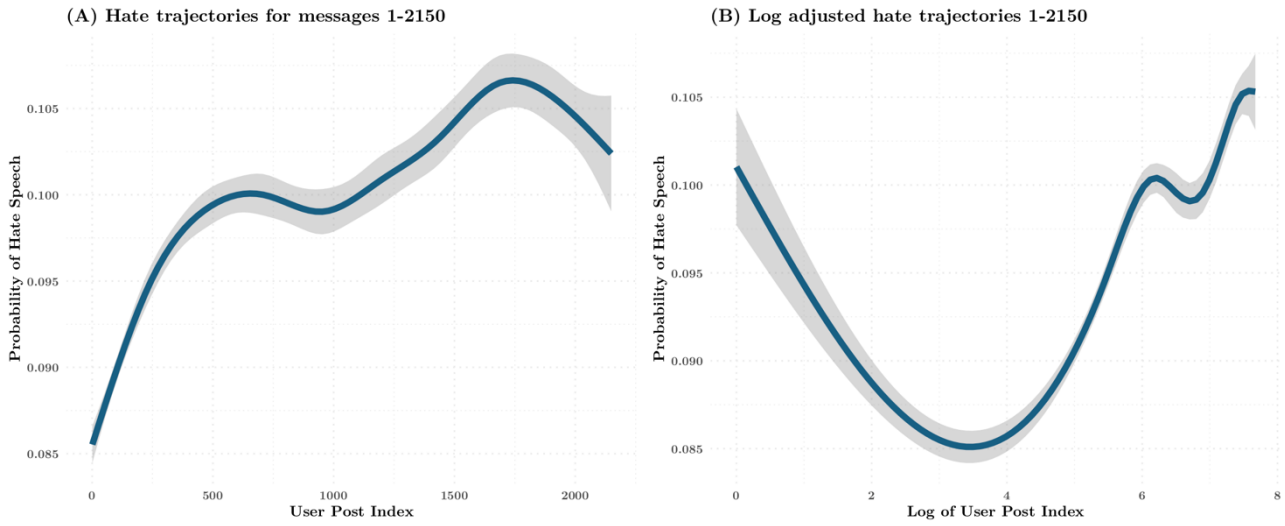
SI Figure 5a shows the average hate trajectory for users on Gab over the period of their first to their 2150th post when controlling for overall density of hate speech on the platform and inter-user differences. Hate increased over the first ~1,000 posts before plateauing (GAMM, n = 18,159,111, df = 1.533, F = 153.3, $p < 0.001$)

SI Figure 5b shows the log-adjusted trajectory of likelihood of hate expression for users on Gab. Taking the log of the user post index reduces the impact of users dropping out from the conversations at the later stages. We find that the overall shape of the trajectory doesn't change by taking the log of the user post index and is broadly comparable– an initial higher chance of a user posting hate when

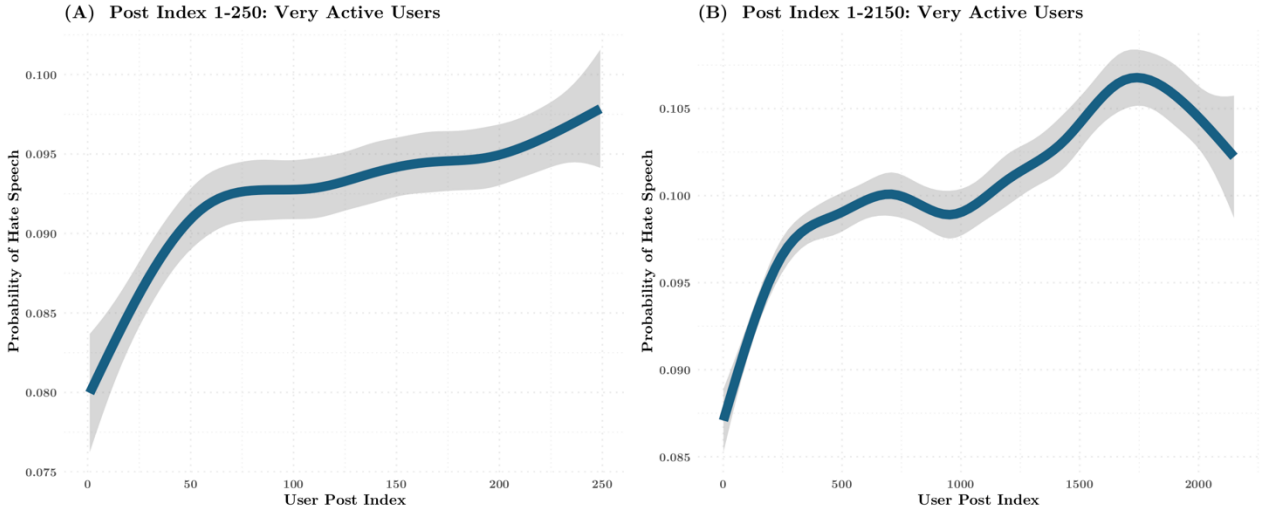
they first join the platform, followed by a decrease before this reverses and likelihood of expressing hate rises over the longer term (GAMM, n = 18,159,111, df = 1.563, F = 169.1, p < 0.001)

3.2 - Removing low activity users

SI Figure SI 6 shows the user hate trajectories after removing users who made fewer than 250 posts to the platform, and retaining over the 'very active users'. We model this over 250 and 2150 posts. In both cases a significant relationship between the user post index and hate speech probability was found (Gab 2150; n = 15,209,090, df = 1.61, F = 11.64, p < 0.001, Gab 250; n = 3,764,000, df = 2.571, F = 16.8, p < 0.001)



SI Figure 5 – (A) Long-term hate trajectory for users on Gab for 1 – 2150 posts, (B) Log-adjusted user hate trajectories over this same period

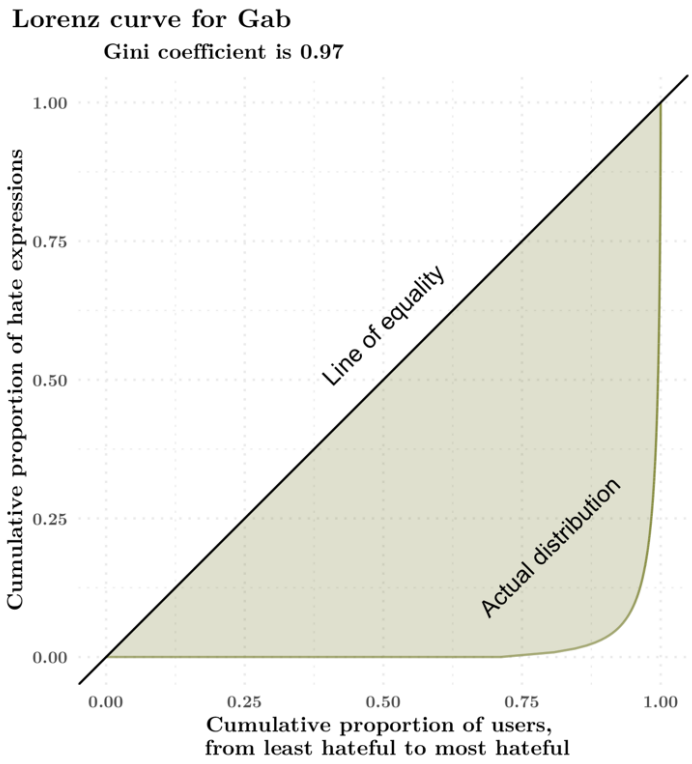


SI Figure 6 – User hate trajectories for 'very active users' who post over 250 times over (A) their first 250 posts and (B) their first 2150 posts

4 – Distribution of hate across users

SI Figure 7 shows the Lorentz curve and the Gini coefficient for hate speech production on Gab. The Gini Coefficient estimates the degree to which hate speech messages are concentrated on a small subset of users as a value between 0 (perfect equality between users) and 1 (one user posting all the hate). This method revealed the dominance of a few 'superusers' in far-right forums (Kleinberg et al., 2020). If all users contributed equally to hate speech on a platform, the cumulative percentage of hate speech posts would increase linearly with the cumulative percentage of users. Conversely, if hate is expressed unevenly, a small proportion of users would produce a large proportion of the hate speech.

We find that hate was very unevenly distributed across users, with a few 'hateful superusers' posting the majority of hate (Gini coefficient=0.97). This means that 50.0% of hate speech posts were generated by only 0.05% of users.



SI Figure 7 – Lorenz curves showing inequality among users for hate speech production for Gab

5 - Short-term exposure effects of hate speech exposure when controlling for longer term exposure trends

SI Table 1 gives the short-term effects of hate speech exposure once controlling for longer-term effects of hate speech exposure over the prior seven days.

SI Table 1 – Short-term (daily, pre-time of adoption) effects of exposure to hate speech once accounting for longer 7-day effects

Hate Speech Exposure Type

BEHAVIOUR (HATE TYPE)	Anti-Semitism			Islamophobia			Misogyny			Anti Black Racism			Anti Immigration			Homophobia		
	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P	ESTIMATE	F	P
Daily exposure pre-toa																		
Anti-Semitism	**8.52**	**3.19**	**0.01**	3.60	2.50	0.10	6.57	0.68	1.00	4.34	0.14	1.00	2.78	1.19	1.00	5.02	0.12	1.00
Islamophobia	1.34	0.92	0.40	2.09	2.13	0.16	**−9.85**	**−3.54**	**0.00**	**6.34**	**2.94**	**0.02**	−2.64	−1.51	0.39	**−9.75**	**−2.90**	**0.02**
Misogyny	9.78	2.16	0.33	9.53	2.84	0.05	−3.06	−0.20	1.00	20.02	1.64	0.83	−20.48	−2.17	0.33	−46.39	−1.56	0.84
Anti-Black Racism	−6.53	−1.17	1.00	**10.54**	**3.64**	**0.00**	−3.23	−0.37	1.00	−21.33	−1.98	0.48	−12.13	−1.59	0.89	7.34	0.66	1.00
Anti-Immigration	2.14	0.92	1.00	**5.93**	**4.49**	**0.00**	6.05	0.89	1.00	2.21	0.24	1.00	6.32	2.67	0.07	−5.66	−0.80	1.00
Homophobia	−15.23	−1.94	0.41	**11.40**	**3.37**	**0.01**	0.65	0.05	1.00	21.99	2.42	0.15	−6.96	−0.99	1.00	**26.72**	**3.43**	**0.01**

6 – Daily example of short-term social contagion

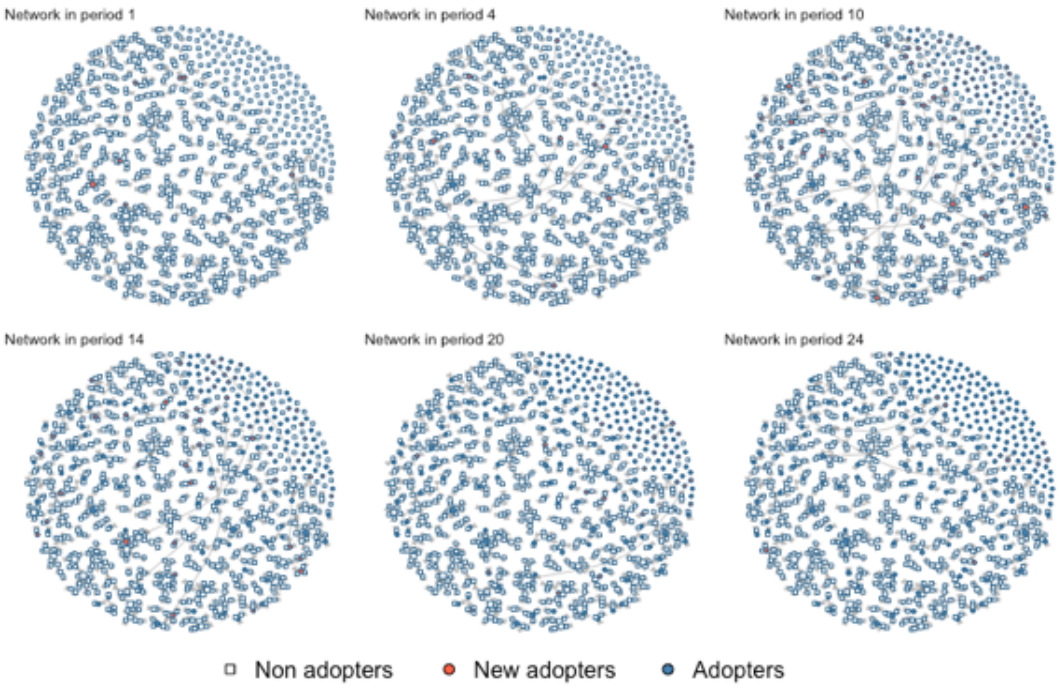
SI Figure 8 shows network slices for a randomly selected day in the dataset (31st March 2018). Users (nodes) and edges (connections) are coloured by their adoption status. From this we estimate the role of connections in shaping the diffusion of behaviours (hate speech expression) through the network.

SI Figure 9 gives the diffusion curve for this contagion pattern. The S-shaped results indicate that early in the diffusion process few people adopt the behaviour, but over time more individuals adopt the behaviour until this reaches a natural plateau / saturation point which is ~30% in this example.

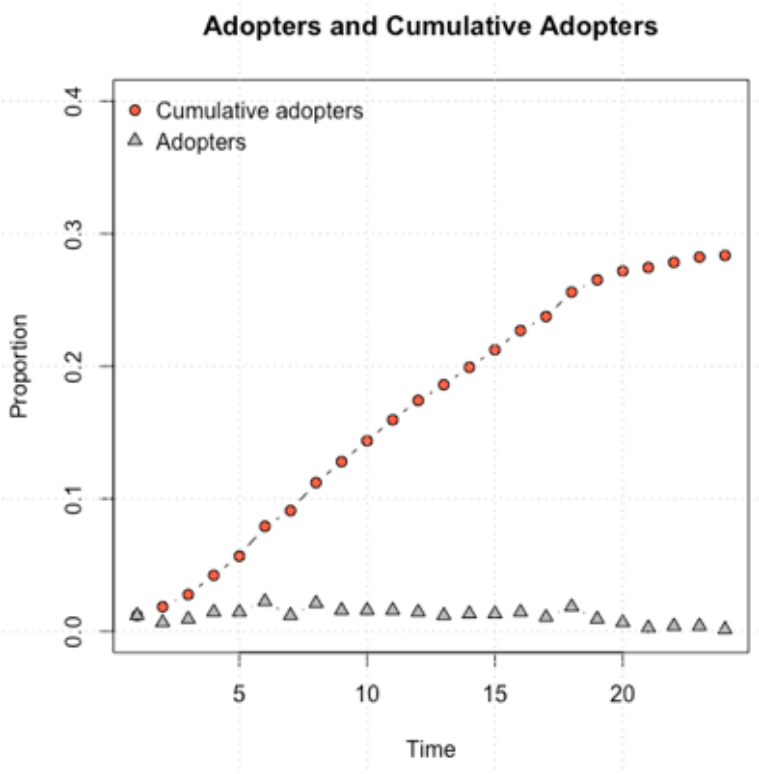
Running a structural dependency test for this day indicates that the network influences the diffusion of behaviour at a level much higher than chance. This tests whether or not a network estimates can be considered a function of the network structure i.e. whether connections and patterns of exposure explain the observed behaviours. By rewiring the graph and calculating a particular statistic t, the test compares the observed mean of t against the empirical distribution of it obtained from rewiring

the network using Monte Carlo simulations. (Simulations = 1,000, nodes = 861, time periods = 24, observed = 0.2839, expected = 0.0004, $p < 0.001$)

SI Table 2 gives the spatial autocorrelation values (Moran's I) for each slice in the network along with the adoption rate.



SI Figure 8 – Example diffusion effects for a single day in the dataset (31st March 2018) across six out of the 24 network slices. Users are shown as nodes, with edges representing network connections in the form of replies, @mentions and re-posts. Nodes are coloured according to their hate adoption state.



SI Figure 9 – Cumulative adopters over time for a single day in the dataset (31st March 2018)

SI Table 2 – Example diffusion effects for a single day in the dataset (31st March 2018) across all 24 network slices

Period	***Adopters***	***Cumulative Adopters (%)***	***Hazzard Rate***	***Moran's I (sd)***
1	9	9 (0.01)	-	0.29 (0.03) ***
2	6	15 (0.02)	0.01	0.39 (0.04) ***
3	12	27 (0.03)	0.01	0.06 (0.04) *
4	11	38 (0.04)	0.01	0.00 (0.04)
5	13	51 (0.06)	0.02	0.04 (0.04)
6	14	65 (0.08)	0.02	(0.04) **
7	15	80 (0.09)	0.02	0.06 (0.04)
8	20	100 (0.12)	0.03	0.06 (0.04)
9	13	113 (0.13)	0.02	-0.05 (0.04)
10	14	127 (0.15)	0.02	-0.07 (0.04) *
11	16	143 (0.17)	0.02	-0.08 (0.04) **
12	21	164 (0.19)	0.03	-0.10 (0.04) ***
13	12	176 (0.20)	0.02	-0.12 (0.04) ***
14	6	182 (0.21)	0.01	-0.13 (0.04) ***
15	6	188 (0.22)	0.01	-0.13 (0.04) ***
16	10	198 (0.23)	0.01	-0.11 (0.04) ***
17	10	208 (0.24)	0.02	-0.11 (0.04) ***
18	6	214 (0.25)	0.01	-0.12 (0.04) ***
19	6	220 (0.26)	0.01	-0.13 (0.04) ***
20	2	222 (0.26)	0	-0.13 (0.04) ***
21	3	225 (0.26)	0	-0.14 (0.04) ***
22	2	227 (0.26)	0	-0.14 (0.04) ***
23	5	232 (0.27)	0.01	-0.12 (0.04) ***
24	4	236 (0.27)	0.01	-0.12 (0.04) ***

Chapter 4

Mutual radicalisation of opposing extremist groups via the Internet

Abstract

Social media has become a common arena for both far-right and Islamic extremist groups to stoke division through the spreading of propaganda and hate speech. This online hate is suggested to drive extremism online, and in some cases lead to offline hate crimes and violence. Whether this online radicalisation happens in isolation within a group, or whether there is an interdependent relationship of mutual radicalisation, is unclear. A possible process by which mutual radicalisation could occur would be if social media incite users to commit offline violence, and if this offline violence in return triggers online reactions from both the target and perpetrator groups. This however has not been tested. This study addresses these questions by investigating the nature of the online-offline relationship of extremist hate. We combine data from the social media platform Gab, variations in Internet search trends, and offline hate crimes in three countries, and test for temporal relationships between opposing extremist groups. Our findings show that online hate from far-right groups both precedes offline violence from these same groups, and spikes following offline violence from opposing Islamic extremist groups. Additionally, far-right Islamophobic violence offline is also followed by increased online interest in Islamic extremist topics. Together, these findings show that the Internet, and specifically hate speech, plays a potential key role in a cyclical process that increases mutual radicalisation.

Introduction

In recent decades there has been a notable rise in Islamic extremism and violence as well as an increase in far-right extremism across much of Europe and the United States of America (Global Terrorism Index, 2019). It has been argued that these two extremist trends have occurred in tandem, with an interdependent relationship of mutual escalation between opposing groups (Guhl & Ebner, 2018). These parallel trends may be an indication of mutual radicalisation, whereby for two opposing groups the actions of one group result in a negative or aggressive reaction from the other, and can lead to these groups shifting in opposing directions (Konaev & Moghaddam, 2010; Moghaddam, 2018). The narratives used by far-right and Islamic extremist groups to motivate their actions demonstrate this trend, with far-right descriptions of *"The west being at War with Islam"* and Islamist extremist narratives *"Muslims are at war with the West"* exhibiting opposing but parallel views (Ebner, 2017, p.197). This may instigate an increasing spiral of hatred and intergroup conflict, with both groups viewing themselves as the victim and the other as the perpetrator (Fielitz, Ebner, Guhl, & Quent, 2018).

The Internet is hypothesised to play an important role in promoting this spiral of intergroup hate, extremism, and violence (Torok, 2013). Social media in particular has historically given extremist groups the unparalleled ability to connect and communicate at scale, both internally and externally, and despite recent crackdowns on the prevalence of extremist material online (Conway, Khawaja, et al., 2019), the online reach of extremist groups is still considerable (e.g. Benigni, Joseph, & Carley, 2017; Berger, 2018b). The role this plays in mutual radicalisation is unclear however, and direct evidence for these effects of mutual radicalisation online is limited. In particular, whether increases in extreme derogation from one group online are associated with subsequent offline violence against these targeted groups, and whether this violence in turns radicalises the targeted groups, is undetermined. However, this role of the Internet may be key to understanding the observed increases in offline violence.

Derogation of the opposing groups is a key feature of the mutual radicalisation process. Online, extreme outgroup derogation is often framed as hate speech, i.e. messages which express hatred towards a targeted group with the intention to attack, disparage, or humiliate members of that group (Davidson, Warmsley, Macy, & Weber, 2017). These messages contain deliberate attacks against (or about) a specific group of people, motivated by (or focused on) aspects of that group's identity (Chapter 2; de Gibert, Perez, García-Pablos, & Cuadros, 2018). The spread of hate speech online is a growing concern. Its propagation across social media is likely to be inflicting harm on the individuals and groups who are targeted and their communities, and the wider society exposed to it (Vidgen, Harris, Cowls, & Guest, 2020). This issue has attracted increasing attention from social media platforms, governments, law enforcement agencies, and civil society. In addition, considerable improvements in automated techniques to detect and measure online hate speech have been made recently, reflecting an increased interest in the study of online hate speech (Vidgen, Tromble, et al., 2019). However, less attention has been given to the role that hate speech plays in dividing online communities, promoting intergroup conflict, and the impact that it may have on individual extremism, group radicalisation, and wider hateful extremist behaviours including offline violence (Kleinberg, Vegt, & Gill, 2020; Siegel, 2020).

There is evidence that online far-right is particularly associated with the creation and distribution of online hate speech (e.g. Berger, 2018b; Conway, Scrivens, & Macnair, 2019; Lee, 2019), and the Internet is suggested to play a central role in breeding and amplifying far-right extremist ideologies

(Ischinger, 2020). Users who are most active and spend more time on far-right fringe social media platforms are more likely to express hate speech themselves (Chapter 3; Javed & Miller, 2019; Mathew, Dutt, Goyal, & Mukherjee, 2018). This effect is likely due to the increased exposure to outgroup hate from other ingroup members (Chapter 3). This suggests that online hate speech can drive radicalisation within groups, however the impact that it has on relations with opposing groups, both those directly targeted and wider groups, is less clear. Furthermore, how reactions to offline attacks from opposing groups can drive this online radicalisation process requires further investigation.

In this study we investigate the relationship between online hate speech and offline hateful extremist behaviour and violence, and what role the Internet and social media platforms are playing in cycles of mutual radicalisation. A conceptual overview of this process is given in Figure 1, illustrating how the Internet may facilitate mutual radicalisation through both reaction and incitement to offline violence. Guided by this idea, we test for evidence of mutual radicalisation at three levels: firstly, whether far-right online hate speech influences far-right offline hate crime (Figure 1 step A), secondly, whether far-right offline violence affects reactionary Internet activity from opposing extremist groups (Figure 1 step B), and finally, whether offline Islamic extremist violence impacts far-right hate speech online (Figure 1 step D). By considering these three links we obtain insights into potential effects of mutual radicalisation between opposing groups via the Internet. Testing the additional stage in this cycle

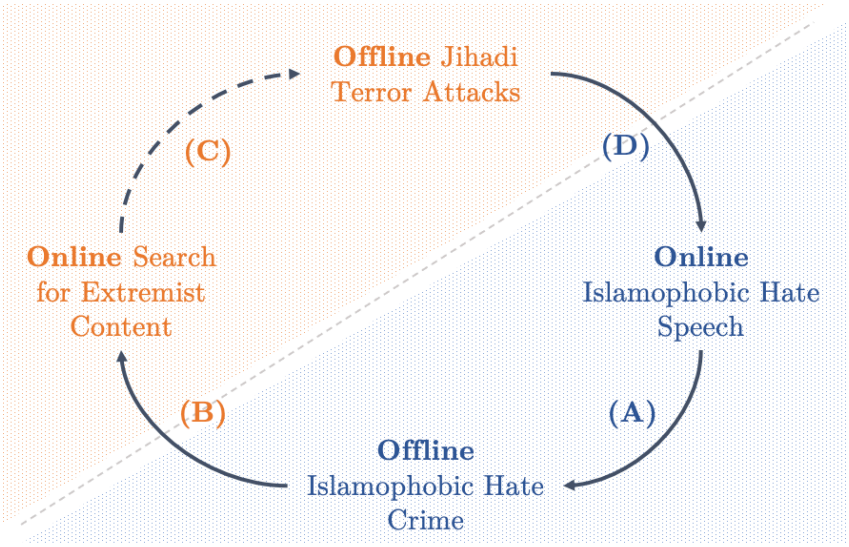


Figure 1 – Proposed cycle of Mutual Radicalisation between opposing far-right and Islamic extremist groups via the Internet. In this example far-right groups are given in Blue and Islamic extremist groups are given in Orange

(Figure 1 step C) whereby interest in joining Islamic extremist groups is linked to subsequent Islamic terror attacks is beyond the scope of this study due to practical limitations around data availability and legal restrictions, the existing evidence for this stage is discussed below.

Mutual radicalisation

Online, hate speech has been linked to driving radicalisation and negative attitudes towards outgroups (Chapter 3). Radicalisation is the process whereby individuals come to increasingly adopt an extreme position through a process of socialisation around shared grievances as the belief that the direct action against the outgroup is morally correct (Reicher et al., 2008; Smith, Blackwood, & Thomas, 2019). This shift towards extremism is characterised by the belief that survival of the ingroup is inseparable from direct action against the outgroup (Berger, 2018a).

For opposing groups taking increasingly extreme positions against one another, it initiates a dynamic process of mutual radicalisation (also referred to as 'co-radicalisation' or 'reciprocal radicalisation'; Knott, Lee, & Copeland, 2018; Pratt, 2015; Reicher & Haslam, 2016), with the groups reacting against real or imagined threats offline, and taking positions further opposing one another (Moghaddam, 2018). This increases intergroup boundaries, perceived intergroup distance, and perceived intergroup threat (Tajfel, 1974; Tajfel & Turner, 1979; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987) and can lead to a greater risk of prejudice, discrimination and conflict (Stephan & Stephan, 2000; Stephan, Ybarra, & Morrison, 2009). This process can manifest as hateful extremist behaviours, i.e. actions that can incite and amplify hate, or make the moral case for violence, while drawing on hateful, hostile or supremacist beliefs directed at an out-group (Commision for Countering Extremism, 2019). Inciting violent reactions from the outgroup can also be a deliberate strategy of terrorist organisations. McCauley & Moskalenko (2008) identify 'jiu-jitsu' radicalisation whereby terror attacks are designed to elicit moral outrage amongst the targeted population and provoke a military response. This military response is then thought to increase the perceived threat posed by this nation and make the local populations more susceptible to local radicalisation.

Moghaddam (2018) identifies three stages in offline mutual radicalisation: i) group mobilisation, ii) extreme ingroup cohesion, and iii) antagonistic identity transformation. This builds on evidence from social psychology that conflict with outgroups can increase ingroup cohesion (Coser, 1956), as the real or imagined threat posed by the outgroup can cause the ingroup to overcome any internal differences

and unite against the outgroup (Stein, 1976). Inside these stages of mutual radicalisation are a number of steps, a subset of which relate directly to hate speech and extreme outgroup denigration. Outgroup denigration acts to (i) increase intergroup differentiation, (ii) normalise public expression of hatred towards the outgroup, (iii) instigate a spiral of increasingly hostile responses to exaggerated outgroup threats, and (iv) shifts the groups to extreme positions against one another. This process has been demonstrated at the national level (Moghaddam, 2018), amongst opposing national and political leadership (Konaev & Moghaddam, 2010; Moghaddam, Harré, & Lee, 2008), amongst Muslims and non-Muslims in Scandinavia (Obaidi, Thompson, & Bergh, 2019), along with others.

Recent evidence of mutual radicalisation between Islamic extremist groups and far-right extremist groups across Europe is increasing (Ebner, 2017; Fielitz et al., 2018). Offline, terrorist violence against civilians predicts an increase in violent behaviour from these same civilian groups against the perceived perpetrators of the original violence (Brandsch & Python, 2020), whilst learning that Islamic extremists dehumanize Westerners increases the likelihood Muslims will be dehumanized in return (Kteily, Hodson, & Bruneau, 2016). Additionally, there is prior evidence for each of the individual links between online and offline radicalisation shown in Figure 1, which this study builds on. However, the role of the Internet in driving this overall cycle of mutual radicalisation remains a central question.

The relationship between online hate and offline violence

It is now increasingly established that online hate speech is linked with offline hate crime. In recent years, numerous high-profile offline far-right terror attacks have been preceded by online activity indicative of the future violence, and while the attackers conducted their attacks alone, they were embedded in global online communities (Ischinger, 2020). Indeed, the concept of 'lone wolf terrorism' is becoming progressively outdated, and evidence suggests that the Internet plays a role in connecting attackers to wider extremist communities (Gill, Horgan, & Deckert, 2014; Schuurman, Bakker, Gill, & Bouhana, 2018; Schuurman et al., 2019).

The attackers at the 2018 Tree of life Synagogue shooting in Pittsburgh, USA, the 2019 synagogue shooting in Halle, Germany, and the 2019 El Paso shooting in the USA all displayed considerable activity on fringe far-right social media platforms prior to their attacks (Evans, 2018; Evans, 2019; Der Spiegel, 2019). Attacks have also been livestreamed over mainstream social media platforms,

including the 2019 terror attack at two mosques in Christchurch, New Zealand (Evans, 2019). Similar evidence has arisen for posts to fringe far-right platforms from attackers responsible for school shootings in the USA, attacks on Black Lives Matter protests, and misogynistic attacks at University campuses (Nagle, 2017), among numerous others. In these events the online environment appears to have played a role in the radicalisation of these attackers, and these attackers then used these same platforms to distribute information of their planned attacks prior to the events (Grover & Mark, 2019).

Beyond these high-profile individual attacks, there is growing evidence that similar trends occur at the group level. Hateful comments increased on Reddit in close temporal proximity to the Unite the Right rally in Charlottesville in August 2017 (Zannettou, ElSherief, Belding, Nilizadeh, & Stringhini, 2020), which left three people dead and multiple injured. The Facebook event page used to organise this event was removed by Facebook prior to the event, due to the high level of hate speech and threats to life (Heath, 2017), and attendees had expressed support for racist neo-Nazi groups and far-right extremist movements prior to attending (Cohen-Almagor, 2018). Beyond western countries, online hate speech has played a leading role in inciting genocide of the Rohingya community in Myanmar (Stecklow, 2018), while Islamophobic hate speech has been linked with deadly mob violence in Sri Lanka (Samaratunge & Hattotuwa, 2014). Online content has also been linked to violence against health workers combatting Polio in Pakistan (Jawaid, 2013), Ebola in Central Africa (Turse, 2019), and most recently to Coronavirus conspiracy theory related violence in the UK (Kelion, 2020).

Systematic evidence that online hate speech predicts offline hate crime (Figure 1 step A)

In addition to these single event observations, systematic research on broader contexts has provided further evidence supporting the link between the online and offline hate environments. Online hate speech and offline racially and religiously aggravated hate crimes are temporally and spatially associated (Williams, Burnap, Javed, Liu, & Ozalp, 2019), even when controlling for known 'trigger events'. In Europe, violence towards immigrants is related to the degree of hate speech expressed on social media in areas where the violence takes place (Müller & Schwarz, 2020a), while in the US, anti-Muslim messages disseminated by President Trump over social media correlate with the number of anti-Muslim hate crimes in states where social media usage is high (Müller & Schwarz, 2020b). Finally, antagonistic online discussions between opposing groups can predict subsequent offline violence between these same groups (Chapter 1: Gallacher, Heerdink, & Hewstone, 2020).

This relationship between online activity and offline violence is also found for more indirect online metrics such as Internet search interest. A higher proportion of racially charged Google search terms is found in areas with higher racial hate crime and racial segregation, and this association increases further with greater broadband availability (Chan, Ghose, & Seamans, 2016). As a result, the Google search volumes for anti-Muslim sentiment can be used to predict anti-Muslim hate crimes (Stephens-Davidowitz, 2017).

Online hate speech may therefore push some users towards offline violence, and while it is unlikely to be the sole cause of this violence, is may act as a 'trigger' or facilitatory mechanism (Briggs & Strugnell, 2011; Gaudette, Scrivens, & Venkatesh, 2020).

Evidence that offline hate crime drives online radicalisation (Figure 1 step B)

Online hate is also found to increase following offline violence, suggesting that offline violence can facilitate online radicalisation within extremist groups. For example, the Charlottesville 'Unite the Right' rally in 2017 led to a 400% increase in search terms indicating a desire to get involved with violent far-right extremist groups in the weeks following the event (Moonshot CVE, 2018), which reflected wider online interest in joining these groups (Tien, Eisenberg, Cherng, & Porter, 2020). Similar effects have been shown for Islamic extremism, and offline Jihadi-inspired terror attacks have led to increases in online posts advocating for further violence amongst Islamic extremist communities (Olteanu, Castillo, Boy, & Varshney, 2018),

This relationship also has a mutual component to it, and areas with larger far-right communities and greater anti-Muslim hostility offline are also linked with greater levels of pro-Islamic State content online (Mitts, 2019), suggesting that offline hostility against a group leads to greater sympathy from members of this group with extremist movements. Similarly, the geographical localisation of anti-Muslim Internet searches is strongly associated with pro-ISIS searches (Bail, Merhout, & Ding, 2018).

As such, the geographical connection between the online and offline environments may, in this context, result in greater offline impacts caused by online radicalisation in areas where use of far-right Internet platforms is high. This warrants further research.

Evidence that offline terror attacks provoke online hate speech (Figure 1 step D)

Offline violence can also cause a reactionary response from opposing groups online. Online Islamophobia has been shown to increase amongst the far-right following Islamic extremist terror attacks (Burnap et al., 2014; Kaakinen, Oksanen, & Räsänen, 2017; Vidgen, Yasseri, & Margetts, 2019; Williams & Burnap, 2016). Similar effects have also been shown whereby online hate speech increases after certain offline 'trigger events' (Awan & Zempi, 2016) including local and national political events (Faris, Ashar, & Gasser, 2016; Saleem, Dillon, Benesch, & Ruths, 2017), and both domestic and overseas terror attacks (Siegel, Tucker, Nagler, & Bonneau, 2017).

This phenomena is further demonstrated by evidence that offline intergroup conflict between far-right extremist groups and opposing groups can lead to increases in the levels of ingroup allegiance for these same far-right groups online, including stronger outgroup sentiments (Bliuc, Betts, Vergani, Iqbal, & Dunn, 2019). However, it is unclear whether these reflexive actions following offline conflict are targeted solely towards the perceived preparators of the violence, or whether they also transfer across to wider outgroups, a process similar to the secondary transfer effect (Pettigrew, 2009). Online, it has been shown that exposure to hate speech against one target group can encourage hate expressions against wider groups (Chapter 3). Whether this is also true for reactions to offline violence deserves investigation, as if so, it could lead to a worsening of intergroup relations across multiple dimensions.

Evidence that online search is related to joining terror groups (Figure 1 step C)

There is growing evidence that Internet searches for how to join terror and extremist groups is related to genuine interest in doing so, and predictive of future recruitment success for these organisations. Bail et al (2018) report a high geographical association between the Google search phrase '*How to join ISIS'* and offline intention to join the terror group. Similarly, online interest for ISIS is shown to be predictive of the onset of offline violence from the group (Johnson et al., 2016). Conversely, campaigns which have aimed to prevent or hinder potential extremists from discovering ISIS-related content on Google have been found to help prevent recruitment into the terror group, providing additional evidence that this is an active route for those looking to join the group (Helmus & Klein, 2018). Although testing this mechanism is beyond the scope of this paper, we include it as a step in our model (Figure 1) for completeness and to help elucidate the entire mutual radicalisation cycle.

The current research

These previous studies demonstrate how the dynamics of online and offline intergroup hate are interconnected, and how the relationship between hate speech and offline violence may be a composite and self-reinforcing process. What remains unclear however is the role of the Internet in facilitating the full-cycle process of mutual radicalisation, whether by allowing for online hate which precedes offline violence or/and by amplifying the impact of offline events from opposing groups by giving users space to express shared grievances against the perceived perpetrator group. Additionally, it remains unclear whether these effects are specific to a particular target group, or whether they are more universal and transfer across the range of minority groups targeted by online hate speech. Determining the role of social media in this process is necessary to understand the impact of increasing online hate speech, and to mitigate for its negative impacts on intergroup conflict.

This study addresses these questions by analysing a combination of hate speech messages from Gab, a fringe social media platform popular with the far-right, data on offline extremist attacks, and Internet search data for extremist terms, in the UK, US and Germany.

Overall analytical approach

We construct our analysis in three steps. First, we identify hate speech in online conversations using a supervised machine learning approach (Chapter 2), and determine the group targeted with this hate (Chapter 3). Secondly, we use Granger causality models to test whether the proportion of online hate speech over the study period relates to offline events including hate crimes and terror attacks from across multiple prejudices, hate targets, and countries, and investigate whether one type of activity consistently precedes the other. Finally, we use Google Trends information to investigate the relationship between offline islamophobia and online Islamic extremist interest. We also use Google Trends data to make local level estimates of interest in Gab and test whether the overall effects between the online and offline environments are stronger at in regions with higher estimated Gab usage.

Gab: a useful platform to study online hate

Fringe social media platforms likely play a key role in the distribution and spread of hate across wider online communities (Nagle, 2017; Zannettou et al., 2017). One of these platforms is Gab, a micro-blogging platform designed as a replica of Twitter, but with lax hate speech moderation policies and

an emphasis on tolerance of 'freedom of speech'. It is known for its high proportion of messages containing explicit hate terms (Lima et al., 2018; Zannettou et al., 2018).

Gab was taken offline in 2018 after the shooter at the Tree of Life Synagogue attack was found to have posted anti-Semitic messages on the platform immediately prior to the attack (Mathew, Dutt, Goyal, & Mukherjee, 2019). It has since returned in a distributed format, making it even less resistant to moderation, and is more popular than ever (Gilbert, 2019; Tech Against Terrorism, 2019). The platform is heavily focused on political content and topics closely follow current affairs, particularly around ideology, race, and terrorism (Zhou, Dredze, Broniatowski, & Adler, 2019).

This makes Gab a useful platform to study the relationship between online and offline violence. Firstly, the known prevalence of far-right activity and connection with offline violence (Evans, 2018) means that any signals of future violence are likely to be present on the platform. Secondly, Gab data is available historically and the platform is not ephemeral unlike other fringe platforms popular with the far-right (e.g. 4Chan, 8Chan), so it is possible to investigate historical trends. Finally, due to the relatively small userbase, compared to mainstream platforms, we are able to measure hate on the platform as a whole without the need to down-sample or focus on any particular subset from the platform which could lead to a biased estimate of overall hate.

This smaller userbase however, has the limitation that much other far-right extremist content is hosted on other fringe platforms, such as Parler (Aliapoulios, Bevensee, Blackburn, & Cristofaro, 2021), Voat (Papasavva, Blackburn, Stringhini, Zannettou, & de Cristofaro, 2020), Telegram (Baumgartner, Zannettou, Squire, & Blackburn, 2020), and others. While it is beyond the scope of this study to include measurements of these platforms, an awareness that similar parallel conversations likely also take place elsewhere is important.

Countries studied

For this analysis we focus on three countries: the United States of America (USA), the United Kingdom (UK) and Germany. These countries were selected as they have the highest levels of Gab usage (Zannettou et al., 2018). Importantly however they differ in primary language (English vs German). The hate speech detection models we use are trained on English language data and so will not detect German language hate speech. We therefore expect the relationship between the online and offline spaces in the USA and UK to be stronger than in Germany, because in the latter country only

a small proportion of the online conversation will take place in English. As such, including Germany will allow us—if indeed we find a weaker association in Germany than in English-speaking countries—to demonstrate more clearly that we are measuring a direct link between online conversation and offline space, rather than an effect driven by an external global variable, which would result in similar associations in all three countries.

Research Aims

This study aims to improve our understanding of how online hate speech relates to offline hateful extremist violence. By measuring the temporal relationship between online hate speech and offline hate crimes, and measing the reactionary responses, we aim to test how the Internet affects the cycle of mutual radicalisation (Figure 1) between far-right and Islamic extremist groups. By focusing on the entire cycle of mutual radicalisation our study aims to provide a more complete picture of the process, complimenting previous studies which have focused on single stages in this cycle.

First, we test whether far-right online hate speech predicts far-right offline hate crimes (Figure 1 step A). Then, we investigate mutual radicalisation between far-right and Islamic extremism, first by testing whether Islamophobic offline hate crimes are followed by greater online interest in Islamic extremist groups (Figure 1 step B), then by testing the relationship between offline Islamic extremist terror attacks and subsequent far-right online hate (Figure 1 step D).

Based on the theory of mutual radicalisation, we hypothesise that all three relationships described above take place and that we will find that (i) increases in far-right online hate speech precede increases in far-right offline violence, (ii) increases in Islamophobic offline hate crime precede increases in Google search trends for terms linked with Islamic extremism, and (iii) increases in attacks linked with or inspired by Islamic extremist groups lead to increases in far-right online hate speech..

In addition, we explore whether this first potential relationship between far-right online hate speech and offline violence varies with usage of fringe social media platforms in the local geographical area. We also explore whether this relationship is specific to certain hate types, i.e. whether the groups targeted online are also targeted offline, or whether the effect is more general and transfers across hate types. If this transfer occurs, we would expect to observe increases in online hate against one group to precede increases in hate against wider groups offline.

Methods

Data collection

We utilise data collected from fringe micro-blogging social media platform Gab. We analyse all the messages from the entire platform posted between its formation on 10th August 2016 and the 29th October 2018 when the platform was taken offline following a terror attack committed by one of its users. This dataset contains 33,089,208 messages posted from 259,598 accounts. This period consists of an amalgamation of data shared by (Zannettou et al., 2018) which covers August 2016 – January 2018 and data from the online repository Pushshift which covers the remainder of this period.

Natural language classification

For each message in the whole corpus we collect two natural language classifications: whether the message contains hate speech, and (if so) what type of hate speech (see Chapter 3 Supplementary Information (SI) Figure 1 for the hate speech processing pipeline).

Hate speech detection

To detect hate speech in the Gab dataset we used a previously developed supervised machine learning approach (see Chapter 2 for details) trained on datasets from Facebook, Twitter, Stormfront and Gab in order to detect various presentations of hate speech against multiple target groups. This approach uses pre-trained contextualised word embedding models (BERT) to identify the semantic features of online hate, along with syntactic features (such as message length and linguistic complexity) and non-linguistic features (such as hate symbols) to identify hate when it is presented in a more nuanced or subtle way. The model classifies all messages into one of three ordered categories; 'clean', 'offensive', or 'hate speech'. The inclusion of this intermediate category is found to improve performance over binary classifiers (Davidson et al., 2017).

This approach has been shown to have a good level of performance for Gab data, with an overall accuracy of 90.9%. The precision and recall values for hate speech in particular are 0.75 and 0.83, indicating that this approach will detect more than 8 out of 10 of the hate speech messages on the platform. Conversely, one out of four of the messages flagged as hate speech will be highlighted incorrectly. Whilst not perfect, these levels of accuracy allow us to make a good approximation of the overall hate speech on Gab over time. The overall proportion of hate speech within the Gab conversations over the period studied is shown in SI Figure 2.

Target of hate detection

In order to identify the type of hate speech expressed in the messages, i.e. the group they target, we use guided topic modelling (Anoop, Asharaf, & Deepak, 2016; Li, Chen, Xing, Sun, & Ma, 2019). Specifically, we used Guided Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), and to ensure consistency we used training data from the same four social media platforms used to train the earlier hate speech detection model outlined above. The details of this approach are described in Chapter 3, and we use the same LDA model here.

Topic modelling identified six distinct types of hate speech present in the dataset: anti-Semitism, anti-immigration, anti-Black racism, Islamophobia, misogyny, and homophobia, along with an 'other' category for messages which could not be classified reliably or contained derogation against other less common groups. These six hate types have been previously highlighted as the most prominent online by a UN report on online hate (Ischinger, 2020), and as the primary groups targeted by the online far-right (Conway, 2018). For each hate speech message, we gathered the type of hate it contained, along with a confidence judgement. Low confidence judgements were re-classified into the 'other' category. Accuracy across the six retained topics was high (average 79.7%) but varied between topics (anti-Semitism 90.0%, anti-Black racism 89.6%, anti-immigration 81.0%, Islamophobia 73.7%, homophobia 72.2%, misogyny 71.8%). Further details including inter-coder reliability checks, equivalence zones, and the most salient words in each of these six hate topics are detailed in Chapter 3.

Offline extremist violence data collection

In order to explore the relationship between online hate speech and offline hateful extremist behaviour we collected data on far-right terror attacks and hate crime from official Government statistics and police reports (e.g. Federal Bureau of Investigation, FBI), civil society crowdsourced datasets (e.g. Anti-defamation league) and academic organisations (e.g. Centre for Research on Extremism Threats / Global Terrorism Database), to build the fullest publicly available picture available of offline hate over the period of interest (August 2016 – October 2018). Details on the sources of offline events along with numbers of events are shown in SI Table 1.

Here we use a definition of hate crime similar to that of the FBI: *"crimes in which the perpetrators acted based on a bias against the victim's race, colour, religion, or national origin*" (FBI Crime Data Explorer, 2020), but we expand it to include any targeted group based on aspects of that group's

characteristics to align with our earlier definition of online hate speech (see additional discussion in Chapter 2). In all cases we collected the date and location of the events, as well as the motivation of each attack and the details on the group targeted to infer the type of hate crime committed (Islamophobia, anti-Semitism etc.). Evens were manually deduplicated across datasets to minimise the impact of hate crimes reported to multiple organisations (see SI 1). We also collected information on the number of perpetrators of each attack and the number of victims.

Overall, we collected 21,151 offline hate crime events from across all countries. The spatial and temporal distributions of these events are shown in Figure 3. Across the entire period, 19,984 of these events occurred in the USA, 470 occurred in the UK and 561 in Germany. In the USA the largest category of attacks were racially motivated anti-Black hate crimes (26.6%), followed by homophobic hate crimes (17.7%%), with anti-immigration (12.5%), anti-Semitism (12.3%), misogyny (8.2%) and Islamophobia (4.6%) following. 'Other' motivations accounted for 27.2% of hate crimes within this data. Importantly, hate crimes can be associated with more than one of these categories, which is why the total percentage does not tally to 100% (see SI 1 for further details on motivation classification). In the UK the most common motivations for attacks were Islamophobia (52.6%) and anti-Immigration (51.9%), followed by 'other' (32.3%), anti-Semitism (16.0%), homophobia (8.9%), misogyny (8.2%), and anti-Black (1%). In Germany the proportions were similar to those in the UK: anti-Immigration (38.0%), Islamophobia (27.8%), 'other' (32.4%), anti-Semitism (19.1%), homophobia (13.5%), misogyny (9.8%), anti-Black (1.1%).

In addition to far-right violence, we also collected data on Islamic extremist-inspired terror attacks perpetrated against the West during the study period (August 2016 – October 2018). Here, we take a definition of 'the West' as Europe, North America and Australasia. We focused on these regions as terror attacks there are likely to have received high media and press coverage in English speaking media (Chalabi, 2018; Wendling, 2016), whereas attack in non-western countries do not receive the same level of attention in the West. This data is collected from the Global Terrorism Database (The National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2020) and Wikipedia. In total we collected data on 117 events. As with far-right violence we collected the location and date of the attack, the number of victims and the number of perpetrators. Unlike with far-right attacks whose specific target we identify, we do not differentiate Islamic extremist attacks by victim/target as these attacks are typically carried out against the secular/Christian West as a whole,

and inspired by a complex mix of religious and historical motivations, as well as perceived military operations and territorial incursions (Institute for Global Change, 2016).

To measure offline violence, it is important to include a measure of the severity of the violence rather than simply the presence/absence of it (Chapter 1). As such, for each event we computed an 'offline violence' metric calculated as the number of perpetrators of the attack multiplied by the number of victims. In cases of attacks against property rather than individuals, this was treated as if the number of victims was equal to one. This effectively treats each interaction between an attacker and a victim as a separate sub-event (1 person attacking 10 victims is given the same violence score as 2 people attacking 5 victims, as in each case there are 10 potential connections in the violence). This approach is similar to violence indices used in other studies, e.g. using the number of victims (Ritchie, Hasell,

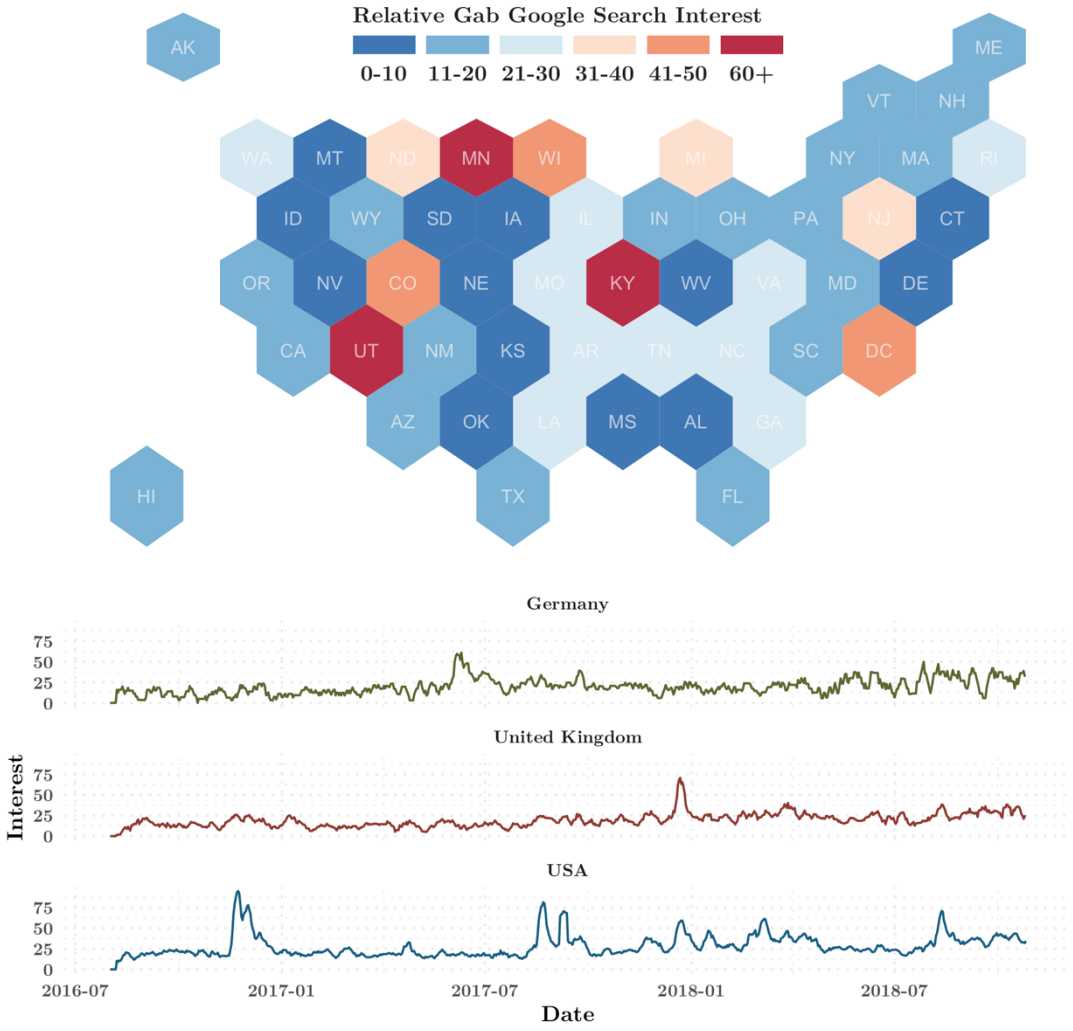


Figure 2 – Relative Google search interest for 'Gab' over the period of interest for the USA, UK and Germany. The map shows the geographical distribution of search interest in the USA for a given month (August 2017) for demonstration purposes.

Appel, & Roser, 2013), while also placing importance on the number of perpetrators as this also reflects the severity of the intention to cause harm (Dugan, LaFree, Cragin, & Kasupski, 2008).

Google Trends data collection

We collected Google Tends data to infer local estimates of Gab interest and to measure local and national interest trends in Islamic extremist topics. Google Trends is an underutilised data resource for the study of online radicalisation and intergroup conflict. The predictive power of search data has been demonstrated in diverse fields (Jun, Yoo, & Choi, 2018). For example, increases in Google search terms associated with influenza symptoms can be a leading indicator of subsequent outbreaks (Ginsberg et al., 2009) and while these results are sometimes overstated (Lazer, Kennedy, King, & Alessandro, 2014), they indicate the power of using search trends to infer wider behaviours.

Measuring relationships between online and offline activity with Granger causality

To investigate the relationship between online hate speech and offline hate crime we constructed time series of both online hate speech and offline hate crime activity and compared the dynamics of these timeseries using Granger causality tests (Granger, 1969). These test whether changes in time series A are predictive of changes in time series B, and construct a test of significance by comparing the predictive power of this relationship between A and B in comparison to using timeseries A to predict itself. All timeseries were normalised to remove any overall trends (upwards / downwards) and any effects of seasonality.

To better understand the directionality of the relationship between online and offline activity, we compared the relationship in both directions. If online information predicts offline events, but not the reverse, then this indicates that the former time series 'Granger causes' the latter. Insight into the directionality of this relationship is the primary benefit of this approach over more simple correlation measures, although we should note that this is not a true causal statement, as other confounding variables may also drive this relationship. This Granger causality is therefore better considered as an approach for forecasting one timeseries using another. This approach has previously been successful in identifying the relationship between online activity and civil unrest (Bastos, Mercea, & Charpentier, 2015).

To control for effects of changing platform popularity over time, we used the proportion of hate speech messages over the whole daily activity on Gab rather than simply the number of hate messages. As such, more active periods on Gab will not necessarily appear as more hateful, but increases in the density of hate on the platform will.

We compared this relationship between online and offline activity at multiple levels and in different countries separately to ensure robustness of results. Initially we compared how the daily proportion of hate speech on Gab relates to the overall daily offline far-right violence metric for the same period, then to the offline Islamic extremist violence metric, and calculated the two-way Granger causality for each. For each Granger causality test, we ascertained the appropriate lag between the two timeseries by computing the relationship for lags ranging from 1-7 days and retaining the optimum selected as that which minimised Akaike Information Criterion (AIC) (Giles, 2011; Thornton & Batten, 1985). We selected seven days as the upper limit of this lag as effects beyond this range are unlikely to be reliable given the fast turnover of online content.

We then performed more targeted analysis, comparing the online activity around specific hate types to the offline hate crimes of the same type. We look at these effects separately for the USA, UK, and Germany. Statistical analysis was performed using the lmtest package (Hothorn et al., 2020) and significance values were adjusted with Holm's correction to control for the effects of multiple comparisons.

Estimates of local Gab usage

We used local Gab Google search interest as a geographical correction factor for our offline violence data. Gab data is not geolocated, and so to take an approximation of Gab usage by country and region over this period, we collected the Google search interest for 'Gab' (taking the specific social media search term rather than keyword) over the period using the Google Trends tool (Google, 2020). This gives us the relative interest in the search term over a given period, and the distribution of this interest over states (USA), and sub-national regions (UK and Germany) (see Figure 2 for an example).

We make this estimate in order to provide additional validation for the potential relationship between online hate and offline violence. If there is an instrumental relationship between the two, we would

expect a more significant association (Granger causality F-statistic) in geographical areas where Gab usage is higher. This is to say that following spikes in online hate, any subsequent offline violence would be more likely to occur in areas with a lot of Gab users present, compared to areas with very few Gab users. As such, we would expect the relationship of online hate on offline violence to be more significant, i.e. a greater predictive power of online hate on offline violence, after weighting by local Gab interest.

Google Trends data is only available at the daily level for periods shorter than 9 months and so in order to collect a correctly normalised dataset over the entire period we collect overlapping 9-month periods (1-month overlap) and re-scale the subsequent dataset to match the trailing/overlapping final month of the previous collection period. This approach accurately estimates daily Google search interest over longer than 9-month periods (Bleher & Dimpfl, 2019; Tseng, 2019). We perform this calculation separately for each geographical area, giving us daily trends in the relative google search interest at the country level. Subsequently, to estimate Google search interest for each country's sub-regions we rescaled the local daily search interest by the country-wide search interest on that day from the national timeseries data. Data collection was performed using the gtrendsR package (Massicotte & Eddelbuettel, 2020) to connect to the Google trends API. We used this Google trends data to weight the violence of each offline event by the relative level of Google search interest in the state/sub-region on the day of the event, to obtain a timeseries of offline violence weighed by local Internet interest in the Gab platform.

We re-ran the Granger Causality analysis outlined above (on the general hate measure only, not separated by type) with this new corrected timeseries, for each country, and compare the significance of that relationship to that of the original model (without correction for local Gab interest).

Estimates of interest in joining Islamic extremist groups

We used the approach from Bail et al (2018) to estimate the local interest in Islamic extremism and in joining Islamic extremist groups, and make use of the search phrase '*How to join ISIS'* which is highly correlated with offline intention to join the terror group. Robustness checks performed by Bail et al show that the level of false positive associations with this term are low, and few people enter this term while searching for unrelated topics. This indicates it is a reasonable proxy for broader Islamic extremist sentiment (Ceron, Curini, & Iacus, 2019). It should be noted however that the relative

power and influence of ISIS as a terror organisation was declining globally over the study period (Burke, 2017), and was of varying relevance to citizens in Europe vs the USA (Byman, 2016) due to geographical distance from the group and the prevalence of terror attacks. We accounted for these effects by ensuring the timeseries for ISIS search interest was de-trended prior to conducting the Granger causality to account for reducing global sway, and by conducting separate analyses for the USA, UK, and Germany.

As with the Gab interest outlined above, we collected country-wide interest over the entire period for the USA, UK and Germany. We then collected search term volumes at the local daily level and combined these with long-term estimates to generate local long-term trends. Due to low search volumes in the UK and Germany at the local daily level, we only collected the local daily trends for the USA. These trends are shown in SI figure 1.

Data analysis

All analysis was done in R (version 3.6.1) using the tidyverse (Wickham et al., 2019) and tidymodels (Kuhn, Wickham, & RStudio, 2020) collections of packages. Where particular additional packages have been used, they have been referenced in the text.

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Reference: SSH_OII_CIA_19_062). All data collection was conducted using open source methods and publicly available data, and hence, informed consent was not explicitly obtained. In order to preserve anonymity, we took a cryptographic hash of all usernames prior to analysis, and real account usernames were not used in the analysis at any point.

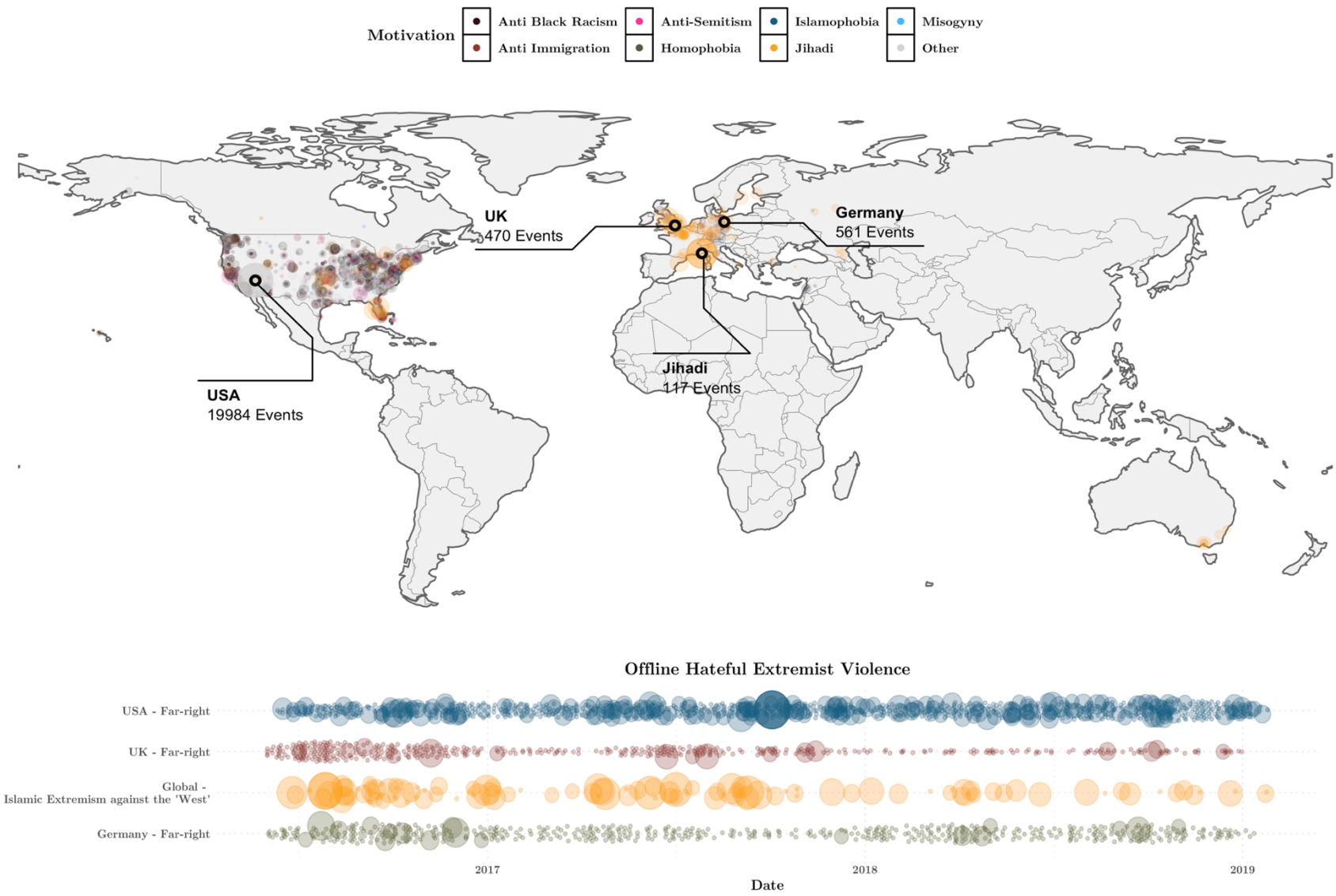


Figure 3 – Offline events used in the Granger causality analysis. Upper panel gives the spatial distribution of events coloured by motivation. Lower panel gives the temporal distribution of events in each country. Larger circles indicate greater levels of violence

Results

Overall, 8.9% of messages on Gab during the study period were classified as hate speech. Islamophobia was the most common type of hate speech with 24.1% of all hate messages classified into this category, followed by anti-immigration sentiment (22.1%) and then anti-Semitism (9.5%). anti-Black racism (2.2%), homophobia (1.9%) and misogyny (1.6%) were less common in this dataset, but still with substantial quantities given the dataset size. The remaining 38.5% of hate speech messages were contained within the 'other' category.

Online hate speech precedes far-right offline violence

When testing for two-way Granger causality between the timeseries in online hate and offline violence, we observed significant Granger causality in the USA and the UK between overall levels of hate speech on Gab preceding far-right hate crime offline (USA; df = 747, F = 6.80, p = 0.008, UK; df = 747, F = 4.93, p = 0.026), while there was no observed effect in the reverse direction. (USA; df = 733, F = 1.31, p = 0.270, UK; df = 738, F = 1.250, p = 0.273). This means that the proportion of hate speech on Gab was predictive of future offline hate crimes, in both countries, with the strength of this effect larger in the USA. In Germany we observed no relationship in either direction, see Table 1.

Looking at these effects for specific hate types, we found significant Granger causality in the USA for anti-Semitism, Islamophobia and anti-immigration, whereby in all three cases spikes in online hate of this type preceded increases in offline hate crimes against the same specific target groups. In the UK, this effect was most significant for Islamophobia, less but still significant for homophobia and anti-immigration, but not significant for anti-Semitism or other hate types. Looking at the reverse direction, we found a relationship where offline Islamophobic hate crimes preceded online hate speech on Gab in the UK. Again, no specific relationships were observed in Germany. Full results across all hate types and countries are given in Table 1.

Offline Islamic extremist violence precedes online far-right hate speech

With regard to extremist terror attacks, spikes in the level of offline Islamic extremist violence preceded spikes in online far-right hate speech ($df = 741$, $F = 82.78$, $p < 0.001$) – the reverse to the effect observed for far-right violence, where online hate speech preceded offline violence. This effect occurred predominately in terms of Islamophobia and anti-immigration sentiment online, but also spread to include misogyny and homophobia (Table 1). Figure 4 demonstrates this effect for Islamophobia in the USA, whereby spikes in Islamic extremist terror attacks preceded increases in Islamophobia on Gab, which in turn preceded increases in offline Islamophobia hate crime. To check whether this primary Granger causality effect (online behaviour preceding offline events) was driven purely by a relationship between offline events we also ran Granger causality tests directly between offline far-right and Islamophobic events, and found a Granger causality effect of Islamic extremist terror attacks preceding increases in Islamophobic hate crimes in the UK ($df = 741$, $F = 9.41$, $p < 0.001$) and Germany ($df = 741$, $F = 22.0$, $p < 0.001$) , but not in the USA ($df = 743$, $F = 0.290$, $p = 0.590$).

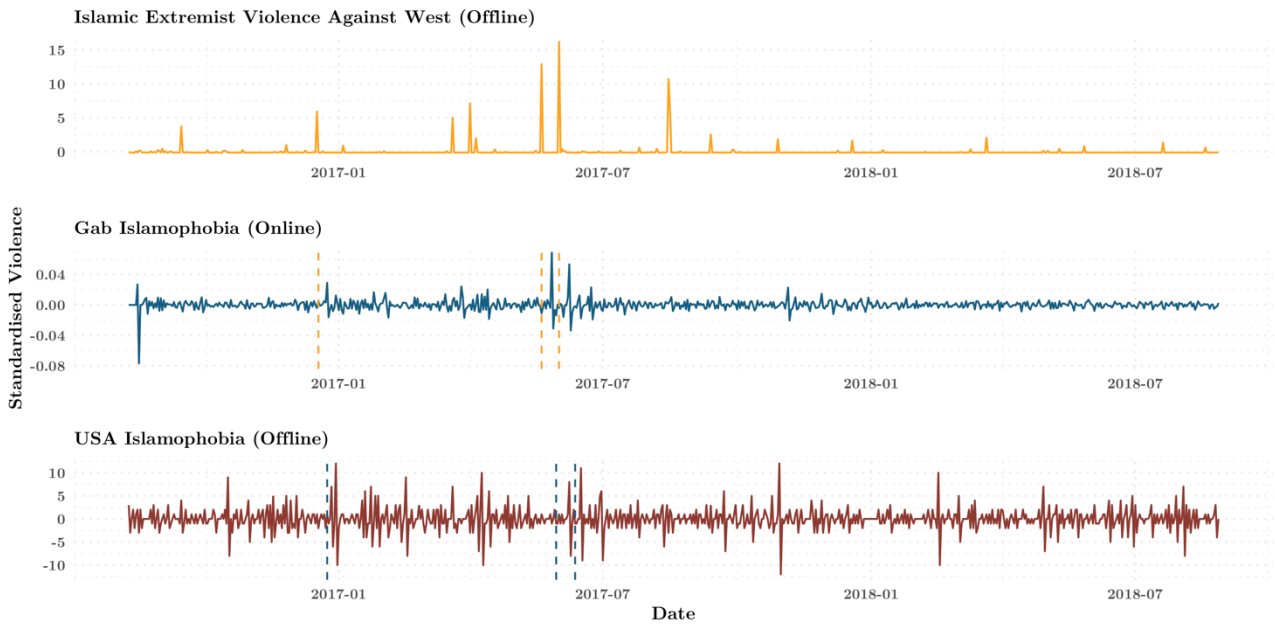


Figure 4 – Example of Granger causality between Islamic extremist offline violence (yellow), Islamophobia on Gab (blue), and offline Islamophobia hate crime (red) in the USA. Vertical dotted lines highlight temporal differences in patterns, with spikes in offline Islamic extremist violence preceding spikes in online islamophobia, which themselves precede spikes in Islamophobic hate crimes.

Table 1 – Two-way Granger causality effects for online hate speech and offline violence across combined and specific online far-right hate types. Sections 1:3 relate to far-right attacks in the US, UK, and Germany; in these sections hate types relate to both the motivation of the attack and the types of online hate speech. The final section relates to Islamic extremist attacks against the 'west', the hate types relate only to online hate types. Hate types are ranked by F statistic for each country and test, darker yellow indicates a larger estimate of F (stronger signal) while darker blue indicates a lower p-value.

	Online Hate → Offline Hate				Offline Hate → Online Hate		
HATE TYPE	DF	ESTIMATE	P	HATE TYPE	DF	ESTIMATE	P
USA - Far-right violence							
Anti-semitism	741	7.107	0.008	Homophobia	740	3.656	0.056
All hate types	747	6.800	0.009	Misogyny	733	1.838	0.120
Anti-immigration	741	5.210	0.023	Anti-black racism	733	1.399	0.233
Islamophobia	733	3.095	0.003	All hate types	745	1.313	0.270
Misogyny	739	2.277	0.132	Islamophobia	735	1.306	0.252
Homophobia	736	1.307	0.271	Anti-semitism	739	0.877	0.416
Anti-black racism	737	0.732	0.481	Anti-immigration	739	0.636	0.530
UK - Far-right violence							
All hate types	747	4.943	0.026	Islamophobia	740	3.021	0.017
Islamophobia	746	4.320	0.038	Misogyny	731	1.971	0.081
Homophobia	732	3.701	0.003	Anti-semitism	728	1.894	0.068
Anti-immigration	738	2.995	0.030	Anti-black racism	738	1.560	0.211
Anti-semitism	734	1.621	0.167	All hate types	735	1.250	0.273
Anti-black racism	740	0.759	0.384	Anti-immigration	732	1.025	0.408
Misogyny	733	0.497	0.738	Homophobia	740	0.160	0.689
Germany - Far-right violence							
Islamophobia	745	3.664	0.392	Anti-black racism	737	4.661	0.068
Anti-immigration	737	1.984	0.690	Misogyny	727	1.768	0.544
All hate types	746	1.195	1.000	Anti-immigration	729	1.581	0.689
Anti-black racism	727	1.126	1.000	Anti-semitism	732	0.940	1.000
Homophobia	740	1.104	1.000	Islamophobia	741	0.706	1.000
Anti-semitism	740	0.719	1.000	Homophobia	740	0.141	1.000
Misogyny	737	0.279	1.000	All hate types	746	0.035	1.000
Global - Islamic extremist attacks against the 'West'							
Islamophobia	737	2.925	0.089	Islamophobia	741	82.784	0.000
Homophobia	728	2.276	0.161	All hate types	742	41.982	0.000
All hate types	738	2.038	0.357	Anti-immigration	741	9.932	0.008
Mysogyny	739	1.481	0.896	Mysogyny	735	4.819	0.010
Anti-immigration	729	1.174	0.947	Homophobia	734	3.068	0.048
Anti-semitism	732	1.007	0.947	Anti-semitism	740	1.461	0.454
Anti-black racism	739	0.415	0.947	Anti-black racism	735	0.810	0.489

Offline-online relationship adjusted for local Gab interest

To test whether these effects were more significant in areas with greater Gab usage, we used data on Google search interest for 'Gab' to make local-level adjustments to the 'violence' timeseries. Locally-adjusted effects of Granger Causality between online hate speech and far-right offline hate crime were more significant in regions with higher Gab interest in the USA and UK, however again we observed no effects in Germany (Table 2).

Both Islamophobic violence and Islamic extremist violence precede online Islamic extremist interest

When testing for two-way Granger causality between the offline far-right hate crime and Google interest (search trends) for Islamic extremist content we observed significant Granger causality in the USA and the UK. In both countries overall hate crime levels preceded Google interest, while there was no relationship for the reverse (Table 3). When looking at hate crimes against specific targeted groups we found that this effect was specific to Islamophobic hate crimes in the USA, while in the UK the effect was most significant for Islamophobic hate crimes but also significant for anti-immigration hate crimes and anti-Semitic hate crimes. The results for all hate crime types are given in SI table 2. Investigations of local state-wide effects in the USA pointed towards potential state-specific relationships, but trends were not sufficiently clear to make a robust conclusion (SI Section 5).

Table 2 – Google adjusted granger causality effects for online hate and offline violence for the USA, UK, and Germany. The difference in model performance for these internet-Gab-search-adjusted models from the unadjusted model is given in the 'Delta F' column. Darker yellow indicates a larger estimate of F (stronger signal) while darker blue indicates a lower p-value.

TEST DIRECTION	DF	ESTIMATE	P	DELTA F
USA - Far-right violence				
Offline Hate → Online Hate	747	11.215	0.001	4.415
Online Hate → Offline Hate	745	1.941	0.144	0.628
UK - Far-right violence				
Offline Hate → Online Hate	747	6.018	0.014	1.075
Online Hate → Offline Hate	735	1.005	0.426	–0.245
Germany - Far-right violence				
Offline Hate → Online Hate	747	2.088	0.149	0.894
Online Hate → Offline Hate	735	0.648	0.716	0.613

We also found significant Granger causality between Islamic extremist terror attacks against the West and Google Islamic extremism interest. In all three countries (USA, UK, and Germany) increases in Islamic extremism interest occurred following increases in offline Islamic extremist violence (Table 3). These effects were most significant for the UK but significant for all three countries.

In Germany (but not elsewhere), we also found a significant Granger causality whereby the Islamic extremist Google search volumes preceded offline far-right hate crimes overall, and specifically Islamophobic hate crimes, anti-immigration attacks, anti-black racism and misogynistic hate crimes (Table 3, SI Table 2).

Table 3 – Granger causality effects between offline hate and offline Islamic extremist google search volumes for the USA, UK, and Germany. Darker yellow indicates a larger estimate of F (stronger signal) while darker blue indicates a lower p-value. Results are given for far-right hate overall and Islamophobic hate specifically, and for Islamic extremist Terror attack violence

	Offline Hate → ISIS Internet Searches			ISIS Internet Searches → Offline Hate			
HATE TYPE	DF	ESTIMATE	P	HATE TYPE	DF	ESTIMATE	P
USA - Far-right violence							
All hate types	744	8.175	0.001	Islamophobia	747	2.534	0.224
Islamophobia	745	4.702	0.019	All hate types	738	2.397	0.072
UK - Far-right violence							
Islamophobia	743	6.555	0.001	Islamophobia	739	0.999	0.417
All hate types	743	5.685	0.002	All hate types	747	0.681	0.410
Germany - Far-right violence							
Islamophobia	742	1.975	0.116	All hate types	734	4.873	0.000
All hate types	742	1.267	0.285	Islamophobia	734	3.932	0.001
Global - Islamic extremist attacks against the 'West'							
UK	738	159.296	0.000	Germany	738	1.246	0.286
Germany	744	34.545	0.000	UK	734	0.926	0.485
USA	738	12.685	0.000	USA	746	0.126	0.723

Discussion

In this study we analysed the dissemination of online hate speech on a fringe social media platform popular with the far-right, and investigated how online hate speech relates to offline hate crime, hateful extremist violence, and mutual radicalisation.

We found a positive association between online hate speech on Gab and far-right offline hate crimes, with the former preceding the latter. The effect was most significant within specific hate types—in other words online hate speech against a specific group more strongly predicted offline attacks against that group, and was more significant in areas with higher levels of interest in fringe social media platforms. Conversely, we found that hate speech against specific groups spiked following offline violence associated with connected groups. Finally, we found that offline Islamophobic violence preceded increases in online interest in Islamic extremism. Together, these results highlight the key role of the Internet in processes of mutual radicalisation and support the cyclical process presented in Figure 1, and is congruent with the idea that the online environment mediates both the instigation of offline violence, and the response to it.

Online hate predicts offline violence

We find a robust relationship between online hate speech and offline hate crimes in the USA and the UK. In both countries, increases in the proportion of online hate speech on Gab temporally preceded increases in offline far-right violence targeting minority groups. The reverse effect did not occur however, and offline far-right hate crimes did not reliably precede increases in online hate speech, demonstrating that this is a unidirectional relationship. This supports prior studies which have shown a relationship between online hateful content and offline violence (Gallacher et al., 2020; Müller & Schwarz, 2020a, 2020b; Williams et al., 2019). Our results help confirm this relationship, while also providing new insight into how this effect is strongest for certain hate types (Islamophobia & anti-Semitism) but also present for a wider range of targeted groups (homophobia, anti-immigration sentiment).

Importantly, while our use of Granger causality models to demonstrate this directionality provides stronger evidence for than simple correlative approaches would have done, this directionality itself does not necessarily imply direct causality between the online and offline hate, nor does it imply that individual users active in the online conversations were the same individuals involved in the offline

violence. Instead, this result indicates that there is something in the hate speech of online conversations which is indicative and predictive of the future violence.

There are three likely explanations for this online-offline relationship. Firstly, it has been shown that there are some individuals radicalised in these types of online conversations, for whom the conversations may act as a trigger for existing views. In these cases, the attackers express indicators of their attacks before the event (possibly instigating wider hate and encouragement from other users), then subsequently commit the attack. In this way, the online hate speech is a direct signal of the future offline violence (Meloy, Hoffman, Guldimann, & James, 2011). This pattern has been shown in a small number of high-profile and high-violence cases (e.g. Minnesota (USA), 2015; Toronto (Canada), 2017; Christchurch (New-Zealand), 2019; Poway (USA), 2019; El Paso (USA), 2019, among others), but these events are rare and this pattern is unlikely to apply to the majority of offline violence. Indeed, this is highly improbable given the relatively small userbase on Gab (~250,00 users for the period studied) when compared to the populations size of the USA and UK.

A second more plausible explanation is that Gab is a representative sample of much wider far-right communities both online and offline. Therefore, measuring the hate on Gab at any one point in time gives an estimate of the overall hateful sentiment from these wider communities. Any individual from these wider communities may engage in violence completely separately from the online conversations, but the motivating factors are mirrored in the online space. This would make hate crime occurrence similar to that of stochastic terrorism (Hamm & Spaaij, 2018; May, 2020; Woo, 2002), where random acts of ideologically-motivated violence are statistically predictable globally (from a fixed point of collection) but individually unpredictable.

A third, and related, potential explanation for this relationship is that the online environment and offline world may be separate leading and lagging indicators of a third instigation to violence – such as a political decision, a terror attack etc (e.g. Awan & Zempi, 2016). In this way, following a trigger event, hate increases quickly online, and then latterly occurs offline against the same targets. This gives the impression of the online spike causing the offline, but in reality, they are both responses to the third variable. While this third explanation is likely in some situations, a relationship between online and offline hate has been shown previously even when controlling for known trigger events (Williams et al., 2019). Similarly, in the current study we found no direct relationship between offline Islamic extremist terror attacks and offline far-right Islamophobic hate crimes in the USA, but we do

find that these events are linked by online activity (discussed below), which supports a more substantial involvement of the Internet in processes of online and offline hate.

The second of these explanations, that Gab is a representative sample of much wider far-right communities, is bolstered by the results for specific hate types individually. In addition to the overall effect of online hate speech preceding offline violence, we also found that increases in specific types on hate of Gab predicted increases in hate crimes against these same groups offline. In the USA, this relationship was significant for anti-Semitism, Islamophobia and anti-immigration, while in the UK the effect was strongest for Islamophobia but also significant for homophobia and anti-immigration. Additionally, the use of local Google search interest in Gab showed that online hate was more predictive of offline overall hate in geographical areas with higher Gab interest. Similar local effects have been shown previously in the relationship between anti-refugee sentiment on mainstream social media platforms and offline violence in Germany (Müller & Schwarz, 2020a), and the relationship between online hate and offline violence on Twitter in the USA (Müller & Schwarz, 2020b). Taken together, these results support the argument that measuring the density of hate speech on Gab can be used to gauge the violent sentiment of wider far-right groups that have similar geographical distributions.

While this relationship between online hate speech and offline hate crime was shown in both the USA and UK, we did not find it in Germany. This lack of significant relationship in Germany compared to its presence in English-speaking countries provides additional support that there is a meaningful relationship between online hate and offline violence in the USA and UK. If the effects in the UK and USA were due to hidden third variables causing both online and offline activity, we would expect to observe a similar relationship in Germany. Instead, given that our hate speech classifier was trained on English language content and will therefore not have performed for German language messages, we have not detected German users exposure to hate speech messages in German but only in English, which likely only represents a small proportion. Additionally, while Germany is the third most common country of origin for Gab users (Zannettou et al., 2018), it trails the USA and UK substantially, with only 5% of accounts belonging to German users (Zhou et al., 2019) compared to 88% from the USA or UK. This means that the sample may simply be too small to detect a relationship, and that much meaningful German fringe far-right conversations are occurring elsewhere (Guhl, Ebner, & Rau, 2020). Finally, there may be high levels of inauthentic accounts or automated activity for German accounts on Gab (Zhou et al, 2019). In future, developing cross-language hate

speech detection approaches would bring additional insight into international variation in these relationships (Vidgen & Derczynski, 2020). The fact that we observe an offline-to-offline effect of global Islamic extremist violence preceding offline far-right hate crime in Germany, as we do in the USA and UK, supports the idea that mutual radicalisation is a more global phenomenon. However, the lack of predictive power of online hate on these German attacks shows that this effect is not mediated by English language Gab content in Germany.

Offline violence predicts online hate

A key finding of our study is that in addition to far-right online hate speech predicting offline hate crimes against minority groups, we also found a significant relationship whereby offline Islamic extremist terror attacks preceded increases in online hate speech on Gab. This effect was most strongly observed for Islamophobia, indicating a reactionary response to the violence, but one which attacks the perceived outgroup as a whole (in this case Muslims) rather than focusing on the attackers themselves. In addition to Islamophobia, we also observed increases in anti-immigration sentiment, misogyny and homophobia. This supports prior evidence that offline violence can be a 'trigger event' for spikes in online hate speech (Burnap et al., 2014; Kaakinen et al., 2017; Williams & Burnap, 2016), and builds on this by demonstrating that while this reaction may be primarily focused on the perceived instigators of the violence, it also spreads across to target wider minority groups. This is suggestive of a transitive effect for intergroup conflict (Pettigrew, 2009; Chapter 3), with offline terror attacks leading to wider conflict with other groups which may be perceived to be similar or associated in some way to the perpetrator group, such as non-Muslim immigrants. In this way, Islamic terror attacks may lead to a reduction in intergroup relations between far-right groups and a range of perceived outgroups.

Importantly, our results also show that offline far-right hate crime is predictive of subsequent increases in Google search volumes for Islamic extremist material. This supports prior suggestions that offline discrimination can be an initial driver of the process of radicalisation (Dechesne, 2009; Moghaddam, 2005; Sageman, 2008). Given that the sharing of common grievances, such as experiences of prejudice and discrimination, is suggested to be a factor driving group extremism (Reicher & Haslam 2008; Smith et al, 2019) this result is particularly concerning. In the USA, this association was found only for Islamophobic hate crimes and no other type of hate, supporting the specificity of this relationship, and suggesting that it is the occurrence or perception of Islamophobic hate crime

itself which drives the relationship rather than broader trends of intolerance against multiple groups. In the UK, this association was strongest following Islamophobic hate crimes, but also occurred after anti-immigration and anti-Semitic hate crimes, suggesting a degree of overlap between targeted groups, or a general sentiment of discrimination against minority groups. Overall, this result supports prior evidence that offline Islamophobia is linked with interest in online Islamic Extremist content in a given location (Mitts, 2019) and that the geographical localisation of Islamophobic Internet searches and Islamic extremist searches is strongly associated (Bail, Merhout, & Ding, 2018). We build on this and show that this association is likely directional, with offline discrimination preceding online interest in these extremist topics.

In addition, we find a strong relationship between Islamic extremist terror attacks and Islamic extremist Google search volumes in all three study countries. This suggests that in addition to prompting reactionary hate from far-right groups, Islamic extremist terror attacks may also inspire sympathisers to investigate how to join the terror group themselves (Berger, 2014; Berger & Morgan, 2015). This result should be considered with caution however, due to the risk of false positive searches caused by interest in the terror group from across the spectrum being detected in the Google trends result (see limitations and future directions).

Role of the Internet in the Mutual Radicalisation cycle

Our results shine new light on the role of the Internet in processes of mutual radicalisation. While the process of mutual radicalisation has been demonstrated for offline intergroup conflict (e.g. Konaev & Moghaddam, 2010; Moghaddam, Harré, & Lee, 2008), we show that the Internet may have a catalyst function which facilitates this process further. Previously, users on fringe far-right platforms have been shown to respond to real world political events with greater online activity (Scrivens, 2020; Scrivens et al., 2020; Zannettou et al., 2018). Our findings expand on this and show that offline Islamic extremist violence predicts online Islamophobic hate and demonstrate a degree of mutual radicalisation between far-right and Islamic extremist groups following Islamic extremist terror attacks (Figure 1 step D). External prompts which are perceived to legitimise radical beliefs have been shown to stoke outgroup derogation (Thomas, Mcgarty, & Louis, 2014). In this way offline terror attacks may be perceived to legitimise Islamophobic attitudes leading to an increase in online hate speech towards the perceived perpetrator group.

This increase in online hate following offline events can itself instigate processes of group radicalisation and extremism as other ingroup members are exposed to this negative sentiment from their ingroup contacts and adopt this behaviour themselves through a process of 'social contagion' (Chapter 3; Ferrara, 2017; Valente, Dyal, Chu, Wipfli, & Fujimoto, 2015). This is reflected in our finding that online hate speech is associated with subsequent increases in offline hate crime across multiple targeted groups. This suggests that any reactionary online responses from far-right groups to perceived outgroup actions may also lead to further offline violence, thereby potentially creating a self-perpetuating cycle of violence (Moghadam, 2018). Prior evidence suggests that over time individual events such as verbal abuse or physical confrontation can lead to 'micro-radicalisations' (Bailey & Edwards, 2017), where negative intergroup contact between far-right extremist group members of Islamic extremist groups leads to mutual escalation. Additionally, direct antagonistic contact between opposing groups in unstructured and relatively unmoderated online spaces may also contribute to mutual radicalisation (Gallacher, Heerdink and Hewstone, 2020). Our results build on these findings and suggest that even in the absence of direct online contact between opposing groups, the Internet can play a role in facilitating the effects of mutual radicalisation, both in reaction to, and an indication of, offline violence (Figure 1 step A).

Finally, we find that Islamophobic offline violence is associated with subsequent online interest in joining Islamic extremist groups (Figure 1 step B). Whether this interest translates directly into the desire to join terror groups is somewhat unclear (Sageman, 2008). Regardless, this evidence helps us to better understand the role of the Internet in driving intergroup conflict by providing easy access to global communities to discuss shared grievances (Smith et al, 2019), and demonstrate the interplay between online and offline activity. By responding with hate speech to terror attacks, online far-right groups increase the chance that the two groups will mutually escalate, increasing the likelihood of further conflict (Figure 1 & Figure 4).

Limitations and future directions

Although we employed statistical methods to provide more robust temporally directional associations than simple correlation (Granger, 1969), this work remains observational and associative in nature, so caution should be taken not to infer strong causal mechanisms. In addition, the statistical and machine learning approaches we employ have limitations which need to be kept in mind when interpreting results.

Topic models for hate speech identification

The topic modelling approach that we use in this work to classify hate speech into distinct types assumes that each message can be classified into exactly one topic, although it may in fact contain multiple types of hate (Burnap & Williams, 2016). The conflation and combination of hate targets may play a role our observed effects of the transfer of hate from one target group to another, but whether the simultaneous expression of hate against multiple targets assists in these transfer effects of hate is unknown. As such, future research should explore this phenomenon in greater depth and include more detailed and nuanced hate type distinctions, for example through a combination of qualitative and quantitative work (e.g., measuring topic overlap in topic modelling approaches in combination with manual inspection).

Data limitations

In this study we analyse data from fringe social media platform Gab, mainly because of the availability of Gab data and how the platform historical data is stored, allowing us to study messages from the entire platform over a >2-year period. However, considering a wider range of social media platforms, and the interplay between them, would allow to for more generalisable results across topics and geographical locations, and should be an interesting avenue for future research. A challenge to overcome to work with fringe platforms popular with the online far-right such as 4chan and 8chan, or closed messaging apps such as Discord and Telegram, relates to the more ephemeral nature of these platforms. Messages and conversation threads within these platforms are deleted after a short period of time, and therefore studying these platforms and the role they play in spreading hate and division will require forward thinking data collection plans, and data contracts which preserve user anonymity and platform user agreements. Progress has been made recently in curating public datasets covering these platforms (Baumgartner et al., 2020; Papasavva, Zannettou, De Cristofaro, Stringhini, & Blackburn, 2020), but as online hate groups migrate between fringe platforms this will present an ongoing challenge. These platforms have been linked with numerous offline terror attacks (Evans, 2019), and so understanding their role in the wider online ecosystem is vital.

Another caveat in our study relates to our use Internet search trends to measure phenomena which would otherwise be very difficult to gauge, such as interest in joining specific extremist groups or local geographic distribution of Gab usage. A primary drawback of this method is that Internet search trends can have a high degree of noise or false positive association (Lazer et al., 2014). The Google

Trends platform provides estimates for phrases based on all possible combinations of keywords in each search phrase (Bail, Merhout, & Ding, 2018), so a query for the phrase 'are they evil' will return the same results as a search for 'they are evil' despite substantial differences in meaning. It is therefore possible that results will include reactionary searches from those interested in the topics from an external or observer position, as well as those genuinely interested in the content. This limitation does not affect our results linked with Google search data for "Gab" as these were for the specific social media platform, but it may affect our results linked with searches for Islamic extremist content. This could potentially explain in part the Granger causality results for an association between ISIS search volume following terror attacks, as these attacks may cause greater interest in these terror groups as well as a greater radicalising effect. Similarly, it is likely that our observed results whereby Islamic extremism search volumes predict subsequent far-right hate crimes in Germany may be caused by false positives being detected from Internet searches on Islamic extremist topics by the far-right. Future work should consider collecting multiple search topics across orthogonal dimensions to create an unbiased estimate of the search topic of interest.

Finally, there are also limitations regarding our dataset of offline hate crimes, including systematic under reporting of certain hate types. This may hinder research efforts to understand how the online environment affects these types of attacks and ultimately the implementation of adequate protective measures. For example, misogynistic attacks on women are underreported (Scott, 2020), and while domestic abuse statistics could be used as a proxy, this is an imperfect solution (Craanen, Berntsson, Ging, & DiBranco, 2020). We identified misogynistic hate speech as one of the top six most common on Gab, but it is not currently considered a protected characteristic under UK legislation (Grierson, 2020). Given a recent increase in popularity of the so-called 'manosphere' online, a conglomerate of online misogynist movements focused on "men's issues" including 'involuntary celibate' groups and 'pick-up artists', this is especially concerning. These groups have been directly linked with at least six large scale hateful extremist violent events since 2014 (Hoffman, Ware, & Shapiro, 2020) as well as numerous large-scale online harassment campaigns (Nagel, 2017). Efforts are therefore needed to address these gaps so that the negative impacts of online hate across the entire spectrum of targeted groups can be understood and then mitigated.

Conclusion

This study provides evidence for the damaging effects of online hate speech in fringe platforms popular with the far-right on intergroup conflict, extremist violence and mutual radicalisation. We show that the online environment does not occur in isolation from the offline world, instead we find a high degree of association between online and offline hate, with online hate speech preceding offline hate crimes of the same motivation, while offline events can both predict subsequent online hate speech against perceived perpetrators and interest in extremist topics from the target groups. This negative feedback loop has serious implications for intergroup relations, intergroup conflict and extremist violence. Together, our results emphasise the potential central role that Internet activity plays in the process of mutual radicalisation, which both drives far-right extremism and goes hand-in-hand with increased Islamic extremism worldwide.

References

Aasland, J. R. (2016). Right-wing terrorism and violence in Western Europe: Introducing the RTV dataset. *Perspectives on Terrorism*, *10*(3), 2–15. Retrieved from http://www.terrorismanalysts.com/pt/index.php/pot/article/view/508

Aliapoulios, M., Bevensee, E., Blackburn, J., & Cristofaro, E. De. (2021). An early look at the Parler online social network. *ArXiv*, 1–8. Retrieved from https://idrama.science/papers/parler-2021-01-07.pdf

Anoop, V. S., Asharaf, S., & Deepak, P. (2016). Unsupervised concept hierarchy learning: A topic modeling guided approach. *Procedia Computer Science*, *89*, 386–394. https://doi.org/10.1016/j.procs.2016.06.086

Anti-Defamation League. (2020). ADL H.E.A.T. Map: Hate, Extremism, Antisemitism, Terrorism. Retrieved from https://www.adl.org/education-and-resources/resource-knowledge-base/adl-heat-map

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Bail, C. A., Merhout, F., & Ding, P. (2018a). Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in U.S. counties. *Science Advances*, *4*(6), 1–10. https://doi.org/10.1126/sciadv.aao5948

Bail, C. A., Merhout, F., & Ding, P. (2018b). Using Internet search data to examine the relationship between anti-Muslim and pro-ISIS sentiment in U.S. counties. *Science Advances*, *4*(6), 1–10. https://doi.org/10.1126/sciadv.aao5948

Bailey, G., & Edwards, P. (2017). Rethinking 'radicalisation': Microradicalisations and reciprocal radicalisation as an intertwined process. *Journal for Deradicalization*, (10), 255–281.

Bastos, M. T., Mercea, D., & Charpentier, A. (2015). Tents, tweets, and events: The interplay between ongoing protests and social media. *Journal of Communication*, *65*(2), 320–350. https://doi.org/10.1111/jcom.12145

Baumgartner, J., Zannettou, S., Squire, M., & Blackburn, J. (2020). The Pushshift Telegram dataset. *ArXiv*, 1–7. Retrieved from https://arxiv.org/pdf/2001.08438.pdf

Benigni, M. C., Joseph, K., & Carley, K. M. (2017). Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter. *PLoS ONE*, *12*(12), 1–23. https://doi.org/10.1371/journal.pone.0181405

Berger, J. M. (2014). How ISIS games Twitter. *The Atlantic*. Retrieved from https://www.theatlantic.com/international/archive/2014/06/isis-iraq-twitter-social-media-strategy/372856/

Berger, J. M. (2018a). *Extremism*. Cambridge, Massachusetts: The MIT Press.

Berger, J. M. (2018b). *The Alt-Right Twitter census: Defining and describing the audience for alt-right content on Twitter*. VOX-Pol Network of Excellence. Retrieved from https://www.voxpol.eu/download/vox-pol_publication/AltRightTwitterCensus.pdf

Berger, J. M., & Morgan, J. (2015). *The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter. The Brookings Project on U.S. Relations with the Islamic World*. Retrieved from https://www.brookings.edu/research/the-isis-twitter-census-defining-and-describing-the-population-of-isis-supporters-on-twitter/

Bleher, J., & Dimpfl, T. (2019). Knitting multi-annual high-frequency Google trends to predict inflation and consumption. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3357424

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *12*(4), 421. https://doi.org/10.2307/3050792

Bliuc, A. M., Betts, J., Vergani, M., Iqbal, M., & Dunn, K. (2019). Collective identity changes in far-right online communities: The role of offline intergroup conflict. *New Media and Society*, *21*(8), 1770–1786. https://doi.org/10.1177/1461444819831779

Brandsch, J., & Python, A. (2020). Provoking ordinary people: The effects of terrorism on civilian violence. *Journal of Conflict Resolution*, 1–31. https://doi.org/10.1177/0022002720937748

Briggs, R., & Strugnell, A. (2011). *Radicalisation: The role of the Internet. A working paper of the PNN*. Retrieved from https://www.isdglobal.org

Burke, J. (2017). Rise and fall of Isis: its dream of a caliphate is over, so what now? *The Guardian*. Retrieved from https://www.theguardian.com/world/2017/oct/21/isis-caliphate-islamic-state-raqqa-iraq-islamist

Burnap, P., & Williams, M. L. (2016). Us and them: Identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science*, *5*(1). https://doi.org/10.1140/epjds/s13688-016-0072-6

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., … Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 1–14. https://doi.org/10.1007/s13278-014-0206-4

Byman, D. (2016). *Europe vs. America: Comparing the terrorism threat. Brookings Institute*. Retrieved from https://www.brookings.edu/blog/order-from-chaos/2016/04/05/europe-vs-america-comparing-the-terrorism-threat/

C-REX - Center for Research on Extremism. (2020). *Codebook: The right-wing terrorism and violence (RTV) dataset, 1990 - 2019*. Retrieved from https://www.sv.uio.no/c-rex/english/groups/rtv-dataset/rtv-codebook-revised-5.5.2020.pdf

Ceron, A., Curini, L., & Iacus, S. M. (2019). ISIS at Its apogee: The Arabic discourse on Twitter and what we can learn from that about ISIS support and foreign fighters. *SAGE Open*, *9*(1), 215824401878922. https://doi.org/10.1177/2158244018789229

Chalabi, M. (2018). Terror attacks by Muslims receive 357% more press attention, study finds | US news | The Guardian. *The Guardian*. Retrieved from https://www.theguardian.com/us-news/2018/jul/20/muslim-terror-attacks-press-coverage-study

Chan, J., Ghose, A., & Seamans, R. (2016). The internet and racial hate crime: Offline spillovers from online access. *MIS Quarterly: Management Information Systems*, *40*(2), 381–403. https://doi.org/10.25300/MISQ/2016/40.2.05

Cohen-Almagor, R. (2018). Taking North American white supremacist groups seriously: The scope and challenge of hate speech on the Internet. *International Journal for Crime, Justice and Social Democracy*, *7*(2), 38–57. https://doi.org/10.5204/ijcjsd.v7i2.517

Commision for Countering Extremism. (2019). Challenging Hateful Extremism, (October).

Conway, M. (2018). *Violent Extremism and Terrorism Online in 2018*. Retrieved from https://www.voxpol.eu/year-in-review-2018/

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2019). Disrupting Daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, *42*(1–2), 141–160. Retrieved from https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513984

Conway, M., Scrivens, R., & Macnair, L. (2019). Right-wing extremists' persistent online presence: History and contemporary trends. *ICCT Policy Brief*. https://doi.org/10.19165/2019.3.12

Coser, L. A. (1956). *The functions of social conflict*. Free Press. https://doi.org/10.4324/9780203714577

Craanen, A., Berntsson, J., Ging, D., & DiBranco, A. (2020). *Tech Against Terrorism: Incels, online misogyny and gender-based terrorism*. Retrieved from https://www.techagainstterrorism.fm/incels-online-misogyny-and-gender-based-terrorism/

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 512–515. Retrieved from http://arxiv.org/abs/1703.04009

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 11–20. https://doi.org/10.18653/v1/w18-5102

Dechesne, M. (2009). Explorations in the experimental social psychology of terrorism: The struggle-violence link and its predictors. *Revue Internationale de Psychologie Sociale*, *22*(3–4), 87–102.

Dugan, L., LaFree, G., Cragin, K., & Kasupski, A. (2008). *Building and analyzing a comprehensive open source database on global terrorist events*. Retrieved from https://www.ncjrs.gov/App/AbstractDB/AbstractDBDetails.aspx?id=245203

Ebner, J. (2017). *The rage: The vicious circle of Islamist and far-right extremism*. London: I.B.Tauris.

Evans, R. (2018). *How the MAGAbomber and the synagogue shooter were likely radicalized. Bellingcat*. Retrieved from https://www.bellingcat.com/news/americas/2018/10/31/magabomber-synagogue-shooter-likely-radicalized/comment-page-2/

Evans, R. (2019). *Shitposting, inspirational terrorism, and the Christchurch mosque massacre. Bellingcat*. Retrieved from https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/

Faris, R., Ashar, A., & Gasser, U. (2016). Understanding harmful speech online. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2882824

FBI Crime Data Explorer. (2020). Downloads & Documentation. Retrieved from https://crime-data-

explorer.fr.cloud.gov/downloads-and-docs

Ferrara, E. (2017). Contagion dynamics of extremist propaganda in social networks. *Information Sciences*, *418–419*, 1–12. https://doi.org/10.1016/j.ins.2017.07.030

Fielitz, M., Ebner, J., Guhl, J., & Quent, M. (2018). *Loving hate. Anti-Muslim extremism, radical Islamism and the spiral of polarization*. Retrieved from https://www.isdglobal.org/isd-publications/hassliebe-muslimfeindlichkeit-islamismus-und-die-spirale-gesellschaftlicher-polarisierung-deutsch/

Fuchs, B. C. (2017). Advocates warn of possible underreporting in FBI hate crime data. *NBC News*. Retrieved from https://www.nbcnews.com/news/asian-america/advocates-warn-possible-underreporting-fbi-hate-crime-data-n830711

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2020). Online contact between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society*, 1–44.

Gaudette, T., Scrivens, R., & Venkatesh, V. (2020). The role of the Internet in facilitating violent extremism: Insights from former right-wing extremists. *Terrorism and Political Violence*, 1–18. https://doi.org/10.1080/09546553.2020.1784147

Gilbert, D. (2019). Here's how big far right social network Gab has actually gotten. Retrieved from https://www.vice.com/en_us/article/pa7dwg/heres-how-big-far-right-social-network-gab-has-actually-gotten

Giles, D. (2011). *Testing for Granger Causality. Econometrics Beat*. Retrieved from https://davegiles.blogspot.com/2011/04/testing-for-granger-causality.html

Gill, P., Horgan, J., & Deckert, P. (2014). Bombing alone: Tracing the motivations and antecedent behaviors of lone-actor terrorists. *Journal of Forensic Sciences*, *59*(2), 425–435. https://doi.org/10.1111/1556-4029.12312

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Global Terrorism Index. (2019). *Global terrorism index 2019: Measuring the impact of terrorism*. Retrieved from https://www.visionofhumanity.org/wp-content/uploads/2020/11/GTI-2019-web.pdf

Google. (2020). Google Trends. Retrieved November 10, 2020, from https://trends.google.com/trends/?geo=US

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.1017/ccol052179207x.002

Grierson, J. (2020). Misogyny "should become a hate crime in England and Wales." *The Guardian*. Retrieved from https://www.theguardian.com/law/2020/sep/23/misogyny-hate-crime-england-wales-law-commission

Grover, T., & Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, 193–204.

Guhl, J., & Ebner, J. (2018). *Islamist and far-right extremists: Rhetorical and strategic allies in the digital age*. Retrieved from https://www.radicalisationresearch.org/debate/ebner-islamist-far-right-extremists-rhetorical-digital-age/

Guhl, J., Ebner, J., & Rau, J. (2020). *The online ecosystem of the German far-right*. Retrieved from https://www.isdglobal.org/wp-content/uploads/2020/02/ISD-The-Online-Ecosystem-of-the-German-Far-Right-English-Draft-11.pdf

Hamm, M. S., & Spaaij, R. (2018). The age of lone wolf terrorism. *The British Journal of Criminology*, *58*(6), 1521–1523. https://doi.org/10.1093/bjc/azy022

Heath, A. (2017). Facebook removed the event page for white nationalist "Unite the Right" rally in Charlottesville one day before it took place. *Business Insider*. Retrieved from https://www.businessinsider.com/facebook-removed-unite-the-right-charlottesville-rally-event-page-one-day-before-2017-8?op=1&r=US&IR=T

Helmus, T. C., & Klein, K. (2018). *Assessing outcomes of online campaigns countering violent extremism: A case study of the redirect method*. Retrieved from https://www.rand.org/pubs/research_reports/RR2813.html

Hoffman, B., Ware, J., & Shapiro, E. (2020). Assessing the threat of incel violence. *Studies in Conflict and Terrorism*, *43*(7), 565–587. https://doi.org/10.1080/1057610X.2020.1751459

Hothorn, T., Zeileis, A., Farebrother, R. W., Cummins, C., Millo, G., & Mitchell, D. (2020). Package "lmtest": Testing Linear Regression Models. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/lmtest/index.html

Institute for Global Change. (2016). *In their own words: Why ISIS hates the West*. Retrieved from https://institute.global/policy/their-own-words-why-isis-hates-west

Ischinger, W. (2020). *Munich Security Report 2020*. Retrieved from https://securityconference.org/en/

Javed, J., & Miller, B. (2019). When content promotes hate: Moral-emotional content, outgroup cues, and attitudes toward violence and anti-muslim policies. *ArXiv*, 1–29. Retrieved from http://www.jeffreyjaved.com/uploads/8/6/1/2/86128854/jm_when_content_promotes_hate_041919.pdf

Jawaid, A. (2013). Pakistan's polio workers targeted for killing. *Aljazeera*. Retrieved from https://www.aljazeera.com/features/2013/12/17/pakistans-polio-workers-targeted-for-killing/

Johnson, N. F., Zheng, M., Vorobyeva, Y., Gabriel, A., Qi, H., Velasquez, N., … Wuchty, S. (2016). New online ecology of adversarial aggregates: ISIS and beyond. *Science*, *352*(6292), 1459–1463. https://doi.org/10.1126/science.aaf0675

Jun, S. P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, *130*, 69–87. https://doi.org/10.1016/j.techfore.2017.11.009

Kaakinen, M., Oksanen, A., & Räsänen, P. (2017). Did the risk of exposure to online hate increase after the November 2015 Paris attacks? A group relations approach. *Computers in Human Behavior*, *78*, 90–97.

Kelion, L. (2020). Mast fire probe amid 5G coronavirus. *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-england-52164358

Kleinberg, B., Vegt, I. Van Der, & Gill, P. (2020). The temporal evolution of a far-right forum. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-020-00064-x

Knott, K., Lee, B., & Copeland, S. (2018). *Briefings: Reciprocal radicalisation*. Retrieved from https://crestresearch.ac.uk/resources/reciprocal-radicalisation/

Konaev, M., & Moghaddam, F. M. (2010). Mutual Radicalization: Bush, Ahmadinejad, and the '"Universal"' cycle of out-group threat/in-group cohesion. In *Words of Conflict, Words of War*. Praeger.

Kteily, N., Hodson, G., & Bruneau, E. (2016). They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization. *Journal of Personality and Social Psychology*, *110*(3), 343–370. https://doi.org/10.1037/pspa0000044

Kuhn, M., Wickham, H., & RStudio. (2020). Package 'tidymodels': Easily Install and Load the "Tidymodels" Packages. *CRAN*, 1–5. Retrieved from https://cran.r-project.org/web/packages/tidymodels/index.html

Lazer, D., Kennedy, R., King, G., & Alessandro, V. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*, 1203–1205. Retrieved from www.sciencemag.orgSCIENCEVOL34314MARCH2014

Lee, B. (2019). *Overview of the far-right. Centre for Research and Evidence on Security Threats (CREST)*. Retrieved from https://www.gov.uk/government/publications/overview-of-the-far-right

Li, C., Chen, S., Xing, J., Sun, A., & Ma, Z. (2019). Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems*, *37*(1). https://doi.org/10.1145/3238250

Lima, L., Reis, J. C. S., Melo, P., Murai, F., Araujo, L., Vikatos, P., & Benevenuto, F. (2018). Inside the right-leaning echo chambers: Characterizing gab, an unmoderated social system. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*, 515–522. https://doi.org/10.1109/ASONAM.2018.8508809

Massicotte, P., & Eddelbuettel, D. (2020). Package "gtrendsR": Perform and Display Google Trends Queries. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/gtrendsR/gtrendsR.pdf

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2018). Spread of hate speech in online social media. Retrieved from http://arxiv.org/abs/1812.01693

Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019). Spread of hate speech in online social media. *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*, 173–182. https://doi.org/10.1145/3292522.3326034

May, R. P. (2020). Early warning? Opportunities and limits of automated internet monitoring. *Global Network on Extremism and Technology*. Retrieved from https://gnet-research.org/2020/05/14/early-warning-opportunities-and-limits-of-automated-internet-monitoring/

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

Meloy, R. J., Hoffman, J., Guldimann, A., & James, D. (2011). The role of warning behaviors in threat assessment: An exploration and suggested typology. *Behavioral Sciences & the Law*, *28*(2), 211–223. https://doi.org/10.1002/bsl

Miller, E., LaFree, G., & Dugan, L. (2018). Global Terrorism Database (GTD). *National Consortium for the Study of Terrorism and Responses to Terrorism*. https://doi.org/10.5260/chara.19.3.14

Mitts, T. (2019). From isolation to radicalization: Anti-muslim hostility and support for ISIS in the west. *American Political Science Review*, *113*(1), 173–194. https://doi.org/10.1017/S0003055418000618

Moghaddam, F. M. (2005). The staircase to terrorism a psychological exploration. *American Psychologist*, *60*(2), 161–169. https://doi.org/10.1037/0003-066X.60.2.161

Moghaddam, F. M. (2018). *Mutual radicalization: How groups and nations drive each other to extremes*. Washington: American Psychological Association. https://doi.org/10.1037/0000089-000

Moghaddam, F. M., Harré, R., & Lee, N. (2008). *Global conflict resolution through positioning analysis*. Springer New York. https://doi.org/10.1007/978-0-387-72112-5_1

Moonshot CVE. (2018). *Searching for hate in America*. Retrieved from http://moonshotcve.com/searching-for-hate-in-america/

Müller, K., & Schwarz, C. (2020a). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *00*(0), 1–37. https://doi.org/10.1093/jeea/jvaa045

Müller, K., & Schwarz, C. (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *SSRN Electronic Journal*, 1–47. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

Nagle, A. (2017). *Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and The Alt-Right*. Zero Books.

Obaidi, M., Thompson, L., & Bergh, R. (2019). "They think we are a threat to their culture": Meta-cultural threat fuels willingness and endorsement of extremist violence against the cultural outgroup. *International Journal of Conflict and Violence (IJCV)*, *12*. https://doi.org/10.4119/UNIBI/ijcv.647

Olteanu, A., Castillo, C., Boy, J., & Varshney, K. R. (2018). The effect of extremist violence on hateful speech online. *ArXiv*. Retrieved from http://arxiv.org/abs/1804.05704

Papasavva, A., Blackburn, J., Stringhini, G., Zannettou, S., & de Cristofaro, E. (2020). Is it a qoincidence?: A first step towards understanding and characterizing the qanon movement on Voat.co. *ArXiv*. Retrieved from https://arxiv.org/pdf/2009.04885.pdf

Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., & Blackburn, J. (2020). Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. *ICWSM*. Retrieved from http://arxiv.org/abs/2001.07487

Pettigrew, T. F. (2009). Secondary transfer effect of contact: Do intergroup contact effects spread to noncontacted outgroups? *Social Psychology*, *40*(2), 55–65. https://doi.org/10.1027/1864-9335.40.2.55

Pratt, D. (2015). Islamophobia as reactive co-radicalization. *Islam and Christian–Muslim Relations*, *26*(2), 205–218. https://doi.org/10.1080/09596410.2014.1000025

Reicher, S. D., & Haslam, S. A. (2016). Fueling Extremes. *Scientific American Mind*, *27*(3), 34–39. https://doi.org/10.1038/scientificamericanmind0516-34

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, *2*(3), 1313–1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Ritchie, H., Hasell, J., Appel, C., & Roser, M. (2013). Terrorism. Retrieved from https://institute.global/policy/their-own-words-why-isis-hates-west

Sageman, M. (2008). *Leaderless Jihad: Terror networks in the twenty-first century*. University of Pennsylvania Press.

Saleem, H. M., Dillon, K. P., Benesch, S., & Ruths, D. (2017). A web of hate: Tackling hateful speech in online social spaces. *ArXiv*. Retrieved from http://arxiv.org/abs/1709.10159

Samaratunge, S., & Hattotuwa, S. (2014). *Liking Violence: A study of hate speech on Facebook in Sri Lanka*. Retrieved from https://www.cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf

Schuurman, B., Bakker, E., Gill, P., & Bouhana, N. (2018). Lone Actor Terrorist Attack Planning and Preparation: A Data-Driven Analysis. *Journal of Forensic Sciences*, *63*(4), 1191–1200. https://doi.org/10.1111/1556-4029.13676

Schuurman, B., Lindekilde, L., Malthaner, S., O'Connor, F., Gill, P., & Bouhana, N. (2019). End of the lone wolf: The typology that should not have been. *Studies in Conflict and Terrorism*, *42*(8), 771–778. https://doi.org/10.1080/1057610X.2017.1419554

Schwencke, K. (2017). Why America fails at gathering hate crime statistics. *ProPublica*. Retrieved from https://www.propublica.org/article/why-america-fails-at-gathering-hate-crime-statistics

Scott, J. (2020). Misogyny: Women "should be protected" under hate crime laws. *BBC News*. Retrieved from https://www.bbc.co.uk/news/uk-politics-54254541?fbclid=IwAR0vLQhcheQkKVcEHGlj2NzEo2h9Pd7IYE_IZ65AIHqTPh-athCv6n7rWRM

Scrivens, R. (2020). Exploring radical right-wing posting behaviors online. *Deviant Behavior*, 1–15. https://doi.org/10.1080/01639625.2020.1756391

Scrivens, R., Burruss, G. W., Holt, T. J., Chermak, S. M., Freilich, J. D., & Frank, R. (2020). Triggered by defeat or victory? Assessing the impact of presidential election results on extreme right-wing mobilization online. *Deviant Behavior*, (August). https://doi.org/10.1080/01639625.2020.1807298

Siegel, A. A. (2020). Online hate speech. In *Social media and democracy: The state of the field and prospects for reform* (pp. 56–88). Cambridge University Press.

Siegel, A., Tucker, J., Nagler, J., & Bonneau, R. (2017). Socially mediated sectarianism. *Unpublished Manuscript*. Retrieved from https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_Sectarianism_January2017.pdf

Smith, L. G. E., Blackwood, L., & Thomas, E. F. (2019). The need to refocus on the group as the site of radicalization. *Perspectives on Psychological Science*, *15*(2), 327–352. https://doi.org/10.1177/1745691619885870

Stecklow, S. (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*. Retrieved from https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Stein, A. A. (1976). Conflict and cohesion: A review of the litrature. *Journal of Conflict Resolution*, *20*(1), 143–165. https://doi.org/10.18574/nyu/9780814786390.003.0007

Stephan, W. G., & Stephan, C. W. (2000). An integrated threat theory of prejudice. In *Reducing prejudice and discrimination* (pp. 23–45). Lawrence Erlbaum Associates Publishers.

Stephan, W. G., Ybarra, O., & Morrison, K. R. (2009). *Handbook of prejudice, stereotyping, and discrimination*. Psychology Press. Taylor and Francis Group. Retrieved from https://archive.org/details/handbookprejudic00nels

Stephens-Davidowitz, S. (2017). *Everybody lies: Big data, new data, and what the internet can tell us about who we really are*. Dey Street Books.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Tajfel, H., & Turner, J. (1979). An integrative theory of intergroup conflict. In M. A. Hogg & D. Abrams (Eds.), *Key readings in social psychology. Intergroup relations: Essential readings* (pp. 94–109). Psychology Press.

Tech Against Terrorism. (2019). *Analysis: The use of open-source software by terrorists and violent extremists*. Retrieved from https://www.techagainstterrorism.org/2019/09/02/analysis-the-use-of-open-source-software-by-terrorists-and-violent-extremists/

The National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2020). Global Terrorism Database. Retrieved November 10, 2020, from https://project-iris.app-staging.cloud/

Thomas, E. F., Mcgarty, C., & Louis, W. (2014). Social interaction and psychological pathways to political engagement and extremism. *European Journal of Social Psychology*, *44*(1), 15–22. https://doi.org/10.1002/ejsp.1988

Thornton, D. L., & Batten, D. S. (1985). Lag-length selection and tests of Granger Causality between money and income. *Journal of Money, Credit and Banking*, *17*(2), 164. https://doi.org/10.2307/1992331

Tien, J. H., Eisenberg, M. C., Cherng, S. T., & Porter, M. A. (2020). Online reactions to the 2017 'Unite the right' rally in Charlottesville: Measuring polarization in Twitter networks using media followership. *Applied Network Science*, *5*(1). https://doi.org/10.1007/s41109-019-0223-3

Torok, R. (2013). Developing an explanatory model for the process of online radicalisation and terrorism. *Security Informatics*, *2*(1), 1–10. https://doi.org/10.1186/2190-8532-2-6

Tseng, Q. (2019). Reconstruct Google trends daily data for extended period. *Towards Data Science*. Retrieved from https://towardsdatascience.com/reconstruct-google-trends-daily-data-for-extended-period-75b6ca1d3420

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell.

Turse, N. (2019). How this pastor of a megachurch is fueling Ebola conspiracy theories. *Time Magazine*. Retrieved from https://time.com/5703662/ebola-conspiracy-theories-congo/

Valente, T. W., Dyal, S. R., Chu, K. H., Wipfli, H., & Fujimoto, K. (2015). Diffusion of innovations theory

applied to global tobacco control treaty ratification. *Social Science and Medicine*, *145*, 89–97. https://doi.org/10.1016/j.socscimed.2015.10.001

Vidgen, B., & Derczynski, L. (2020). Directions in abusive language training data: Garbage in, garbage out. *ArXiv*, 1–26. Retrieved from http://arxiv.org/abs/2004.01670

Vidgen, B., Harris, A., Cowls, J., & Guest, E. (2020). *An agenda for research into online hate*. Retrieved from https://www.turing.ac.uk/sites/default/files/2020-10/an_agenda_for_research_into_online_hate.pdf

Vidgen, B., Tromble, R., Harris, A., Hale, S., Nguyen, D., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. *3rd Workshop on Abusive Language Online*, 1–14. Retrieved from https://www.aclweb.org/anthology/W19-3509/

Vidgen, B., Yasseri, T., & Margetts, H. (2019). Trajectories of Islamophobic hate amongst far right actors on Twitter. *ArXiv*, 1–20. Retrieved from https://arxiv.org/pdf/1910.05794

Wendling, M. (2016). Do terror attacks in the Western world get more attention than others? *BBC Trending*. Retrieved from https://www.bbc.co.uk/news/blogs-trending-35886051

Wickham, H., Averick, M., Bryan, J., Chang, W., D', L., Mcgowan, A., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, M. L., & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Woo, G. (2002). Quantitative terrorism risk assessment. *Journal of Risk Finance*, *4*(1), 7–14. https://doi.org/10.1108/eb022949

Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). What is Gab? A bastion of free speech or an alt-right echo chamber? *ArXiv*. https://doi.org/10.1145/3184558.3191531

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., … Blackburn, J. (2017). The web centipede: Understanding how web communities influence each other through the lens of mainstream and alternative news sources. *Proceedings of the 2017 Internet Measurment Conference*. Retrieved from http://arxiv.org/abs/1705.06947

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. *ArXiv*. Retrieved from http://arxiv.org/abs/2005.07926

Zhou, Y., Dredze, M., Broniatowski, D. A., & Adler, W. D. (2019). Elites and foreign actors among the alt-right: The Gab social media platform. *First Monday*. https://doi.org/10.5210/fm.v24i9.10062

Supplementary Information (SI) for Chapter 4:

Mutual radicalisation of opposing extremist groups via the Internet

1 - Offline violence data collection

In order to estimate the level of offline hate crime over the study period, we collected data from a number of sources and combined these datasets together. SI Table 1 shows the data sources used, along with the number of events contributed from each dataset. In all cases the events were de-duplicated across sources to ensure that no event was double counted, this deduplication process was performed by manually checking events which were reported to have occurred on the same day and location and retaining only a single entry for events which appeared twice.

Far-right hate crime and violence

For the USA, we collected data on far-right terror attacks from the Global Terrorism Database (GTD, Miller, LaFree, & Dugan, 2018) and from Wikipedia. The Global Terrorism Database (GTD) is an open-source database including information on Global terrorist events since 1970 and is updated through to the end of 2018, therefore covering the period of interest for the current study (August 2016 – October 2018). For Wikipedia, we used an automated collection technique to collect all events listed on the global 'List of terrorist incidents' page. In both cases we filtered the datasets by location and our time period of interest.

USA hate crime data was collected primarily from the Federal Bureau of Investigation (FBI) dataset. This contains all hate crime reported to the FBI in the USA and gives detailed information on the type of attack, motivation or the attacker, and details of the victim(s). As this dataset is known to under-report certain hate crimes (Fuchs, 2017; Schwencke, 2017), we additionally collected

community-reported data on hate crimes. Community reported data helps to partially solve the problems of under-reporting, as victims, or observers, of hate crime can report these directly to the civil society organisation, which in turn collates this information and makes it publicly available. In addition, often other open-source information is used by organisations compiling these datasets to build a picture of the level of ongoing hate crimes in a region, including news and media reports, victim reports, extremist-related sources, as well as direct investigations (Anti-Defamation League, 2020). Due to the specific interests and focus areas of different civil society organisations, they may still over/under report certain types of hate crime, and this should be considered when interpretating the results.

For the USA, these community reports were sourced from the Anti-defamation league (ADL) and the Organization for Security and Co-operation in Europe (OSCE), which collates numerous reports from civil society organisations including the United Nations and the Council of Europe amongst others. All data is available at the daily level. In the case of the OSCE data, we did not include hate crime reports from official government statistics, only from 'other' sources in order to avoid duplicate reports, although it should be noted that if a hate crime is reported both to a civil society organisation and separately to the police, but with insufficient information for manual deduplication or with different date/location information then this event may appear in both datasets.

For the United Kingdom and Germany, we collected data from a similar range of sources, with the GTD and Wikipedia providing information on larger scale terror attacks and more notable hate crimes. We also included data from the Right-wing terrorism and violence in Western Europe dataset (RTV) collated by the Centre for Research on Extremism (CREX) (Ravndal, 2016). In the case of the RTV data, this also included plots which were foiled prior to the attack being carried out, which adds to the richness of the data. As with the USA, we supplemented this data with community-reported data on hate crimes from the OSCE, whilst excluding OCSE data from official government sources to minimise duplication of events.

Far-right motivation classification

Ensuring that all the events included in this analysis had a far-right motivation is challenging. In the case of the larger and more notable events, we included an attack in the analysis if the prosecuting authority, police, or media indicated a far-right, nationalist or white supremacist motivation. For the community-reported data, we included an attack if the victim type matched a typical far-right

targeted outgroup. It is therefore possible that this latter data includes attacks without a far-right motivation. To mitigate for this in the UK, we made sure to remove any entries which referred to Irish separatist / Irish sectarian violence, while in Germany we also removed any left-wing attacks and attacks attributed to the Kurdistan Workers' Party.

In order to classify the events into specific hate types (e.g. Islamophobia, anti-Semitism etc) we used a combination of official designation and inference. In the case of the FBI dataset for example, the hate motivation was provided through both the offender bias and the victim designation and so these classifications were used. The OCSE and ADL data was also broken down by hate motivation. In the case of the GTDB, C-Rex and Wikipedia data, the hate type was extracted from the description of the event, and details on offender background, motivation and history, along with details on the victims. This process of classification of events by hate type was performed manually.

Hate crimes could be classified as belonging to multiple categories and these are not mutually exclusive (e.g. an attack on women at a Mosque would be classified as Islamophobia and Misogyny). Matching with the Centre for Research on Extremism codebook, victims were also coded reflecting perpetrator intention (e.g. if a Sikh individual was targeted because the perpetrator believed the victim was a Muslim, target group will be coded as Muslim) (C-REX - Center for Research on Extremism, 2020).

We manually checked all datasets to ensure that the descriptions of hate type were consistent across all.

Islamic extremist violence and terror attacks

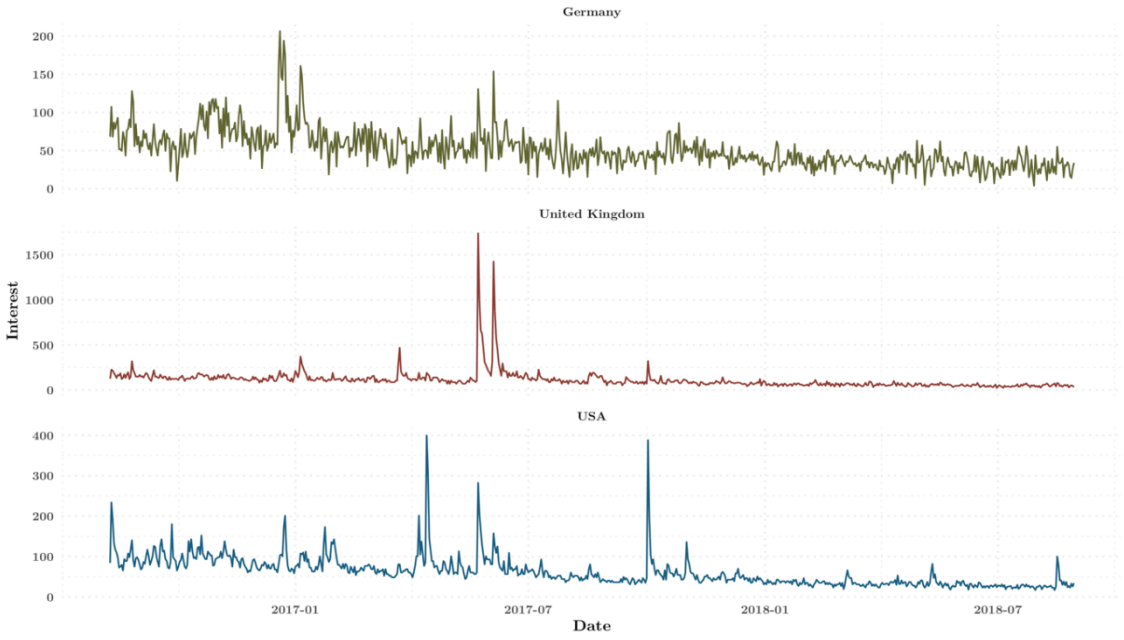
We collected Islamic extremism data from a combination of the GTD and Wikipedia. Islamic extremist terrorism consists of either direct attacks claimed by Al-Qaeda and Islamic State (ISIS), or events where the attacker either claimed to have been inspired by these groups or where authorities have formally made this connection. We restricted these events to those perpetrated in the 'West' (Europe, North America or Australasia) for reasons discussed in the main text.

SI Table 1 – Sources for offline violence collection and the number of events each source contributed to the collection after de-duplication

Country	*Source*	*Number of Events*	*Link to source*
United States of America	Global Terrorism Database	188	https://www.start.umd.edu/gtd/
	Wikipedia	2	https://en.wikipedia.org/wiki/List_of_terrorist_incidents
	FBI Hate Crime	18,534	https://crime-data-explorer.fr.cloud.gov/downloads-and-docs
	ADL	1,190	https://www.adl.org/education-and-resources/resource-knowledge-base/adl-heat-map
	OCSE – Incidents reported by 'other' sources	89	https://hatecrime.osce.org/united-states-america
United Kingdom	Global Terrorism Database	54	https://www.start.umd.edu/gtd/
	Wikipedia	1	https://en.wikipedia.org/wiki/List_of_terrorist_incidents
	OCSE	401	https://hatecrime.osce.org/united-kingdom
	C-Rex RTV	14	https://www.sv.uio.no/c-rex/english/topics/online-resources/rtv-dataset/index.html
Germany	Global Terrorism Database	43	https://www.start.umd.edu/gtd/
	Wikipedia	7	https://en.wikipedia.org/wiki/List_of_terrorist_incidents
	OCSE – Incidents reported by 'other' sources	470	https://hatecrime.osce.org/germany
	C-Rex RTV	41	https://www.sv.uio.no/c-rex/english/topics/online-resources/rtv-dataset/index.html
Islamic extremist Terrorism	Global Terrorism Database	92	https://www.start.umd.edu/gtd/
	Wikipedia	25	https://en.wikipedia.org/wiki/List_of_terrorist_incidents

2 - Google search volumes for 'How to join ISIS'

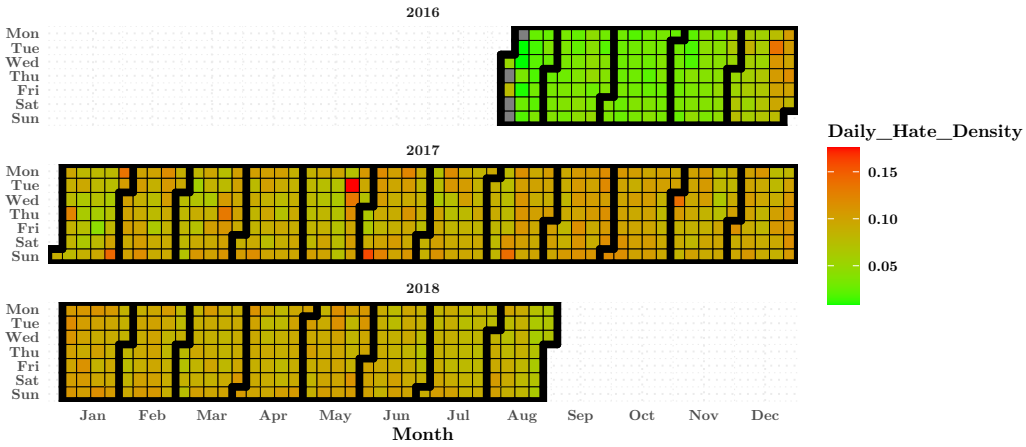
SI Figure 1 shows the Google search trends for the phrase 'How to join ISIS' for Germany, the United Kingdom and the United States of America (USA). In each case the data was collected for a series of 9-month overlapping periods and then normalised across the entire period (see methods section), which gives rise to the differential upper bounds for each country.



SI Figure 1 – Google search trends for the phrase 'How to join ISIS' for Germany, the United Kingdom and the United States of America (USA)

3 - Overall density of hate speech on Gab

SI Figure 2 gives the overall density of hate speech within the Gab conversations over the period August 2016 – October 2018. The density rose quickly over the first 4 months before remaining fairly level but with significant variation from day-to-day.



SI Figure 2 – Daily hate density for Gab over the period August 2016 – October 2018

4 - Wider effects of violence on ISIS search volumes

SI Table 2 shows the Granger causality effects for offline hate and offline Islamic extremist Internet interest for the USA, UK, and Germany across specific hate types not outlined in the main results. Darker yellow indicates a larger F statistic (stronger signal) while darker blue indicates a lower p-value.

We observed no effects for any hate types other than Islamophobia in the United States of America. In the UK we observed effects for anti-immigration sentiment and anti-Semitism, but these were less significant than for Islamophobia.

In Germany we observed an association between offline anti-Black hate crimes and subsequent Islamic extremism Google interest, but not for any other offline hate types. We do not read too far into this result and consider that it might be a spurious false positive. We do however observe interesting results whereby the Google search volumes for Islamic extremist content in Germany consistently preceded hate crimes across anti-immigration, misogynistic and anti-Black racism motivations.

SI Table 2 – Granger causality effects for offline hate and offline Islamic extremist google search volumes for the USA, UK, and Germany across specific hate types. Darker yellow indicates a larger estimate of F (stronger signal) while darker blue indicates a lower p-value.

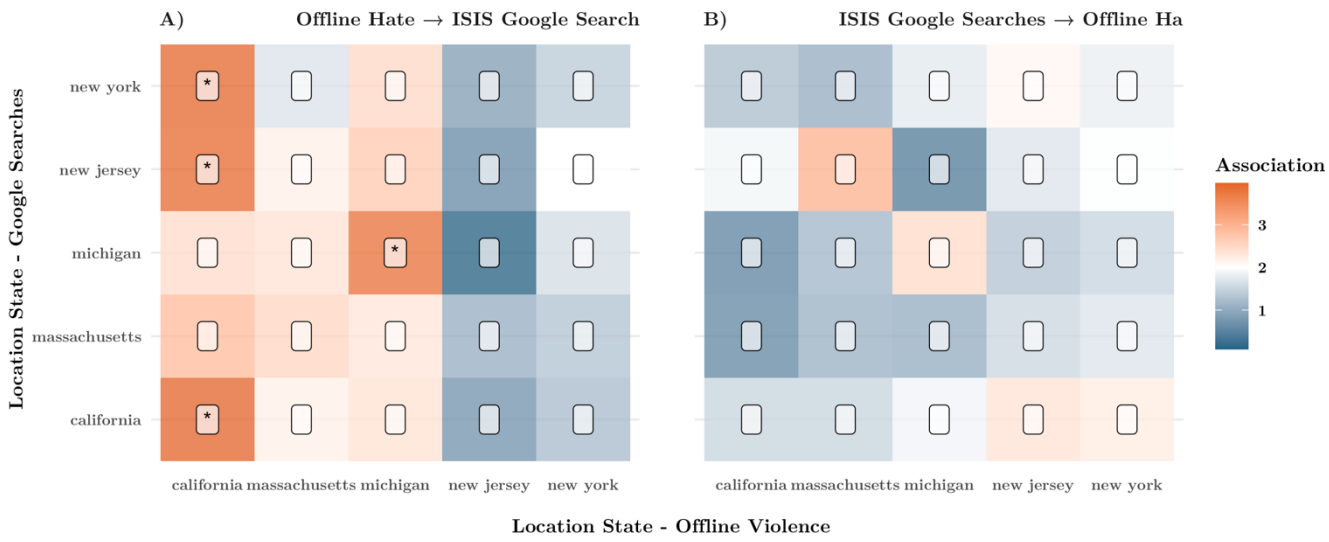
	Offline Hate → ISIS Google Searches				ISIS Google Searches → Offline Hate		
HATE TYPE	DF	ESTIMATE	P	HATE TYPE	DF	ESTIMATE	P
USA - Far-right violence							
Anti-immigration	746	3.236	0.145	Misogyny	736	1.681	0.245
Anti-black racism	746	2.809	0.188	Anti-semitism	738	1.404	0.441
Anti-semitism	746	1.262	0.523	Homophobia	736	1.377	0.443
Homophobia	740	0.780	1.000	Anti-black racism	742	1.168	0.642
Misogyny	740	0.475	1.000	Anti-immigration	742	0.896	0.886
UK - Far-right violence							
Anti-immigration	742	5.913	0.002	Anti-immigration	734	0.915	0.886
Anti-semitism	742	5.327	0.004	Anti-semitism	734	0.909	0.499
Homophobia	746	0.069	1.000	Misogyny	746	0.397	0.529
Misogyny	746	0.067	1.000	Homophobia	746	0.373	0.541
Anti-black racism	746	0.038	0.847	Anti-black racism	746	0.021	0.884
Germany - Far-right violence							
Anti-black racism	740	4.149	0.007	Anti-immigration	734	4.597	0.000
Anti-immigration	742	1.824	0.145	Misogyny	734	3.280	0.006
Misogyny	742	1.656	0.525	Anti-black racism	734	2.607	0.035
Homophobia	742	1.381	0.742	Homophobia	734	2.372	0.063
Anti-semitism	743	0.893	0.523	Anti-semitism	737	2.213	0.120

5 - Local USA estimates of Islamic extremist search interest and offline Islamophobia

In order to investigate the local effects of Islamophobic hate crime on Islamic extremist search interest we separately compared the timeseries data from the five states with the highest number of Islamophobic hate crimes in the USA over the study period. For each state we ran two-way Granger Causality tests looking at the predictive power of offline Islamophobic hate crimes on Islamic extremist Internet search interest in that state and vice-versa. We then compared the hate volumes in each of these top five states to the Islamic extremist search volumes in each of the other five states. We limited the analysis to the top five states for two reasons; firstly, because with the highest offline crime volumes these states will lead to the most reliable statistical measures, and secondly because including more states in the analysis increases the likelihood of spurious effects due to the high number of combinations of states that would be need to be calculated. In all cases we corrected for multiple comparisons using Holms corrections.

Here, we follow the method of Bail et al (2018) and excluded Washington DC due to the high likelihood that investigative journalists or intelligence personnel may type skew the Google search volumes for ISIS related content as part of research into ISIS recruitment tactics. This left the following states included in the analysis; California, Massachusetts, Michigan, New jersey and New York.

Results showed a degree of local specificity in the associations between offline hate crime and online search activity (SI Figure 3). Panel A shows how offline Islamophobic hate crime in Michigan predicted Google search volumes in that state, and only in that state, while offline crime in California predicted the Google search volume in California. The activities in these two states were mutually independent. However, the offline activity in California also predicted Google search volumes in New Jersey and New York suggesting that the results are not completely clear cut. We observed no effects in Massachusetts, and in none of the states did Google search volumes predict offline violence.



*SI Figure 3 – Local state-wide effects of Islamophobic hate crime in the USA on Islamic extremist Google search volumes for the top 5 states by Islamophobic hate. Panel A shows the Granger Causality effects of offline violence on online searches across, while panel B shown the reverse. Darker Orange indicates a larger estimate of F (stronger signal) while darker blue indicates a smaller value. Significance values are given by * for $p < 0.05$ and ** for $p < 0.01$.*

Chapter 5

How hostile Information Operations increase the polarisation, intergroup antagonism, and hate of online conversations

Gallacher, J. D., & Heerdink, M. W

Measuring the effect of Russian Internet research Agency information operations in online conversations (2019) *Defence Strategic Communication., vol. 6, p.155:198*

https://www.stratcomcoe.org/jd-gallacher-m-w-heerdink-measuring-effect-russian-internet-research-agency-information-operations

Abstract

The Internet has given new opportunities to those who wish to interfere and disrupt society through the systematic manipulation of social media. One goal of these cyber-enabled information operations is to increase polarisation in Western societies by stoking both sides of controversial debates. Whether these operations are successful remains unclear. This paper describes how novel applications of computational techniques can be used to test the impact of historical activity from the Russian Internet Research Agency (IRA) on two social media platforms: Twitter and Reddit. We show that activity originating from the Russian IRA had a measurable effect on the subsequent conversations of genuine users. On Twitter, increases in Russian IRA activity predicted subsequent increases in the degree of polarisation of the conversation surrounding the Back Lives Matter movement. On Reddit, comment threads started by Russian IRA accounts contained more toxic language and identity-based attacks than comparable threads started by genuine users – demonstrating a lower conversational quality. We use causal analysis modelling to further show that Russian IRA activity in existing threads caused measurable changes in the conversational quality metrics of the following 25-100 posts. By developing methods to measure the impact of information operations in online conversations and demonstrating a measurable effect on genuine conversations, our study provides an important step in developing effective countermeasures.

Introduction

The rapid development of the Internet has enabled people everywhere to connect, communicate, and distribute information globally at an unprecedented scale. However, some use this opportunity for connection to divide rather than to bring people together. In recent years, a great deal of attention has been focused on groups that conduct deliberate social media activities to divide and polarise societies. These activities include the use of fake social media accounts, paid advertisements, and automated scripts designed to spread of disinformation—deliberately misleading information designed to influence public opinion (Guess et al., 2018). These activities are constituents of wider information operations campaigns that seek to gain a competitive international advantage over traditional adversaries (Paul & Matthews, 2016). While the approach itself is not new—similar methods targeting the psychology of civilian populations can be traced back to the Roman, Persian, and Chinese empires (Weedon, Nuland, & Stamos, 2017) —these methods have transformed in the digital age and now increasingly rely on social media platforms that provide global reach and can target individuals directly for a fraction of the cost of traditional methods (Lucas & Pomerantsev, 2016). This

phenomenon is characterised by sustained and pervasive efforts which peak around election cycles, although elections are not the sole focus (Gallacher & Fredheim, 2019). This persistent engagement, short of traditional thresholds for conflict, makes it difficult to construct robust responses (Kello, 2017).

This rapid spread of disinformation online was identified by the World Economic Forum in 2014 as one of the top 10 perils to society (Howell et al., 2014). Since this warning, the deliberate spread of misleading information has been linked to political earthquakes such as the 2016 US Election (Intelligence Community Assesment, 2017), the 2016 UK Brexit referendum (UK Department for Digital Culture Media and Sport Committee, 2018), and the rise of populist parties across Europe (Ferrara, 2017), as well as to political violence in Brazil (Arnaudo, 2017), Myanmar (Stecklow, 2018), and India (Rani, 2018). All these events are connected by one consistent trend—an increase in social polarisation, defined as the process of increased segregation into distinct social groups, separated along racial, economic, political, religious or other lines (Castree, Kitchen, & Rogers, 2013). Hostile information operations on social media are ongoing and show no evidence of slowing down (Karan, Barojan, Hall, & Brookie, 2019), while social media platforms stand accused of failing to act decisively in combatting this threat (UK Department for Digital Culture Media and Sport Committee, 2019). Understanding the consequences of these activities is essential to developing effective defences. In-depth knowledge about the consequences of these hostile narratives should inform policy decisions aimed at countering them, yet very little is known about the effect these activities have on the online conversations of genuine citizens, and whether or not they achieve their goals.

In this study we developed methods to address this question and to measure the effect of artificial social media manipulation on subsequent human conversations, using publicly attributed information operations from the Russian state as a case study. Recently, evidence shows that the Russian government has been engaged in a substantial effort to sway public opinion on a number of key topics, at home and abroad, through a prolonged information campaign (Intelligence Community Assesment, 2017; Weedon et al., 2017). This campaign includes disinformation, artificial social media accounts imitating a grass-roots movement, paid advertisements, and automated scripts designed to hijack filtering algorithms in order to disseminate content to the widest possible audience (DiResta et al., 2018; Howard, Ganesh, Liotsiu, Kelly, & Camille François, 2018). These accounts also promoted real-world protests and demonstrations, often encouraging both sides of controversial topics (Bertrand, 2017). While the 2016 US presidential election seems to have been one important focus for

these activities, the wider intention appears to have been to polarise online conversations and sow social division along social as well as political lines (Bay et al., 2019; DiResta et al., 2018).

The relationship between disinformation and polarisation

People increasingly use social media as their primary source for news and information, with two-thirds of Americans and half of adults in the developing world getting their news from social media platforms (Shearer & Gottfried, 2017). Ideological alignment with specific groups and ideas is often more obvious in online environments than it is offline (Lee, 2007; Postmes, Spears, & Lea, 1998), either due to structural features, such as profile pictures or group memberships, or because of the content shared by users. For this reason, separation into groups of likeminded people is more likely to occur online than offline. This facilitates group polarisation, a social-identity based phenomenon where individuals endorse more extreme ideological positions following a discussion with other in-group members (Turner, Davidson, & Hogg, 1990).This increased polarisation may encourage group members to take a more extreme position on certain issues, or may result in an increased dislike of members of other groups without a change in issue position (Mason, 2015).

Increased polarisation can both stem from, and facilitate the spread of, disinformation online. Online content which emphasises inter-party conflict has been shown to reinforce social polarisation and is easy to distribute in online environments. Messages containing strong partisan cues that match an individual's beliefs can encourage them to accept and share inaccurate information (Garrett, Weeks, & Neo, 2016), while messages that agree with pre-held stereotypes can facilitate an individual's acceptance of inaccurate information about an out-group (Druckman, Peterson, & Slothuus, 2013; Gvirsman et al., 2014; Kosloff, Greenberg, Schmader, Dechesne, & Weise, 2010; Weeks, 2015). Equally, polarised conversations can lead to increased dissemination of disinformation. People are more likely to trust inaccurate information if it elicits anger and aligns with their existing opinions (Weeks, 2015). Content that is highly controversial or elicits greater moral outrage is more likely to be shared by social media users (Chapter 3; Brady, Wills, Jost, Tucker, & Van Bavel, 2017), while fake content can be made more sensational than true content and therefore more likely to inspire fear and disgust and propagate faster and farther through the networks (Crockett, 2017; Vosoughi, Roy, & Aral, 2018).

There is some early evidence that online environments may create 'echo chambers'—networks of like-minded people who confirm each other's opinions instead of promoting critical thinking (Conover,

Ratkiewicz, & Francisco, 2011; Yardi & Boyd, 2010)—exacerbating these effects. Disinformation spreads more quickly within these closely connected groups due to a lack of dissenting voices (Pariser, 2011). This may facilitate the creation of a society that is increasingly polarised (Sunstein, 2017) and misinformed (Vosoughi, Roy, & Aral, 2018) as people are more likely to be affected by inaccurate information if they see it more frequently, especially in cases where recent exposure influences decision-making (Berinsky, 2017; Pennycook, Cannon, & Rand, 2018).

Recently, evidence suggests that echo-chambers may not be forming as often as first expected (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015; Bright, 2018), and users are, in fact, exposed to more cross-cutting information online than they would select purely based on choice (Bakshy, Messing, & Adamic, 2015). Even so, this cross-cutting information may not have a positive effect. Users with more extreme ideological positions are more active on social media (Barberá & Rivero, 2015; Preoţiuc-Pietro, Liu, Hopkins, & Ungar, 2017) and exposure to opposing views online can also increase polarisation by highlighting areas of disagreement (Chapter 1; Bail et al., 2018). Both situations provide opportunities for those who wish to leverage the polarising effects of social media, either through infiltrating echo chambers to spread negative messages about an out-group without opposition, or by engaging with someone while posing as an out-group member in order to antagonise and create a negative impression of the out-group as a whole.

In this study we investigate the impact that the deliberate manipulation of the online environment has on the conversation of genuine users, and whether information operations from state-backed organisations cause measurable increases in online polarisation, hostility of conversations, and antagonistic intergroup relationships. We do this by using the activities of the Russian Internet Research Agency on social media as a case study.

The Russian Internet Research Agency and online polarisation

From as early as 2012, information operations conducted over social media have been targeting citizens in the West (Howard et al., 2018). These operations originate from the St Petersburg 'troll farm' run by the Russian Internet Research Agency (Russian IRA). The agency aims to influence online conversations about regional, national, and international issues that affect Russian foreign and domestic policy interests (Gerber & Zavisca, 2016; Sanovich, 2017). Online manipulation can take the form of 'trolling' orchestrated from human-controlled accounts or propaganda spread by automated accounts (bots) (Fredheim, 2019; Kumar, Cheng, Leskovec, & Subrahmanian, 2017; Woolley &

Howard, 2016). Since 2012, these campaigns have grown steadily in number and scale (Howard et al., 2018), and have gained much international attention, particularly surrounding the 2016 US presidential election (Intelligence Community Assesment, 2017).

Over the course of 2018, large, open-source datasets detailing the posts from accounts attributed the Russian IRA were published, making it possible to conduct a detailed analysis of how Russia ran these information campaigns (Hindman & Barash, 2018; Linvill & Warren, 2018). The data showed that the campaign was not restricted to the 2016 US election but rather sought to divide online groups along racial, ethnic, social, and political lines, aiming to weaken and exacerbate social vulnerabilities in Western societies, and continued long after the election was decided (Gallacher & Fredheim, 2019). Both sides of numerous controversial debates were inflamed by Russian IRA activity, especially conversations surrounding provocative race issues such as the Black Lives Matter movement in the United States.

The present research: Measuring the effect of these information operations

While the intention behind this activity is clear, measuring its impact is complex. Trolls have been shown to manipulate the opinions of users in online forums (Mihaylov, Georgiev, & Nakov, 2015) and to steer conversations on blogging platforms (Sobolev, 2018). While at times these accounts have garnered greater popularity than those of organic users (Howard et al., 2018), the impact they have on the wider online ecosystem is hard to measure. Some calculations show that Russian IRA accounts were influential in spreading targeted URLs across Twitter (Zannettou et al., 2019), but that this activity did not carry over to other web communities (Reddit, 4Chan, etc) (Zannettou et al., 2018). Twitter's key role in these campaigns is also illustrated by the fact that in the run-up to the 2016 US Election, more hyperlinks to websites hosting disinformation were shared on Twitter than the top sixteen mainstream media outlets combined (Barberá, 2018). What is not clear from this evidence however, is what effect the Russian IRA accounts have had on fuelling both sides of controversial online discussions to exacerbate online polarisation and reduce intergroup relations.

In this paper we use a two-part strategy to measure the effect of information operations on online conversations. In Part 1 we focus on a case study of the Black Lives Matter (BLM) movement which was targeted by Russian IRA accounts. This social movement has spread both online and offline to protest the systematic violence perpetrated against African-Americans, particularly by police officers

(Freelon, McIlwain, & Clark, 2016). Opposition movements to BLM (which is associated with the #BlackLivesMatter hashtag) have criticised it for failing to appreciate the value of all races (leading to the #AllLivesMatter hashtag) or for failing to respect the value of police officers and the risk they take in course of their work (represented by #BlueLivesMatter) (Stewart, Arif, Nied, Spiro, & Starbird, 2017). These hashtags can shape how information flows through the wider network and therefore play a significant role in the spreading of ideas (Gallagher, Reagan, Danforth, & Dodds, 2018; Leo G. Stewart, Arif, & Starbird, 2018). Russian IRA accounts imitated authentic users on both sides of this debate to disseminate provocative messages to various target audiences and to foster antagonism between opposing groups (Arif, Stewart, & Starbird, 2018). This is likely to have contributed to the polarisation of the #BlackLivesMatter conversation online; Russian IRA accounts were in the top percentile of retweeted accounts in both supporting and opposing sides of the Twitter conversation (Stewart et al., 2018). We investigated the global effect of the Russian IRA tweets on the entire #BlackLivesMatter conversation by testing whether the daily degree of polarisation of the Twitter conversation positively correlates with earlier Russian IRA activity surrounding the #BlackLivesMatter hashtag.

In Part 2 we look at the impact of Russian IRA activity on Reddit using natural language processing and causal impact modelling to analyse the effect of >16,000 Reddit posts attributed to the Russian IRA. We take three measures of conversation quality (integrative complexity, toxicity, and identity attacks) and use these to investigate how the nature of the online conversations that Russian IRA accounts started and participated in was altered compared to unmanipulated conversations on similar topics. Following revelations about the scope of Russian IRA manipulation of social media platforms in 2016, Reddit was the only social media company to keep this activity publicly visible on the platform rather than removing it, so it is the only platform where the immediate response to Russian IRA content can be analysed directly. We measure the response to known artificial activity and predict that Russian IRA activity causes a measurable decrease in the quality on discussion threads.

Methods

The following analysis is composed of two parts. In Part 1 we measure the polarisation of Twitter conversations to investigate how the degree of daily polarisation of the #BlackLivesMatter conversation on Twitter correlates with Russian IRA activity within this conversation. In Part 2 we look to Reddit and use natural language processing to measure the direct effect of Russian IRA activity on the conversations.

Importantly, across both parts of this study we do not make any attributions to which accounts were operated from the Russian IRA. Instead, the accounts were identified and attributed by the social media platforms themselves using information that is not available to the public.

Part 1 - Polarisation of Twitter Conversations

Data collection and sampling

Twitter is a popular social media platform built on a microblogging format. Users can share short messages, or 'tweets', with their followers who can in turn 'retweet' these messages to their own followers. Tweets can sometimes contain hashtags indicating that it is part of a broader conversation. In late 2018 Twitter averaged 321 million active monthly users (Statista, 2019).

We collected Twitter data relating to the Black Lives Matter conversation from an archive complied by the digital chronicling organisation 'Documenting the Now' (docnow.io, 2018). The dataset contains 17,292,130 tweet IDs for tweets collected from the Twitter streaming API for #BLM and #BlackLivesMatter between 29 January 2016 and 18 March 2017 (Summers, 2017). Twitter's terms of service don't allow public redistribution of tweets; however, they do allow datasets of tweet IDs to be shared. We then recovered the full tweet from each tweet ID by performing a search through the Twitter search API (also known as 'hydration') using DocNow's Hydrator software (Documenting the Now, 2019). Only tweets which were still publicly available at the time of the search could be recovered; we could not recover tweets that had been deleted by Twitter or by the users themselves. We hydrated the dataset of tweet IDs on 24 November 2018, which led to a collection of 9,531,526 tweets, or 55% of available tweet IDs (45% of the original tweets had been deleted since publication). This value is higher than might be expected from prior work (Almuhimedi, Wilson, Liu, Sadeh, & Acquisti, 2013), and may reflect the contentious nature of the topic, the delay between the conversations taking place and re-hydration, as well as Twitter's increased crackdown on automated

English-language accounts (far beyond just Russian IRA activity) over the period studied (Fredheim, 2019). While our dataset, therefore, does not represent the full conversation, it is the best approximation available given the limits that Twitter places on data sharing. Importantly, this dataset does not contain the tweets from Russian IRA, as this information was removed from the platform at point of attribution by Twitter, prior to collection. Therefore, our measure of polarisation reflects the polarisation of the conversation of genuine (i.e. non-Russian IRA) accounts without potential artificial inflation from Russian IRA tweets.

Data on the activity of known Russian IRA accounts were collected by Linvill and Warren (Linvill & Warren, 2018), and made publicly available by the team at fivethirtyeight.com (Roeder, 2018). This dataset contains 2,973,371 tweets from 2,848 Twitter accounts spanning the period from 2015–2018.

Measuring polarisation

We measured the degree of daily polarisation on Twitter using a novel application of correspondence analysis, implemented in the FactoMineR package for R (Husson, Josse, Le, & Maintainer, 2018). Correspondence analysis is a statistical method that makes it possible to map contingency tables to expose underlying relationships between objects in the data (Hirschfeld & Wishart, 1935). All analyses were performed in R (version 3.4.4, R Core Development Team 2017).

For each day of the dataset, we used a retweet matrix as the contingency table to show the relationship between active users within the dataset (rows) and tweets (columns) (see example in Table 1). A retweet matrix is a good starting point for discovering the structure of Twitter conversations, as retweets have been shown to closely represent the expression of agreement with a particular message and, under certain conditions, support of the messenger (Metaxas & Twittertrails Research Team, 2017). Given this, we assumed that if a user retweets messages expressing support or opposition for a given position, this reflects the user's own beliefs.

By summarising contingency tables via dimensionality reduction correspondence analysis allows us to identify if certain users are associated with groupings of tweets. It interprets the retweet matrix across a number of dimensions whereby the largest amount of variability in the data is captured in dimension 1, the next largest amount of variability is captured in dimension 2, the third largest

Popular Tweets ---->

<---- Active Users	μ^1	μ^2	μ^3	μ^4
ϱ^1	1	2	0	0
ϱ^2	0	0	1	0
ϱ^3	3	0	0	1
ϱ^4	1	2	0	0
ϱ^5	0	1	0	5

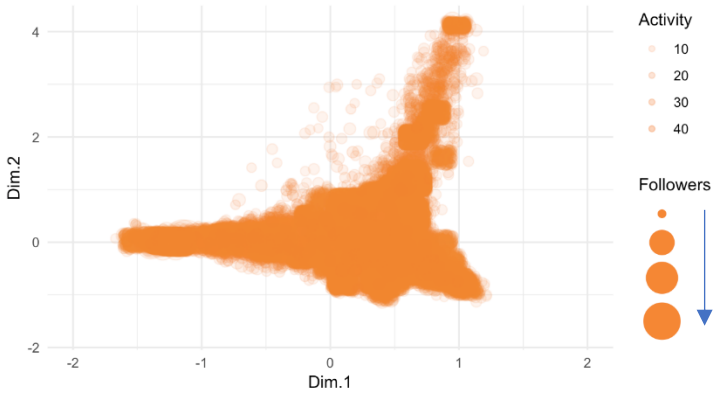


Table 1 and Figure 1 – Simplified retweet matrix for popular tweets and active users for the #BlackLivesMatter Twitter conversation on 07/07/2016 and the correspondence analysis results placing users on dimensions one and two.

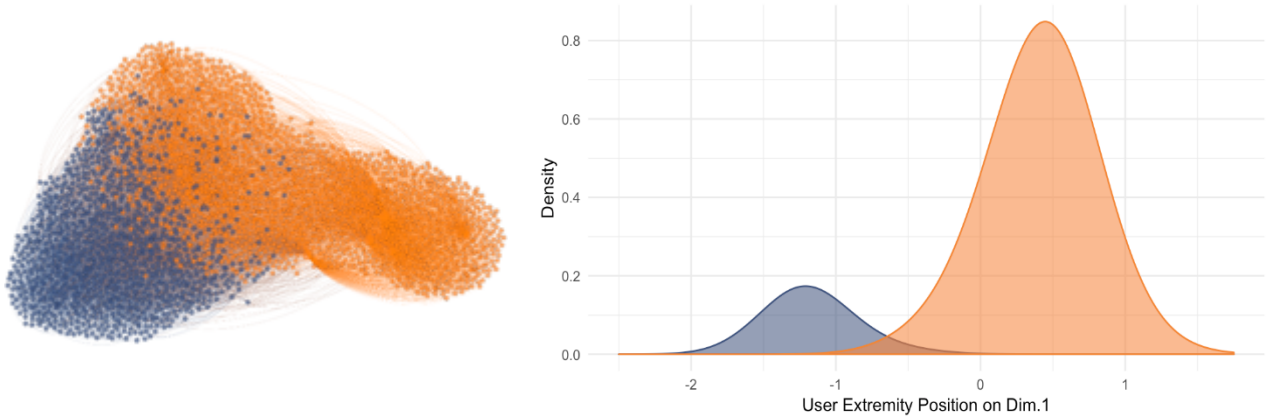


Figure 2(a) and (b) – Visualisation of the polarised retweet network and matching bimodal distribution for dimension 1 for the #BlackLivesMatter conversation on 07/07/2016

amount of variability is captured in dimension 3, etc. The scores for dimension 1 were used to calculate the position of each tweet on the dimension 1 scale in relation to the other tweets for that day. As explained below, dimension 1 generally distinguishes between tweets that were either for or against the #BlackLivesMatter movement; the greater the score in dimension 1 the stronger the support or criticism. Opposition tweets were often framed as part of a counter-movement, such as #BlueLivesMatter, co-opting BLM-related hashtags (#BlackLivesMater or #BLM) to inject opposing opinions into the conversation.

This approach has the benefit over alternative measures of polarisation of Twitter networks, such as clustering-based network community detection (Bakillah, Li, & Liang, 2015; Silva, Santana, Lobato, & Pinheiro, 2017), of allowing us to measure the position of each user in the conversation on a continuous scale and along a single dimension, and therefore relative to one-another, which facilitates better inter-user comparisons compared with discrete groupings which may struggle to differentiate between polarisation and fragmentation of conversations. Additionally, this analysis can be performed

based only on activity and interactions within the network without the need to gather any further user-level variables. This makes it easier and more robust when comparing network measures over time across multiple days.

We focused only on the dimension that demonstrated the greatest variance in the daily activity, dimension 1, because it was the most stable across multiple days and was the most reliable indication of the level of support or opposition for the Black Lives Matter movement indicated by the tweet. We verified the consistency of this dimension by taking a random sample of 50 days from the dataset and selecting the 10 tweets with the highest and lowest scoring days on dimension 1. We manually coded whether the majority of the content presented in these tweets represented opposing sides. This was the case for 85% of the days. Further inspection of the remaining 15% of days showed the most popular tweets on these days appeared to be relevant to both sides of the debate, supporting the idea that these conversations may have been more homogenous and less polarised, at least in terms of popular tweets the sides of the debate interacted with. We are therefore confident that dimension 1 was an appropriate measure for the level of support or opposition to the Black Lives Matter movement.

To successfully perform a correspondence analysis, the contingency table had to represent a well-connected subgraph of the retweet network to avoid a small subset of users, peripheral to the main conversation, generating large scores on the important dimensions (similar to the k-core within network theory). We therefore used thresholds to filter out less popular tweets (as assessed by the number of retweets) and 'inactive' users (who did not retweet many popular tweets). These thresholds depend on daily conversation size and are shown in Table 2, with higher thresholds for larger conversation sizes. For total conversation sizes below 5,000 tweets we applied a threshold of 1 for both measures, and so did not reduce the conversation. This latter case applied to 30 of the 378 days in the dataset. (See Supplementary Information (SI) Figure 1 for distribution of conversation size over days).

Table 2 – Thresholds selected for the number of retweets needed for a tweet to be 'popular', and the number of these tweets a user needs to interact with to be considered 'active'.

	Total Conversation Size (Nb of Tweets)				
	> 200,000	> 100,000	> 10,000	> 5,000	< 5000
Re-tweet threshold	10	6	4	2	1
Active user threshold	5	3	2	1	1

Selecting the correct values for these thresholds is important for achieving stable results. We selected suitable thresholds dynamically for each day according to two rules: (a) thresholds should not produce extreme scores for a subset of users on dimension 1 ($|z| > 10$), and (b) when applying back to scores from the subgraph to all users, thresholds should allow for >25% of daily users in the conversation to be classified as belonging to dimension 1. In rare cases the standard thresholds did not fit; for these days slightly lower/higher thresholds were applied. This was necessary as for some days certain tweets went 'viral', changing the relationship between conversation size and the overall activity of the average user, therefore reducing the coverage of the correspondence analysis and violating rules a and b given above. In those rare cases, while slightly altering the thresholds appropriately improved the coverage of users and avoided artificially high z scores for subsets of users for each given day, overall it did not alter results substantially.

After ranking all popular tweets along dimension 1, we used the results to estimate the dimension 1 score for each user compared to all other users, based on the average of all the tweets they had retweeted. This last step could be performed for all users who interacted with at least one 'popular tweet', not only those defined as 'active' in the earlier step of the correspondence analysis.

The spread of users across dimension 1 reflects how their opinions are distributed, and whether users formed distinct 'camps'—something we would expect if the conversation were polarised. We measured the degree of this polarisation using Hartigans' dip test (Hartigan & Hartigan, 1991; Maechler, 2015), which measures how bimodal a sample is, with higher scores indicating higher bimodality. Formally, the dip test measures bimodality in a sample by calculating the maximum difference, over all sampled points, between the observed distribution and the theoretical unimodal distribution that would minimise the difference. The higher this residual error, the less unimodal the distribution is considered to be. We operationalised polarisation as the bimodality of each daily distribution of user scores on dimension 1 (Figure 2). The more bimodal this distribution was, the more polarised the conversation.

Relation to Russian Troll Farm Activity

After measuring the degree of polarisation in the daily conversation from genuine accounts, we related it to the artificial activity originating from accounts associated with the Russian IRA using (lagged) Pearson's correlations. Russian IRA activity is measured as the number of posts using a BLM-related hashtag from the public dataset released in summer 2018.

Russian IRA activity is unlikely to have an immediate effect on the degree of polarisation of the conversation, especially as the direct responses to this activity were unavailable. To measure the correlation between Russian IRA activity and the subsequent level of polarisation in the conversation, taking into account cumulative effects of sustained activity over time and delayed effects in the changing dynamics of the conversation, we compared the cumulative Russian IRA activity for a period of 1–7 days prior to each focal day in the dataset with the mean degree of polarisation over the subsequent 1–20 days. Days were selected as the unit of measurement for this analysis for two primary reasons. Firstly, they reflect a natural cycle for the conversations as the activity peaks during the day and falls at night. Secondly, they allow for a substantial amount of activity to have occurred within each window of observation, which increases the robustness of the approach, compared to using hours for example, without the granularity loss that would occur if measuring activity at a weekly scale.

To test if the association between Russian IRA activity and subsequent polarisation was significantly higher than expected by chance, we used a permutation test. For each given level of lag in polarisation (1–20 days) and cumulative period of Russian IRA activity (1–7 days), we simulated a new dataset where Russian IRA activity for each day was paired with a level of polarisation randomly sampled (with replacement) from our real dataset. We then calculated the correlation coefficient between the Russian activity and the lagged polarisation. This was repeated for 10,000 iterations. This circumvented the problem of autocorrelation associated with the lagged time-series as the lagged polarisation was calculated after the randomisation. To avoid biased coefficients arising from right-skewed distributions of activity and polarisation, we normalised the data using box-cox transformations in the R package 'MASS' (Box & Cox, 1964; Venables, Bates, Firth, & Ripley, 2018).

We measured effect sizes for each cumulative period and lag period combination by taking the mean for the total 10,000 simulations. Significance values were calculated as the proportion of simulations where the simulated correlation was higher than the observed correlations (Bishara & Hittner, 2012).

Part 2 - Measuring the direct effect on Reddit conversations

Data collection and sampling

Reddit is a social media platform that focuses on news aggregation and discussion. Content is crowd-sourced, with members submitting text, images, or external hyperlinks, which are then voted up or down by other members. This content is organised into specific 'subreddits', user-created boards covering a wide variety of topics. In February 2018, Reddit had 542 million monthly visitors, ranking as the #3 most visited website in US and #6 globally (Alexa, 2019).

In the summer of 2018 Reddit released the identity of Russian IRA accounts. This totalled 16,821 Reddit posts from 944 accounts (Huffman, 2018). We extracted our dataset in November 2018 from a publicly available repository of historical Reddit data on pushshift.io (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020). The data are available in the form of a Google Big Query Database, which can be queried by users to download specific sections of the entire database. Here we study the period from January–December 2016, the period during which the Russian accounts were most active. We selected subreddits on which Russian IRA accounts posted at least 50 submissions during 2016. These span a range of topics, allowing us to explore differential effects in different areas of the social media platform. Previous research (Gallacher & Fredheim, 2019) has demonstrated that some of these subreddits were used by the Russian accounts for political manipulation, while others were used for more mundane purposes such as generating realistic account histories or 'karma' (platform specific credits that give a user more credibility in their comments). We selected the followed 12 subreddits; r/funny, r/uncen, r/Bad_Cop_No_Donut, r/AskReddit, r/PoliticalHumor, r/news, r/worldnews, r/aww, r/gifs, r/politics, r/The_Donald, r/racism. Of these, the subreddit r/uncen had received only submissions from Russian IRA accounts and no comments on the posts and was therefore not included in the analysis. Pushshift collects data at the point that it is posted to Reddit. This means that the dataset is unaffected by subsequent deletion of posts, however it also means that it does not capture edits made to comments after they are posted (a feature available on Reddit but not on Twitter).

Text Analysis Measures

The impact of Russian IRA activity on the conversational quality on Reddit was operationalised using three text analysis measures, which were applied to each post included in the analysis: Integrative Complexity, Toxicity, and Identity Attacks. Together, changes in these measures of conversation

quality would reflect the conversations becoming more polarised, more antagonistic and adversarial and potentially more hateful. Measuring the nature of the conversations in this more focused way allows us to measure more granular changes to complement the broader network approach outlined in Part 1.

Integrative Complexity (IC) is a social-psychological measure of how much an individual presents the ability to think and reason with input from multiple perspectives (Streufert & Suedfeld, 1965). Higher IC is associated with more accurate and balanced perceptions of other people, lower prejudice, the use of more information when making decisions, as well as less extreme views (Smith, Suedfeld, Conway, & Winter, 2008; Tetlock, Peterson, & Berry, 1993). Lower IC in discussions is associated with prediction of future violence and intergroup conflict (Chapter 1; Guttieri, Wallace, & Suedfeld, 1995; Suedfeld & Bluck, 1988). We used AUTO IC (Gideon, Conway, & Houck, 2014; Houck, 2014) to get IC scores for each Reddit post. The system produces a score from 1 to 7 for each comment, with lower scores representing lower levels of complexity. AUTO IC has been used successfully for the study of online terrorist content, demonstrating the validity of applying the measure to the digital domain (Houck, Repke, & Conway, 2017).

We measured the level of Toxicity of each Reddit post with the Google Perspective API[1]. This classification tool was designed by Google's 'Project Jigsaw' and 'Counter Abuse Technology' teams with the aim of promoting better discussions online (Wulczyn, Thain, & Dixon, 2017). The tool uses machine learning models to score the perceived impact a comment might have on a conversation. Comments defined as being ruder, more disrespectful, or more unreasonable receive a higher Toxicity score. The model gives a Toxicity score for each comment on a scale ranging from 0 to 1.

The Google Perspective API also provides additional classifiers that are more specific and can provide further insight into the nature of comments. The Identity Attack option measures the degree to which a comment demonstrates negative or hateful comments targeting someone because of their identity. This is especially useful in the current study as it measures specific intergroup aggression and conflict based on who people are and the social groups they are perceived to belong to. This aligns closely with existing measures of hate speech discussed in Chapters 2-4 (Davidson, Warmsley, Macy, &

[1] Google Project Jigsaw, 'Perspective', accessed 23 March 2018.

Weber, 2017; de Gibert, Perez, García-Pablos, & Cuadros, 2018). As with Toxicity, the model provides an Identity Attack score for each post on a scale ranging from 0 to 1.

Analysis of submissions and comments

Russian IRA activity consisted of submissions and comments. A submission is the starting post for a new conversation—i.e. threads started by Russian IRA accounts—while a comment is a post made on an existing conversation thread started by a genuine user. We analysed submissions and comments separately. We tested whether threads started by Russian IRA posts differed from those started by genuine users, and if Russian IRA comments injected into an existing thread had an impact on the subsequent conversation.

To measure the impact of submissions from Russian IRA accounts, we collected all comments made on threads started by Russian IRA accounts, including the initial submission starting the conversation, from the eleven subreddits identified above. In total this included 2,368 submissions and 30,112 comments. To test whether these conversations differed from genuine conversations, we collected a similar number of random 'control' submissions to the same subreddits within the same time frame. As with the Russian IRA submissions, we collected all the responses to this sample of genuine submissions, with a resulting total of 1,872 submissions and 22,503 comments. The lower number of genuine submissions is due to the exclusion of some submissions which received no subsequent comments. We then compared the conversation qualities for these two types of threads (those started by Russian IRA posts vs genuine submissions). As each subreddit was likely to include both types of conversation, we compared like-for-like conversations in each subreddit independently. For each comment in a thread we calculated a number of metrics relating to the measures used to determine the quality of the conversation, namely Integrative Complexity, Toxicity and Identity Attack.

To measure the impact of Russian IRA comments on existing genuine threads (rather than on new threads), we collected the comments from all threads that received at least one comment from a Russian IRA account. The sample of unmanipulated comment threads above was also used as the control for this sample. This dataset contained 455,300 comments from 826 threads, 1,253 of which came from Russian IRA controlled accounts. For each thread we numbered all comments in chronological order, with the injected Russian IRA post numbered as index position zero, subsequent posts incremented positively and previous posts negatively. We limited our analysis to threads

containing ≥ 20 comments and to the 100 posts either side of the injected Russian IRA post. For each of these 200 comments we calculated the three text analysis measures and averaged these for each position in the thread across all threads, to show the average trend of the conversations. The data were then analysed using a causal analysis model (see details below) to detect changes in the three metrics after the injection of a Russian IRA comment. The analysis was performed across all subreddits for each metric. To explore whether the effect differed between political and non-political conversations, it was then run separately on political and non-political subreddits (Political_Subreddits; 'The_Donald', 'politics', 'Bad_Cop_No_Donut', 'PoliticalHumor', 'racism', 'news', 'worldnews', Other_Subreddits; 'aww', 'gifs', 'funny', 'AskReddit'). We investigated both the immediate and the overall impact of a content injection by running the analysis on the first 25 comments as well as on all 100 comments post-injection.

Statistical Methods

We investigated differences between conversations started by Russian IRA accounts compared to controls by using linear mixed models (LMMs) with the lme4 package. We investigated differences in Integrative Complexity, Toxicity, and Identity Attacks between the Russian IRA-started and genuine threads, including subreddit ID as a random effect. Significance levels of group differences were obtained by comparing the full model to the null model with an $\chi 2$ test. The difference between Russian IRA-started and genuine threads was also compared in each of the 11 individual subreddits using Welch two sample t-tests comparing the differences in mean conversation qualities. Toxicity and Identity Attack measures were square-root-transformed to ensure normality. Integrative Complexity could not be normalised, and so a Wilcoxon rank sum test with continuity correction was used. We corrected for multiple comparisons by adjusting the p-values with a Bonferroni-Holm correction.

We calculated the impact of a single artificial comment on an existing thread by constructing Bayesian structural time-series models of both manipulated and unmanipulated 'control' threads, and used impact modelling to estimate the causal effect of the manipulation relative to the control on conversation quality (Brodersen, Gallusser, Koehler, Remy, & Scott, 2015). Specifically, the time-series information used in this analysis reflects the conversation quality (individually across the three text analysis measures) as it progresses over time along the thread before and after the injection of the Russian IRA comment. We normalised all threads such that this comment injection occurred at thread index position zero, and inspect (up to) the 100 preceding and 100 following messages in the thread. This analysis was conducted using the CausalImpact package (Brodersen & Hauser, 2017).

This method allowed us to make causal inferences even though performing a randomised experiment was not possible. Through the construction of a time-series model, this method predicts a counterfactual of how the response metric would have evolved after the intervention if the intervention had never occurred (Brodersen, Gallusser, Koehler, Remy, & Scott, 2015) – in this case how the qualities of conversation would have evolved in the absence of any manipulation from Russian IRA accounts. Implementing this causal impact modelling requires a control time-series of similar data unaffected by the intervention—here we used the unmanipulated threads described above, which are sampled from the same subreddits in the same timeframe, and measured on the same three elements of conversation quality. In matching control and manipulated threads in this way, we can therefore expect similar conversational evolution across both conditions, and any 'natural' topic-specific patterns of increase/decline of conversation quality are controlled for.

By calculating the relationship between the control and manipulated time-series on the 100 posts prior to the intervention we are able to project forwards the predicted time-series over the subsequent 100 posts, had there been no injection of Russian IRA comment. We then calculated the observed pointwise differences between manipulated and predicted threads after the intervention occurred. Summing these pointwise differences over a given time window, the model provides a measure of the size of this cumulative difference over time, which was tested for statistical significance with a Bayesian one-sided tail area probability test.

Any differences in overall thread quality between control and manipulated threads are controlled for in this initial 100 post period, and it is the change from baseline relationship between control and manipulated threads which is instead measured. For example this would account for any effects whereby Russian IRA accounts may be posting to threads which are already generally more toxic.

Ethics

All research was conducted in accordance with the University of Oxford Ethics Committee (Ethics Reference: R57579/RE001). All data collection was conducted using open-source methods and publicly available data, and hence, informed consent was not explicitly obtained.

Results

Polarisation of Twitter conversations

Correlations between Russian IRA activity and subsequent polarisation of the Twitter conversation related to Black Lives Matter were significantly higher than expected by chance (permutation test, Figure 3b). This effect did not occur immediately following Russian IRA activity, but rather occurred predominantly between 3 and 10 days after the conversation manipulation had taken place. More specifically, it increased over time until reaching a peak around 7–9 days following the activity, and then gradually returned to the initial base level (Figure 3a). The effect started earlier, lasted longer, and was more pronounced when we considered Russian IRA activity over a longer time window (Figure 3, Table 3, see Table S1 for individual significance scores and correlation effect sizes). When looking at the longest period of cumulative activity—seven days—this trend appeared to last for almost two weeks from day two until day 14. Importantly, there was no general increasing or decreasing trend over time for either Russian IRA activity or polarisation and so our results were not due to long-term matching trends between the two variables.

The distributions of daily Russian IRA activity showed a long right tail (SI Figure 2c), suggesting this activity was uncommonly large on certain days. We tested whether these spikes in Russian IRA activity had an especially large effect on subsequent conversation polarisation by taking the top 10 days with the highest degree of polarisation, and testing whether each of these days had been preceded by a spike in Russian IRA activity (defined as a day with over 100 tweets) within a period of 10 days. We found that in 80% of these most polarised days, a spike in activity had preceded the polarisation.

The highest peaks in Russian IRA activity were fairly evenly distributed throughout the period studied. The mean Russian IRA activity across all days was 27 tweets, but this spiked as high as 592 tweets in a single day and 16 days had over 100 posts.

Table 3 – Statistical results for the highest correlation in the lagged permutated test across each activity window. For full results see SI Table 1.

Activity Window (Days)	Number Significant Days	Start Day (Lag)	End Day (Lag)	Day of Max Correlation	Max Correlation	p
1	*0*	*NA*	*NA*	*1*	*0.011*	*0.419*
2	*0*	*NA*	*NA*	*1*	*0.009*	*0.428*
3	*0*	*NA*	*NA*	*7*	*0.041*	*0.217*
4	*0*	*NA*	*NA*	*7*	*0.078*	*0.069*
5	*4*	*5*	*8*	*7*	*0.107*	*0.019*
6	*8*	*3*	*10*	*6*	*0.136*	*0.005*
7	*11*	*2*	*11*	*6*	*0.156*	*< 0.001*

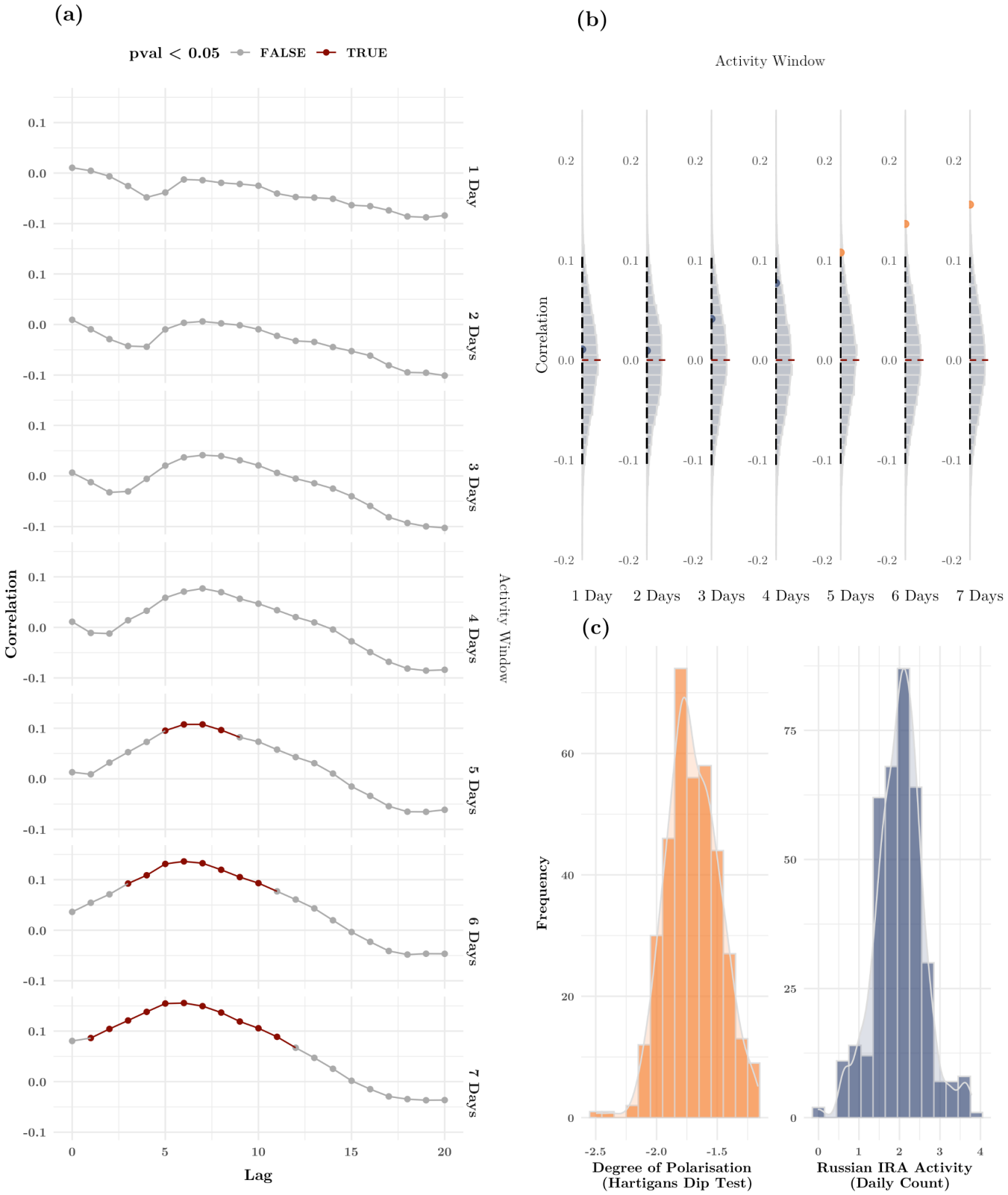


Figure 3(a-c) – a) Correlations between the degree of daily polarisation in BLM conversations on Twitter and preceding total Russian IRA Activity over various periods (1-7 days). Red dots show significant correlations.
b) Significance effects for max correlations for each activity window compared to distribution obtained by chance (grey) as calculated with a permutation test. (orange : $p < 0.05$, blue = non significant)
c) Normalised distributions of polarisation and activity (see appendix for raw distributions)

Reddit Submissions

The conversation quality on threads started by Russian IRA-operated accounts differed substantially from that of genuine conversations, but the direction of these differences varied between subreddits and thus between topics of conversation. Overall, posts on threads started by Russian IRA account had higher Toxicity (Russian IRA: 0.48 ± 0.001 vs genuine: 0.47± 0.002, n = 56,249, $\chi 12 = 28.34$, $p < 0.001$) and Identity Attacks (Russian IRA: 0.42 ± 0.001 vs genuine 0.40± 0.001, $\chi 12 = 85.33$, $p < 0.001$) but showed no overall change in Integrative Complexity (Russian IRA: 1.37± 0.004 vs genuine 1.36 ± 0.004, $\chi 12 = 2.39$, $p = 0.122$).

Further analyses performed on individual subreddits showed that threads started by Russian accounts within r/news, r/gifs, r/funny and r/Bad_Cop_No_Donut had higher average Toxicity scores than genuine threads in the same subreddits (Figure 4b, Table 4). Other subreddits showed no differences. We found a similar pattern with regard to levels of Identity Attack. Threads started by Russian accounts within r/racism, r/news, r/gifs, r/funny and r/AskReddit had higher average Identity Attack scores than genuine threads in the same subreddits (Figure 4c, Table 4), while artificial

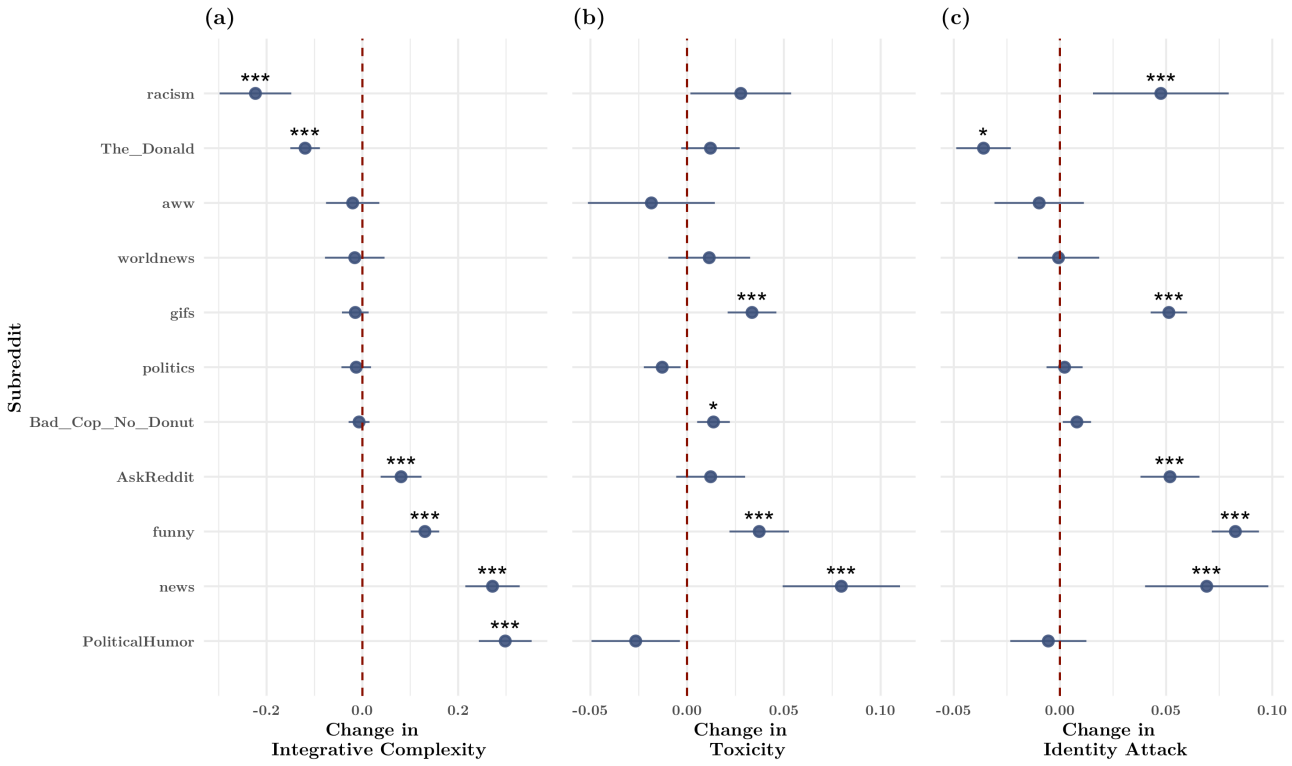


Figure 4(a-c) – Differences in mean conversation quality scores for threads started by Russian IRA Reddit accounts compared to genuine comment threads within the same subreddit. Higher values indicate Russian IRA started conversations scored higher on that conversation metric.

Table 4 – Statistical results for pared sample t-tests comparing differences in mean conversation quality scores for threads started by Russian IRA Reddit accounts compared to organic comment threads within the same subreddit

Text Analysis Measures

Subreddits	Integrative Complexity					Toxicity					
	Mean IRA Started	Mean Genuine	W	p	d	Mean IRA Started	Mean Genuine	df	t	p	d
r/racism	1.18±0.02	1.40±0.03	132938	< 0.001	0.318	0.55±0.01	0.52±0.01	987	-2.09	0.037	0.127
r/The_Donald	1.15±0.01	1.27±0.01	1626926	< 0.001	0.149	0.48±0.01	0.47±0.01	3646	1.56	0.471	0.051
r/aww	1.09±0.02	1.11±0.02	32751	1	0.064	0.36±0.01	0.38±0.01	292	-1.11	0.541	0.103
r/worldnews	1.41±0.03	1.43±0.02	426810	1	0.024	0.46±0.01	0.45±0.01	1023	1.06	0.541	0.033
r/gifs	1.22±0.01	1.23±.01	3296930	1	0.029	0.45±0.004	0.42±0.01	4900	-5.22	< 0.001	0.145
r/politics	1.44±0.01	1.45±0.01	9524468	1	0.018	0.45±0.004	0.46±0.002	4083	2.66	0.06	0.06
r/Bad_Cop_No_Donut	1.38±0.01	1.39±0.01	18150512	1	0.011	0.53±0.003	0.51±0.003	11,717	-3.16	0.013	0.058
r/AskReddit	1.34±0.02	1.26±0.01	858343	< 0.001	0.149	0.43±0.01	0.42±0.01	2413	1.34	0.541	0.054
r/funny	1.29±0.01	1.16±0.01	1690132	< 0.001	0.259	0.46±0.004	0.42±0.001	2143	4.75	< 0.001	0.164
r/news	1.42±0.001	1.15±0.03	1296950	< 0.001	0.406	0.49±0.002	0.41±0.02	186	-5.18	< 0.001	0.334
r/PoliticalHumor	1.53±0.02	1.23±0.02	664965	< 0.001	0.405	0.44±0.002	0.41±0.01	726	4.75	0.136	0.164

Subreddits	Identity Attack					
	Mean IRA Started	Mean Genuine	df	t	p	d
r/racism	0.61±0.01	0.56±0.01	984	-2.92	0.022	0.177
r/The_Donald	0.39±0.01	0.43±0.01	3792	5.5	< 0.001	0.178
r/aww	0.29±0.01	0.30±0.01	294	-0.92	1	0.085
r/worldnews	0.42±0.01	0.42±0.01	991	-0.07	1	0.004
r/gifs	0.36±0.003	0.31±0.003	5208	-11.7	< 0.001	0.314
r/politics	0.42±0.003	0.42±0.002	3766	0.5	1	0.012
r/Bad_Cop_No_Donut	0.43±0.003	0.42±0.003	11,644	-2.35	0.095	0.043
r/AskReddit	0.37±0.01	0.32±0.004	2112	-7.31	< 0.001	0.3
r/funny	0.40±0.004	0.32±0.004	3155	-14.6	< 0.001	0.429
r/news	0.44±0.002	0.34±0.02	185	-4.7	< 0.001	0.314
r/PoliticalHumor	0.39±0.003	0.40±0.01	735	-0.6	1	0.031

comment threads started within r/TheDonald by comparison had a lower average Identity Attack scores than genuine threads. Other subreddits showed no differences. While we found no difference in Integrative Complexity overall, artificial threads started by Russian IRA accounts received comments with lower IC scores in r/TheDonald and r/racism, but higher IC scores in r/PoliticalHumor, r/news, r/funny and r/AskReddit (Figure 4a, Table 4).

Reddit Comments

Across all subreddits and comment threads, Russian IRA comments led to a small drop in the Integrative Complexity of the subsequent conversation over a period of 100 comments by a factor of

1% ± 0.51 (Figure 5a-c). For the period after a Russian IRA comment injection, the average Integrative Complexity was 1.41 ± 0.004. In the absence of any intervention, the causal analysis model predicted an average value of 1.42 ± 0.006, significantly higher than the observed value (Bayesian one-sided tail area probability p = 0.035). In other words, on average a Russian IRA comment caused a 0.01 decrease in IC compared to predictions.

Additionally, Russian IRA comment injections lead to short term increase in the Integrative Complexity of conversations in non-political subreddits by a factor of 2% ± 0.77 over the subsequent 25 comments (p = 0.005). There were no measurable differences in the effect of Russian IRA comment injection on Integrative Complexity in political subreddits when considered in isolation, or in non-political subreddits over longer periods of time (SI Table 2).

Russian IRA comment injections also affected the Toxicity of subsequent conversations, but these effects occurred only in political subreddits and for short periods. While there was no significant effect

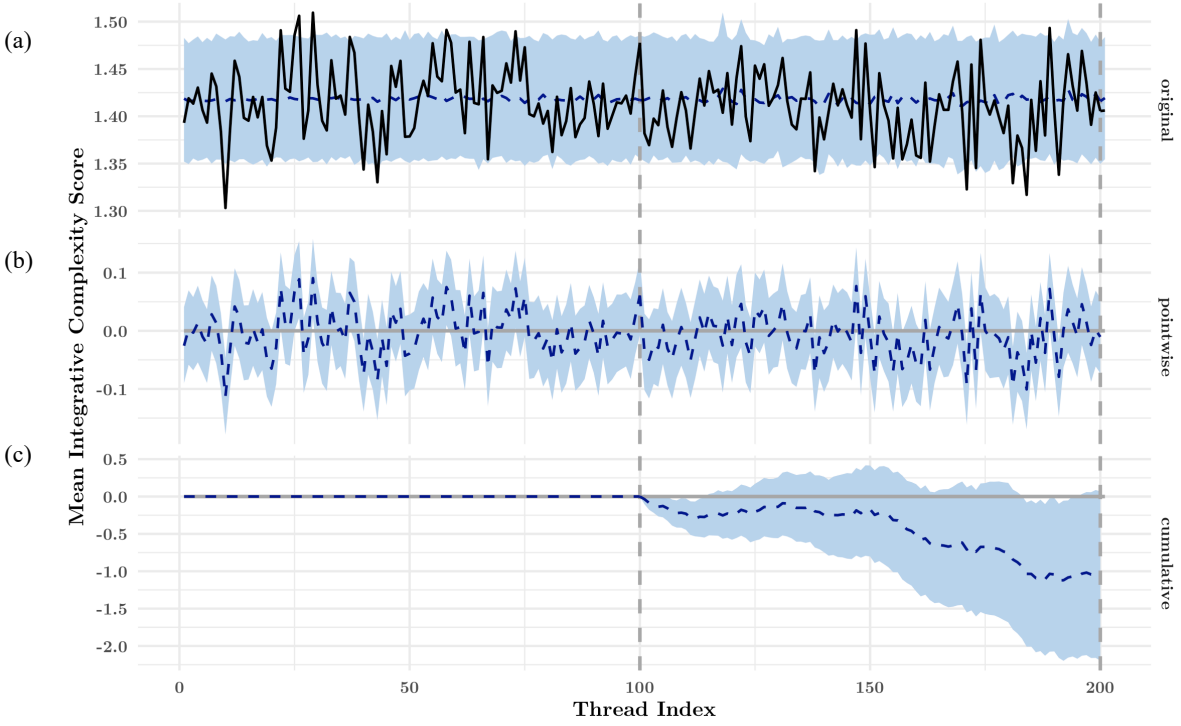


Figure 5(a-c) – Causal Impact analysis of Artificial Russian Reddit account comment injection on the Integrative Complexity of the conversation. Panel (a) - the observed trend for average IC over the course of the conversations, along with the counterfactual prediction period after the intervention if it had not occurred. Panel (b) - the pointwise difference this counterfactual prediction and the observed data. Panel (c) - the cumulative pointwise difference overtime, giving an indication of the overall effect of the intervention on the IC of the conversation.

of Russian IRA comment injection on Toxicity if considered over the entire post-intervention period of 100 comments, comment injections did increase Toxicity of the conversation over the next 25 comments by a factor of 3% ± 1.53 ($p = 0.019$), but this effect subsequently disappeared over the following 75 comments (Figure 6).

Similarly, the impact of the degree of Identity Attacks taking place in conversations after a Russian IRA comment injection also depended on whether the comments occurred in political or non-political subreddits. In non-political subreddits, comment injection was followed by a marked short-term increase in Identity Attacks over the next 25 comments by a factor of 10% ± 2.04 ($p = 0.001$), and this effect subsequently dissipated over time. There was no change in the degree of Identity Attacks following a Russian IRA comment injection in a political subreddit.

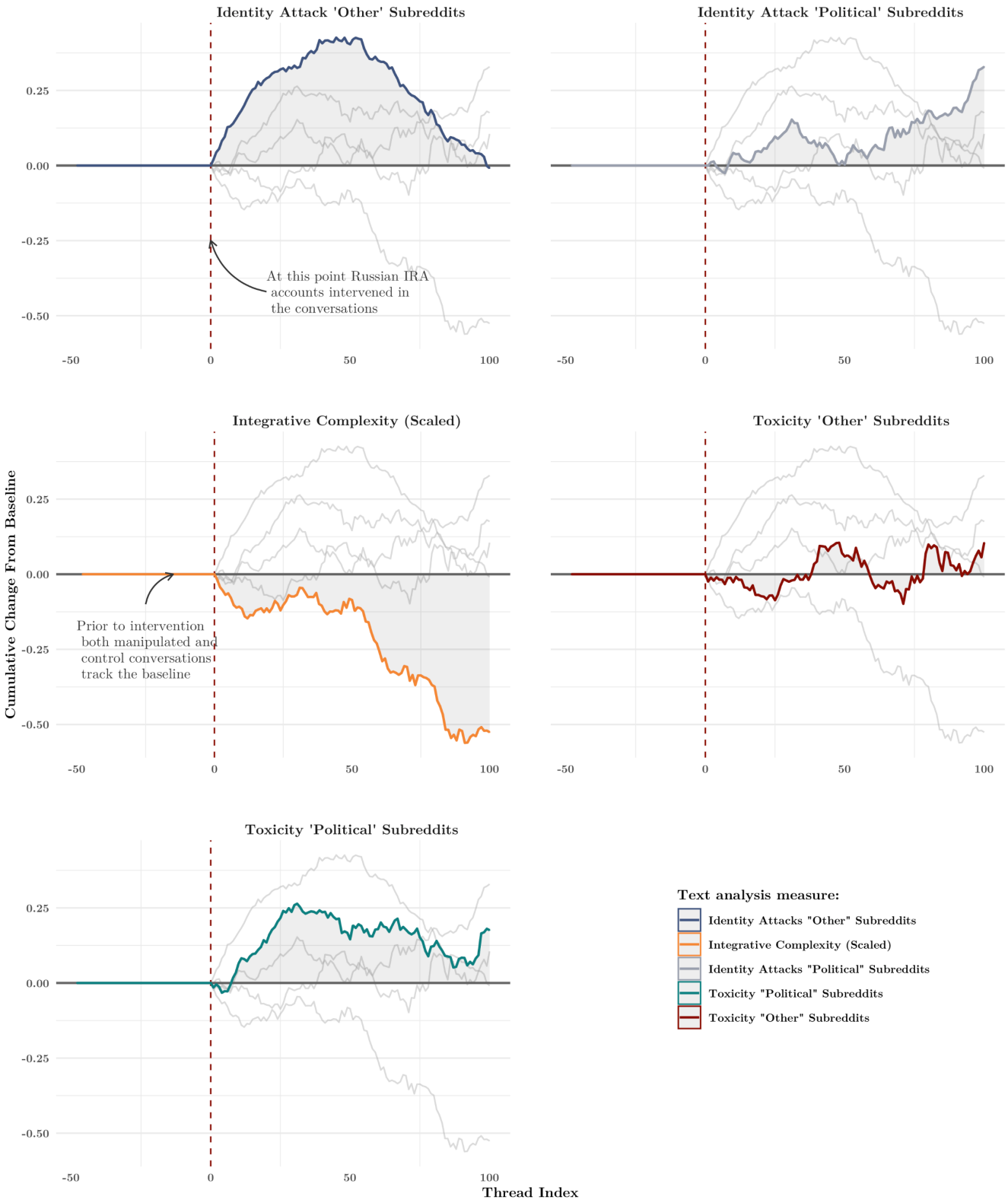


Figure 6 – Cumulative impact of artificial Russian IRA comment injections on the toxicity, identity attack (IA) , and integrative complexity (IC) of subsequent conversation on Reddit in political and non-political subreddits.

Discussion

In this study we examined whether social media activity from artificial accounts run by the Russian IRA led to measurable changes in the conversation of genuine users on Twitter and Reddit. Our results show that Russian IRA activity indeed predicted changes in the conversations taking place on both platforms, but the exact effects differed between platforms and the type of manipulation taking place.

On Twitter, higher amounts of Russian IRA activity in the Black Lives Matter conversation predicted increases in the subsequent conversational polarisation of genuine Twitter users. This increase in polarisation peaked approximately one week after the injection of Russian IRA content and the association was most pronounced around the periods of highest Russian activity, suggesting that large spikes in Russian IRA activity had the greatest influence on the subsequent conversation. The gradual build-up of these effects over a week may reflect a structural property of Twitter—that more a tweet is retweeted, the more influence it gains on the network (Lim, Achananuparp, & Zhu, 2011). On days with higher numbers of tweets from Russian IRA accounts there was a greater likelihood that one of the tweets would go 'viral' and be exposed to a much larger audience—either by simply manually increasing the number of tweets or by mass (automated) retweeting through the use of connected bot accounts (Fredheim, 2019; Kumar et al., 2017). This orchestration of inauthentic accounts may also play a role in the gradual build-up of the effects of these information operations we observed in the week following Russian IRA activity, with the nature or provocativeness of the campaign potentially building over time as it gains an audience. The exact mechanism underlying this delayed effect remains unknow however, and future research should investigate what (if any) the immediate effects of information operations on Twitter are, and also the role of sustained activity over the long-term.

Earlier research has found that Russian IRA accounts embed themselves into both for and against sides of the Black Lives Matter debate (Arif et al., 2018); our results show that this may have contributed to the polarisation of both sides of the debate. It is noteworthy that we find this effect despite the high attrition rate within our Twitter data; 45% of Tweets were deleted before data collection. Deleted tweets are more likely to contain negative sentiment or profanity (Bhattacharya & Ganguly, 2016) or to be 'regretted' by their author (Zhou, Wang, & Chen, 2016), and so the exclusion of these tweets likely muted the observed effects of Russian IRA activity on polarisation.

On Reddit we found that threads started by Russian IRA accounts were generally more Toxic than conversations started by genuine users whilst also showing more instances of Identity Attacks. Higher Toxicity reflects that these conversations were more rude, aggressive, or disrespectful, and more likely to inflame other users (both targets and observers), while conversations with higher Identity Attack scores contained a greater number of hostile comments made against people due to group membership, including race and political affiliation (Wulczyn et al., 2017). Both of these measures indicate that Russian IRA activity was effective in promoting hostile conversations among other users, likely increasing divisions among group lines.

The effects of Russian IRA activity on Integrative Complexity were more complicated. While there was no overall difference in the Integrative Complexity of threads started by the Russian IRA compared to genuine threads, there were differential effects of Integrative Complexity depending on the subreddit in which a conversation was started. Conversations started by Russian IRA accounts in r/racism and r/The_Donald showed reductions in Integrative Complexity, (less complex conversations with less nuance, demonstrating reasoning from fewer viewpoints (Streufert & Suedfeld, 1965), while conversations started in r/AskReddit, r/funny, r/news and r/PoliticalHumor displayed higher Integrative Complexity compared to genuine conversation threads in these subreddits. One interpretation of these results is that they are related to the partisan nature of the political subreddits, which may facilitate a reduction in complexity due to a lack of opposing voices (Sunstein, 2017), compared to the 'general interest' subreddits, which may enable greater intergroup discussion because of their non-partisan nature. These and other explanations need direct testing, however, and merit further research.

We also found evidence suggesting a causal relationship between Russian IRA activity and conversation quality by studying the impact of comments from Russian IRA accounts injected into existing genuine conversations. Across all subreddits, Russian IRA comment injections led to a decrease in the Integrative Complexity of the conversation over the subsequent 100 comments. Additionally, there was a shorter-lived effect, detectable on the 25 subsequent comments, which led to an increase in Toxicity in political subreddits and an increase in the level of Identity Attacks in non-political subreddits. Although these findings are less clear-cut than those described above, they similarly demonstrate that any measurable effects of Russian IRA activity are in the direction of undermining conversational quality. Cumulatively, these small effects have the power significantly shape a conversation. They also suggest that different dynamics unfold in political and non-political

online conversations, which is in line with previous findings (Barberá et al., 2015; Garimella, Morales, Gionis, & Mathioudakis, 2018), and that distinguishing between these conversational domains remains important in future research. We found that in the absence of manipulation the control threads within political subreddits had higher Integrative Complexity, Toxicity and Identity Attacks than non-political subreddits, suggesting that political conversations are characterised by both increased engagement and increased hostility. These characteristics may relate to findings that echo chambers form primarily in the political domain (Barberá et al., 2015; Garimella et al., 2018), but whether these are causes or effects remains to be tested.

Comparing the results across platforms, we found that the effects of Russian IRA activity manifested more quickly on Reddit than on Twitter. On average, the effects detected over 25 and 100 Reddit posts following manipulation peaked around 3.5 days and 5 days after submission respectively, while on Twitter the association between Russian IRA activity and polarisation peaked after 7 days. This is likely due to the structural differences between the platforms. On Twitter the impact of content is measured by popularity—how many people react to it—and therefore tweets that go viral can have a large effect on the overall conversation (Lim et al., 2011). On Reddit a single comment cannot go viral, and impact results from the cumulative effect of many posts, or of many users 'upvoting' a thread (Salihefendic, 2015). On Twitter, tweets can take longer to go viral, compared to the direct responses which occur on Reddit threads, that have a shorter-lived visibility. Given these considerations, it would also be interesting to study the consequences of more sustained periods of Russian activity in a single Reddit thread. Our analytical procedure did not allow us to identify these consequences as we could only model a single intervention at a time, but we expect that repeated co-ordinated activity within a single thread would lead to increased cumulative effects (Berger & Morgan, 2015; Ferrara, 2015). Including this co-ordinated behaviour may mean that the consequences of comments in existing threads more closely resemble the observed differences in total conversations following genuine submissions and Russian IRA submissions.

By increasing the polarisation of conversations on Twitter and undermining the quality of conversations on Reddit, Russian IRA activity is likely to be effective in increasing the distance between social groups, fuelling both ideological and affective polarisation (Mason, 2015). This in turn provides ideal circumstances for the distribution of disinformation (Del Vicario, Gaito, Quattrociocchi, Zignani, & Zollo, 2017; Garrett et al., 2016) because it increases the acceptance of (inaccurate) information that confirms prior views—a phenomenon known as 'confirmation bias' (Nickerson, 1998)

—and facilitates repeated exposure to the same inaccurate information because alternative perspectives are eliminated from discussion by design (Berinsky, 2017; Pennycook et al., 2018). Western societies that focus more on internal strife from polarised domestic communities tend to focus less on international issues, illustrating that this activity may be part of a larger geopolitical strategy (Kirchick, 2017; Singer & Brooking, 2018).

In this study we focused on activity originating from publicly attributed Russian IRA accounts and their effect on two key social media platforms. Future research should consider including other platforms, and also other groups engaged in information operations. Russian IRA activity accounts for a fraction of all possible information operations activities worldwide, and many other groups produce similar content for a range of different purposes. This includes pursuing international strategic goals (as demonstrated by Iranian actions (Karan et al., 2019)), focusing attention on perceived domestic concerns (utilised by far-right groups (Gleicher, 2019)), and quashing dissent (as tactic favoured by China (King, Pan, & Roberts, 2017)). Our study only begins to unveil the negative effect of information operations globally. If fuelling arguments on both sides of controversial topics works to increase polarisation in these conversations, then pushing only a single side may work to decrease polarisation or even to stifle active debate. This might be the goal for a regime that wishes to quash dissent or opposition. For example, evidence of Chinese government involvement in online discussions shows that across ~450 million social media posts per year the strategy is not to engage with controversial topics or with sceptics of government, but rather to change the subject of conversations with vocal cheerleading for pro-China positions to overwhelm opposition voices (King et al., 2017). The Kremlin takes a similar approach towards a domestic audiences, using troll farms such as the Russian IRA to produce vast quantities of pro-regime messages in Russian for local consumption (Gallacher & Fredheim, 2019).

While it remains to be seen whether these online effects translate into offline actions, there is evidence that online activities can have substantial effects on real world behaviour ranging from exercise and smoking to consumer trends (Althoff, Jindal, & Leskovec, 2017; Depue, Southwell, Betzner, & Walsh, 2015; Muralidharan & Men, 2015). Our research also shows that online interaction between groups predicts offline violence (Gallacher, Heerdink, & Hewstone, 2020), while other research demonstrates how online aggression towards disadvantaged groups can precede offline hate crimes (Chapter 4; Müller & Schwarz, 2018, Williams et al, 2019). By demonstrating that information operations promote social polarisation and can have measurable impacts on online conversations more broadly,

our study also highlights the risk of potential future vulnerabilities. The ability of hostile actors to create polarising content is increasing at pace, thanks to advances in machine-generated text that closely resembles human speech (Radford et al., 2018). If this technology is paired with malicious intent to drive communities apart using social media platforms, then the volume of content may well expand and increase the severity of the challenge to detecting inauthentic content and oppose it (Brundage et al., 2018).

It is therefore essential to design solutions that address and counter the negative effects of hostile information operations. Identifying the impact of information operations is only the first step in creating counter measures. Evidence suggests that organised attempts to challenge the veracity of disinformation on Twitter are generally ineffective (Margolin, Hannak, & Weber, 2018; Shin, Jian, Driscoll, & Bar, 2017), while spontaneous fact-checking on Facebook is rare and generally unsuccessful (Friggeri, Adamic, Eckles, & Cheng, 2014). Other technical solutions should therefore focus on the early detection of artificial content before it can manipulate online conversations (Wright & Anise, 2018), or on methods to inoculate citizens from the effects of disinformation (Roozenbeek & Linden, 2018). Structural changes to social media platforms promoting positive exposure to members of opposing groups will also likely reduce and dilute the impact of efforts to divide these same groups through negative content injections (Brown & Hewstone, 2005; Pettigrew & Tropp, 2008). Addressing the challenge of disinformation is so broad that designing effective interventions will require interdisciplinary efforts at multiple levels of analysis (Lazer et al., 2018).

Conclusion

Our study reveals that the malicious use of social media by ‘fake’ accounts can measurably affect the subsequent conversations held by genuine users. Using the activity of the Russian Internet Research Agency on Twitter and Reddit as case studies, we have shown that this effect differed between social media platforms. The effect of Russian activity on Twitter was to increase polarisation after a one-week delay, while there was a more immediate effect on Reddit, immediately altering the quality of subsequent conversations. By developing methods to measure the impact of information operations in online conversations, our study provides an important step in developing effective countermeasures.

Acknowledgements

This research was supported by grants from EPSRC and the University College Oxford Radcliffe Scholarship. The second author's contribution to this project was partially supported by a grant from the Netherlands Organisation for Scientific Research (NWO 446-16-015). The funding bodies played no further role in designing or implementing the research, and the authors declare no competing interests. We thank members of the Oxford Centre for the Study of Intergroup Conflict for their helpful feedback.

References

Alexa. (2019). reddit.com | Competitive analysis, marketing mix and traffic. Retrieved December 2, 2020, from https://www.alexa.com/siteinfo/reddit.com

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets are forever: A large-scale quantitative analysis of deleted tweets. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 897–907. https://doi.org/10.1145/2441776.2441878

Althoff, T., Jindal, P., & Leskovec, J. (2017). Online actions with offline Impact. *WSDM*, 537–546. https://doi.org/10.1145/3018661.3018672

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *2*, 1:26. https://doi.org/10.1145/3274289

Arnaudo, D. (2017). Computational propaganda in Brazil: Social bots during elections. *Computational Propaganda Research Project*, *8*, 1–39.

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., … Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.17605/OSF.IO/4YGUX

Bakillah, M., Li, R. Y., & Liang, S. H. L. (2015). Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: The case study of typhoon Haiyan. *International Journal of Geographical Information Science*, *29*(2), 258–279. https://doi.org/10.1080/13658816.2014.964247

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Barberá, P. (2018). Explaining the spread of misinformation on social media: Evidence from the 2016 U.S. presidential election. *Comparative Politics Newsletter*, *28*(2), 7:11.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *ArXiv*. Retrieved from https://arxiv.org/pdf/2001.08435.pdf

Bay, A. S., Bertolin, G., Biteniece, N., Christie, E. H., Rolf, E., Gallacher, J. D., … Marchenko, T. (2019). *Responding to cognitive security challenges*. NATO Strategic Communications Centre of Excellence. Retrieved from https://stratcomcoe.org/responding-cognitive-security-challenges

Berger, J. M., & Morgan, J. (2015). *The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter. The Brookings Project on U.S. Relations with the Islamic World*. Retrieved from https://www.brookings.edu/research/the-isis-twitter-census-defining-and-describing-the-population-of-isis-supporters-on-twitter/

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*(2), 241–262. https://doi.org/10.1017/S0007123415000186

Bertrand, N. (2017). Russia organized 2 sides of a Texas protest and encouraged "both sides to battle in the

streets." *Business Insider*. Retrieved from https://www.businessinsider.com/russia-trolls-senate-intelligence-committee-hearing-2017-11

Bhattacharya, P., & Ganguly, N. (2016). Characterizing deleted Tweets and their authors. *ICWSM*, 10–13. Retrieved from http://parantapa.net/mypapers/bhattacharya-icwsm16.pdf

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods*, *17*(3), 399–417. https://doi.org/10.1037/a0028087

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Bright, J. (2018). Explaining the emergence of echo chambers on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23*, 17–33. https://doi.org/10.2139/ssrn.2839728

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applies Statistics*, *9*(1), 247–274. https://doi.org/10.1214/14-AOAS788

Brodersen, K. H., & Hauser, A. (2017). Package "CausalImpact": Inferring causal effects using bayesian structural time-series models. *CRAN*, 1–8. https://doi.org/10.1214/14-AOAS788>.See

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., … Amodei, D. (2018). *The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation.* https://doi.org/10.1002/adma.201405087

Castree, N., Kitchen, R., & Rogers, A. (2013). *A dictionary of human geography*. Oxford University Press.

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *133*(26), 89–96. https://doi.org/10.1021/ja202932e

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, 1–3. https://doi.org/10.1038/s41562-017-0213-3

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*, 512–515. Retrieved from http://arxiv.org/abs/1703.04009

de Gibert, O., Perez, N., García-Pablos, A., & Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. *Proceedings Ofthe Second Workshop on Abusive Language Online (ALW2)*, 11–20. https://doi.org/10.18653/v1/w18-5102

Del Vicario, M., Gaito, S., Quattrociocchi, W., Zignani, M., & Zollo, F. (2017). Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the Italian Referendum. *ArXiv*. Retrieved from http://arxiv.org/abs/1702.06016

Depue, J. B., Southwell, B. G., Betzner, A. E., & Walsh, B. M. (2015). Encoded exposure to tobacco use in social media predicts subsequent smoking behavior. *American Journal of Health Promotion*, *29*(4), 259–261. https://doi.org/10.4278/ajhp.130214-ARB-69

DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., … Johnson, B. (2018). *The tactics & tropes of the Internet Research Agency*. Retrieved from https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper-121718.pdf

docnow.io. (2018). Documenting the Now. Retrieved December 2, 2020, from https://www.docnow.io/

Documenting the Now. (2019). GitHub - Hydrator: Turn Tweet IDs into Twitter JSON & CSV. Retrieved December 2, 2020, from https://github.com/DocNow/hydrator

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review*, *107*(01), 57–79. https://doi.org/10.1017/s0003055412000500

Ferrara, E. (2015). Manipulation and abuse on social media. *SIGWEB Newsletter*. https://doi.org/10.1145/2749279.2749283

Ferrara, E. (2017). Desinformation an bots oparations on the run up to the 2017 French presidential election. *SSRN Electronic Journal*. https://doi.org/10.5210/FM.V22I8.8005

Fredheim, R. (2019). *Robotrolling 2019. Issue 1*. Riga, Latvia. Retrieved from https://www.stratcomcoe.org/robotrolling-20191

Freelon, D., McIlwain, C. D., & Clark, M. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2747066

Friggeri, A., Adamic, L., Eckles, D., & Cheng, J. (2014). Rumor cascades. *ICWSM 2014 International Conference on Weblogs and Social Media*, 101–110. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8122/8110

Gallacher, J. D., Heerdink, M. W., & Hewstone, M. (2020). Online contact between opposing political protest groups via social media is linked to physical violence of offline encounters. *Social Media + Society*, 1–44.

Gallacher, J. D, & Fredheim, R. E. (2019). Division abroad, cohesion at home: How the Russian troll factory works to divide societies overseas but spread pro-regime messages at home. In *Responding to Cognitive Security Challenges* (p. 60:79). Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Gallagher, R. J., Reagan, A. J., Danforth, C. M., & Dodds, P. S. (2018). Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLoS ONE*, *13*(4), 1–23. https://doi.org/10.1371/journal.pone.0195644

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *International World Wide Web Conference*, *2*. Retrieved from http://arxiv.org/abs/1801.01665

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. *Journal of Computer-Mediated Communication*, *21*(5), 331–348. https://doi.org/10.1111/jcc4.12164

Gerber, T. P., & Zavisca, J. (2016). Does Russian propaganda work? *The Washington Quarterly*, *39*(2), 79–98. https://doi.org/10.1080/0163660X.2016.1204398

Gideon, L., Iii, C., Conway, K. R., & Houck, S. C. (2014). Automated Integrative Complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Gleicher, N. (2019). *Removing coordinated inauthentic behavior from the UK and Romania*. Retrieved from https://newsroom.fb.com/news/2019/03/removing-cib-uk-and-romania/

Google Project Jigsaw. (2018). Perspective. Retrieved March 23, 2018, from https://www.perspectiveapi.com/#/

Guess, A., Barber, P., Vaccari, C., Kingdom, U., Nyhan, B., Seigel, A., … Stukal, D. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature*. Retrieved from https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/

Guttieri, K., Wallace, M. D., & Suedfeld, P. (1995). The Integrative Complexity of American decision makers in the Cuban missile crisis. *Journal of Conflict Resolution*, *39*(4), 595–621. https://doi.org/10.1177/0022002795039004001

Gvirsman, S. D., Garrett, R. K., Dal, A., Neo, R., Tsfati, Y., & Johnson, B. K. (2014). Implications of pro- and counterattitudinal information exposure for affective polarization. *Human Communication Research*, *40*(3), 309–332. https://doi.org/10.1111/hcre.12028

Hartigan, J, A., & Hartigan, P, M. (1991). The dip test of unimodality. *Annals of Statistics*, *19*(3), 1403–1433.

Hindman, M., & Barash, V. (2018). *'Fake news' and influence campaigns on Twitter*. Retrieved from https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter/

Hirschfeld, H. O., & Wishart, J. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(04), 520. https://doi.org/10.1017/S0305004100013517

Houck, S. C. (2014). Automated Integrative Complexity: Current challenges and future directions. *Political Psychology*, *35*(5), 647–659. https://doi.org/10.1111/pops.12209

Houck, S. C., Repke, M. A., & Conway, L. G. (2017). Understanding what makes terrorist groups' propaganda effective: an integrative complexity analysis of ISIL and Al Qaeda. *Journal of Policing, Intelligence and Counter Terrorism*, *12*(2), 105–118. https://doi.org/10.1080/18335330.2017.1351032

Howard, P. N., Ganesh, B., Liotsiu, D., Kelly, J., & Camille François, G. (2018). *The IRA, social media and political polarization in the United States, 2012-2018*. Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf

Howell, L., Gmur, M. N., Bisanz, P., de Sola, I., Von, K., Prampart, B., … Lanois, A. (2014). *Outlook on the global agenda 2014*. Geneva, Swizerland. Retrieved from www.weforum.org

Huffman, S. (2018). Reddit's 2017 transparency report and suspect account findings. Retrieved February 25, 2019, from https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/

Husson, F., Josse, J., Le, S., & Maintainer, J. M. (2018). Package "FactoMineR": Multivariate Exploratory Data Analysis and Data Mining Author. *CRAN*. https://doi.org/10.1201/b10345-2>

Intelligence Community Assesment. (2017). *Assessing Russian activities and intentions in recent US elections*. Retrieved from https://www.dni.gov/files/documents/ICA_2017_01.pdf

Karan, K., Barojan, D., Hall, M., & Brookie, G. (2019). *#TrollTracker: Outward influence operation from Iran*. Retrieved from https://medium.com/dfrlab/trolltracker-outward-influence-operation-from-iran-cc4539684c8d

Kello, L. (2017). *The virtual weapon and international order*. Yale University Press. https://doi.org/10.2307/j.ctt1trkjd1

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, *111*(03), 484–501. https://doi.org/10.1017/s0003055417000144

Kirchick, J. (2017). *Russia's plot against the West. Politico*. Retrieved from https://www.politico.eu/article/russia-plot-against-the-west-vladimir-putin-donald-trump-europe/

Kosloff, S., Greenberg, J., Schmader, T., Dechesne, M., & Weise, D. (2010). Smearing the opposition: Implicit and explicit stigmatization of the 2008 U.S. presidential candidates and the current U.S. President. *Journal of Experimental Psychology: General*, *139*(3), 383–398. https://doi.org/10.1037/a0018809

Kumar, S., Cheng, J., Leskovec, J., & Subrahmanian, V. S. (2017). An army of me: Sockpuppets in online discussion communities. *International World Wide Web Conference Committee*. https://doi.org/10.1145/3038912.3052677

Lazer, D. M. J., Schudson, M., Benkler, Y., Zittrain, J. L., Thorson, E. A., Watts, D. J., … Rothschild, D. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, E. J. (2007). Deindividuation effects on group polarization in computer-mediated communication: The role of group identification, public-self-awareness, and perceived argument quality. *Journal of Communication*, *57*(2), 385–403. https://doi.org/10.1111/j.1460-2466.2007.00348.x

Lim, E.-P., Achananuparp, P., & Zhu, F. (2011). On modeling virality of Twitter content. *ICADL*, 212:221. https://doi.org/10.1017/CBO9781107415324.004

Linvill, D. L., & Warren, P. L. (2018). Troll factories: The Internet Research Agency and state-sponsored agenda building. *(In Press)*, 21. Retrieved from https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building

Lucas, E., & Pomerantsev, P. (2016). *Winning the information war: Techniques and counter-strategies to Russian propaganda in central and Eastern Europe* (Vol. 14). https://doi.org/10.1108/13590790710828136

Maechler, M. (2015). *Package "diptest": Hartigan's dip test statistic for unimodality. CRAN*. Retrieved from https://cran.r-project.org/web/packages/diptest/diptest.pdf

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196–219. https://doi.org/10.1080/10584609.2017.1334018

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Metaxas, P., & Twittertrails Research Team. (2017). Retweets indicate agreement, endorsement, trust: A meta-analysis of published Twitter research. *ArXiv Preprint*. Retrieved from http://cs.wellesley.edu/~pmetaxas/WorkingPapers/Retweet-meaning.pdf

Mihaylov, T., Georgiev, G., & Nakov, P. (2015). Finding opinion manipulation trolls in news community forums. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, (July), 310–314. https://doi.org/10.18653/v1/K15-1032

Müller, K., & Schwarz, C. (2018). Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3082972

Muralidharan, S., & Men, L. R. (2015). How peer communication and engagement motivations influence social media shopping behavior: Evidence from China and the United States. *Cyberpsychology, Behavior, and*

Social Networking, *18*(10), 595–601. https://doi.org/10.1089/cyber.2015.0190

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. https://doi.org/10.1007/BF00316552

Pariser, E. (2011). *The Filtter Bubble: What the internet is hiding from you*. New York, New York, USA: The Penguin Press.

Paul, C., & Matthews, M. (2016). *The Russian "firehose of falsehood" propaganda model*. https://doi.org/10.7249/PE198

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465.supp

Pettigrew, T. F., & Tropp, L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology*, *38*, 922–934. https://doi.org/10.1002/ejsp

Postmes, T., Spears, R., & Lea, M. (1998). Building or breaching social boundries? SIDE effects of computer mediated communication. *Communication Research*, *25*(6), 689–715.

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. *Proceedings Ofthe 55th Annual Meeting Ofthe Association for Computational Linguistics*, 729–740. https://doi.org/10.18653/v1/p17-1068

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2018). Language models are unsupervised multitask learners. *ArXiv*. Retrieved from https://github.com/codelucas/newspaper

Rani, N. (2018). Social media in India: A human security perspective. *The Research Journal of Social Sciences*, *9*(10), 43–52.

Roeder, O. (2018). *Why we're sharing 3 million Russian troll tweets. FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/

Roozenbeek, J., & Linden, S. Van Der. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *9877*, 1–11. https://doi.org/10.1080/13669877.2018.1443491

Salihefendic, A. (2015). How Reddit ranking algorithms work – Hacking and Gonzo – Medium. Retrieved March 14, 2019, from https://medium.com/hacking-and-gonzo/how-reddit-ranking-algorithms-work-ef111e33d0d9

Sanovich, S. (2017). *Computational propaganda in Russia: The origins of digital misinformation* (No. 2017.3). Retrieved from https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Russia.pdf

Shearer, E., & Gottfried, J. (2017). *News use across social media platforms 2017*. Retrieved from http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2017). Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New Media and Society*, *19*(8), 1214–1235. https://doi.org/10.1177/1461444816634054

Silva, W., Santana, Á., Lobato, F., & Pinheiro, M. (2017). A methodology for community detection in Twitter. *Proceedings of the International Conference on Web Intelligence - WI '17*, 1006–1009. https://doi.org/10.1145/3106426.3117760

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization Of Social Media*. New York, USA: Houghton Mifflin Harcourt Publishing Company.

Smith, A. G., Suedfeld, P., Conway, L. G., & Winter, D. G. (2008). The language of violence: Distinguishing terrorist from nonterrorist groups by thematic content analysis. *Dynamics of Asymmetric Conflict*, *1*(2), 142–163. https://doi.org/10.1080/17467580802590449

Sobolev, A. (2018). *Fantastic beasts and whether they matter: How pro-Government "trolls" influence political conversations in Russia*. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/266563_26294ad8ef0f47b2b94bea9e33eb8f88.html#(1)

Statista. (2019). Twitter: Number of active users 2010-2018. Retrieved February 25, 2019, from https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Stecklow, S. (2018). Why Facebook is losing the war on hate speech in Myanmar. *Reuters*. Retrieved from https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

Stewart, Leo G., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network. *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*., 6. https://doi.org/https://doi.org/10.475/123_4

Stewart, Leo Graiden, Arif, A., Nied, A. C., Spiro, E. S., & Starbird, K. (2017). Drawing the lines of contention: Networked frame contests within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*, *1*(CSCW), 1–23. https://doi.org/10.1145/3134920

Streufert, S., & Suedfeld, P. (1965). Conceptual structure, information search, and information utilization. *Journal of Personality and Social Psychology*, *2*(5), 736–740. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5838772

Suedfeld, P., & Bluck, S. (1988). Changes in Integrative Complexity prior to surprise attacks. *Journal of Conflict Resolution*, *32*(4), 626–635. https://doi.org/10.1177/0022002788032004002

Summers, E. (2017). BlackLivesMatter Tweets 2016 . Retrieved December 2, 2020, from https://archive.org/details/blacklivesmatter-tweets-2016.txt

Sunstein, C. R. (2017). *#Republic: divided democracy in the age of social media*. Princeton, New Jersey, United States: Princeton University Press.

Tetlock, P. E., Peterson, R. S., & Berry, J. M. (1993). Flattering and unflattering personality portraits of integratively simple and complex managers. *Journal of Personality and Social Psychology*, *64*(3), 500–511. https://doi.org/10.1037/0022-3514.64.3.500

Turner, J. C., Davidson, B., & Hogg, M. A. (1990). Polarized norms and social frames of reference: A test of the self-categorization theory of group polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101

UK Department for Digital Culture Media and Sport Committee. (2018). *Disinformation and "fake news": Interim Report*. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/363.pdf

UK Department for Digital Culture Media and Sport Committee. (2019). *Disinformation and "fake news": Final report*. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf

Venables, B., Bates, D. M., Firth, D., & Ripley, M. B. (2018). Package "MASS": Support functions and datasets for venables and Ripley's MASS. *CRAN*. https://doi.org/ISBN 0-387-95457-0

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*, 1146–1151.

Weedon, J., Nuland, W., & Stamos, A. (2017). *Information operations and Facebook*. https://doi.org/10.1016/B978-1-4377-2003-7.00058-3

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, *65*(4), 699–719. https://doi.org/10.1111/jcom.12164

Woolley, S. C., & Howard, P. N. (2016). Political communication, computational propaganda, and autonomous agents. *International Journal of Communication*, *10*, 4882–4890.

Wright, J., & Anise, O. (2018). Don 't @ me: Hunting Twitter bots at scale. *Black Hat*, 1–43. Retrieved from https://duo.com/blog/dont-me-hunting-twitter-bots-at-scale

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal attacks seen at scale. *International World Wide Web Conference*, 1–9. https://doi.org/10.1145/3038912.3052591

Yardi, S., & Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, *30*(5), 316–327. https://doi.org/10.1177/0270467610380011

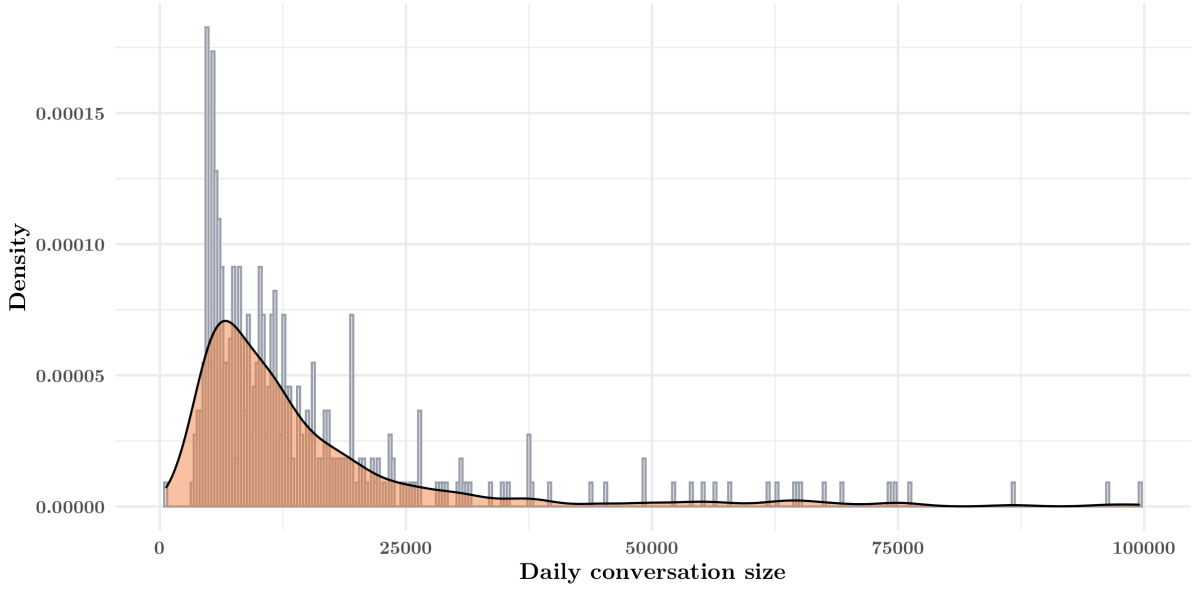
Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G., & Blackburn, J. (2018). Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the Web. *ArXiv*. https://doi.org/10.1007/s11082-007-9133-1

Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G., & Blackburn, J. (2019). Who let the trolls out? Towards understanding state-sponsored trolls. *ArXiv*. Retrieved from https://arxiv.org/pdf/1811.03130.pdf

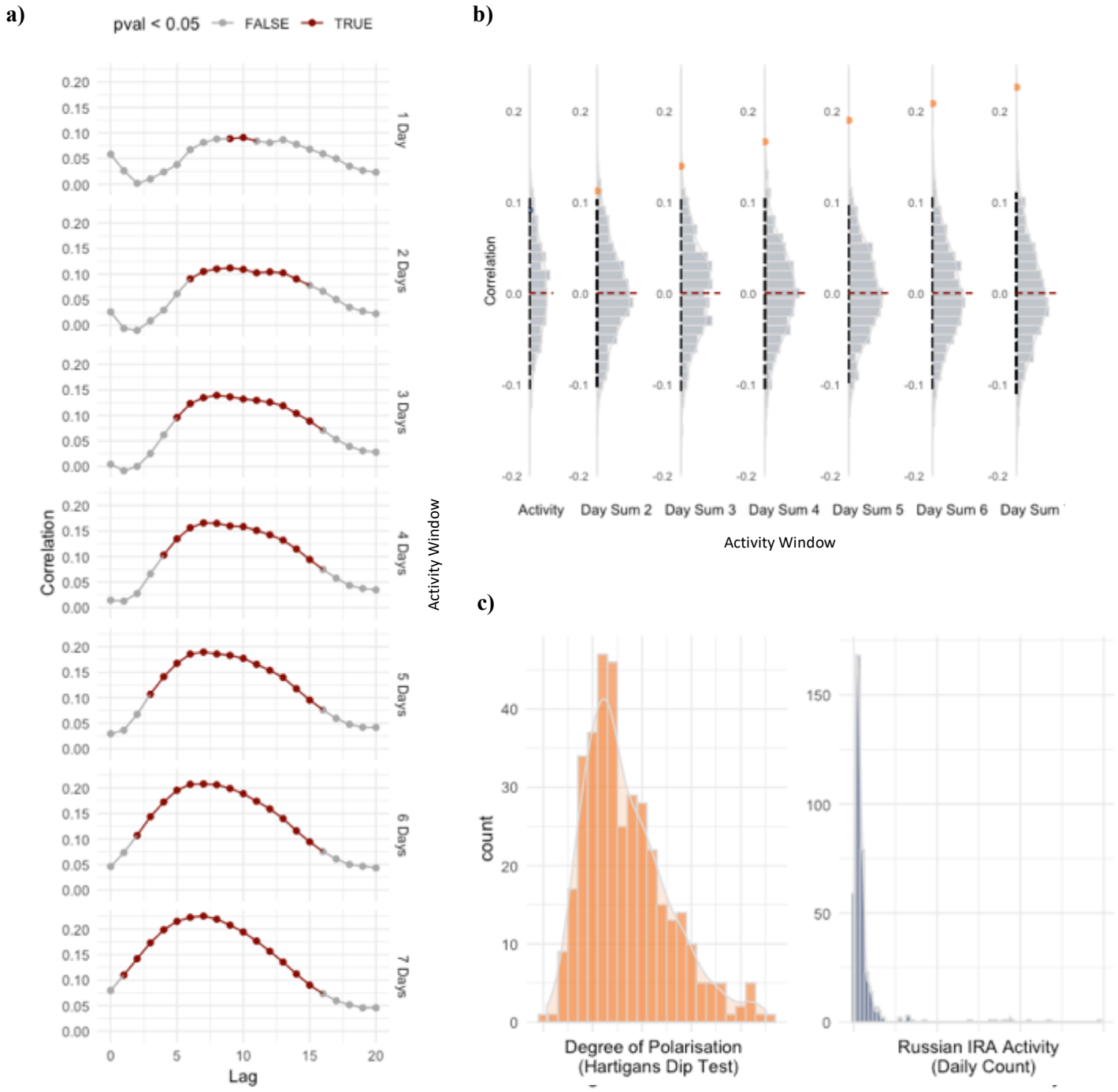
Zhou, L., Wang, W., & Chen, K. (2016). Tweet properly: Analyzing deleted Tweets to understand and identify regrettable ones. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 603–612. https://doi.org/10.1145/2872427.2883052

Supplementary Information (SI) for Chapter 5:

How hostile Information Operations increase the polarisation, intergroup antagonism, and hate of online conversations



SI Figure 1 –Distribution of daily activity in the #BlackLivesMatter / #BLM Twitter conversation for the collection period between 29 January 2016 and 18 March 2017.
Additional datapoints not shown for visibility: seven days with > 100,000 messages and < 200,000 and six days with > 200,000 messages



SI Figure 2(a-c) –
Results from Twitter lagged permutation test using raw un-transformed data
a) Correlations between the degree of daily polarisation in BLM conversations on Twitter and preceding total Russian IRA Activity over various periods (1-7 days). Red dots show significant correlations.
b) Significance effects for max correlations for each activity window compared to distribution obtained by chance (grey) as calculated with a permutation test. (orange : $p < 0.05$, blue = non significant)
c) Raw data distributions of polarisation and activity

SI Table 1 – Statistical results for the lagged permutation test across activity window and lag period. Bold indicates statistical significance at the p=0.005 level

Sum Period (Days)	Lag Period (Days)	Correlation	p
1	0	0.011	0.419
	1	0.005	0.467
	2	-0.006	0.548
	3	-0.026	0.688
	4	-0.048	0.820
	5	-0.038	0.768
	6	-0.013	0.599
	7	-0.014	0.610
	8	-0.019	0.648
	9	-0.022	0.665
	10	-0.025	0.690
	11	-0.040	0.784
	12	-0.047	0.821
	13	-0.049	0.822
	14	-0.051	0.832
	15	-0.063	0.887
	16	-0.065	0.894
	17	-0.074	0.920
	18	-0.086	0.951
	19	-0.088	0.955
	20	-0.084	0.945
2	0	0.009	0.428
	1	-0.009	0.564
	2	-0.029	0.709
	3	-0.043	0.788
	4	-0.044	0.798
	5	-0.010	0.573
	6	0.003	0.473
	7	0.006	0.451
	8	0.002	0.484
	9	-0.001	0.514
	10	-0.009	0.575
	11	-0.022	0.663
	12	-0.032	0.728
	13	-0.034	0.740
	14	-0.045	0.803
	15	-0.053	0.841
	16	-0.062	0.878
	17	-0.081	0.934
	18	-0.095	0.962
	19	-0.095	0.964
	20	-0.101	0.971
3	0	0.007	0.441
	1	-0.012	0.594
	2	-0.032	0.739
	3	-0.031	0.723
	4	-0.006	0.544
	5	0.020	0.354
	6	0.037	0.242
	7	0.041	0.217
	8	0.039	0.227
	9	0.031	0.280
	10	0.021	0.348
	11	0.006	0.459
	12	-0.005	0.541
	13	-0.014	0.611
	14	-0.025	0.680
	15	-0.040	0.772
	16	-0.059	0.864
	17	-0.082	0.936
	18	-0.093	0.960
	19	-0.100	0.969
	20	-0.103	0.971

Sum Period (Days)	Lag Period (Days)	Correlation	p
4	0	0.011	0.409
	1	-0.011	0.581
	2	-0.012	0.590
	3	0.014	0.394
	4	0.033	0.264
	5	0.059	0.131
	6	0.071	0.087
	7	0.077	0.069
	8	0.070	0.092
	9	0.057	0.141
	10	0.047	0.184
	11	0.034	0.259
	12	0.020	0.345
	13	0.010	0.416
	14	-0.004	0.527
	15	-0.028	0.691
	16	-0.049	0.816
	17	-0.068	0.893
	18	-0.082	0.930
	19	-0.085	0.940
	20	-0.084	0.940
5	0	0.013	0.405
	1	0.009	0.438
	2	0.032	0.279
	3	0.053	0.156
	4	0.073	0.080
	5	**0.095**	**0.035**
	6	**0.108**	**0.020**
	7	**0.108**	**0.020**
	8	**0.096**	**0.033**
	9	0.082	0.058
	10	0.073	0.082
	11	0.058	0.136
	12	0.043	0.209
	13	0.031	0.283
	14	0.010	0.422
	15	-0.015	0.613
	16	-0.034	0.741
	17	-0.054	0.848
	18	-0.065	0.890
	19	-0.065	0.894
	20	-0.061	0.877
6	0	0.036	0.248
	1	0.054	0.151
	2	0.071	0.087
	3	**0.092**	**0.040**
	4	**0.109**	**0.018**
	5	**0.131**	**0.006**
	6	**0.136**	**0.005**
	7	**0.133**	**0.006**
	8	**0.120**	**0.010**
	9	**0.105**	**0.023**
	10	**0.093**	**0.037**
	11	0.077	0.073
	12	0.061	0.125
	13	0.043	0.208
	14	0.020	0.360
	15	-0.003	0.534
	16	-0.023	0.674
	17	-0.041	0.785
	18	-0.048	0.823
	19	-0.046	0.814
	20	-0.047	0.814

Sum Period (Days)	Lag Period (Days)	Correlation	p
7	0	0.081	0.058
	1	**0.086**	**0.049**
	2	**0.104**	**0.023**
	3	**0.121**	**0.009**
	4	**0.138**	**0.003**
	5	**0.155**	**0.001**
	6	**0.156**	**< 0.001**
	7	**0.150**	**0.002**
	8	**0.137**	**0.004**
	9	**0.119**	**0.012**
	10	**0.106**	**0.022**
	11	**0.089**	**0.046**
	12	0.067	0.101
	13	0.047	0.185
	14	0.026	0.313
	15	0.002	0.489
	16	-0.015	0.611
	17	-0.029	0.705
	18	-0.035	0.734
	19	-0.037	0.749
	20	-0.036	0.747

SI Table 2 – Statistical results for causal impact analysis across the three conversation measures; Integrative complexity, toxicity and identity attack and across two time periods; 100 and 25 comments

			Text Analysis Measures											
			Integrative Complexity				Toxicity				Identity Attack			
Subreddits		Time Span (Comments)	Mean Observed	Mean Predicted	P	Cumulative Change	Mean Observed	Mean Predicted	P	Cumulative Change	Mean Observed	Mean Predicted	P	Cumulative Change
All	**100**		***1.41 ± 0.004***	***1.42 ± 0.006***	***0.035***	***-1% ± 0.51***	*0.26 ± 0.004*	*0.27 ± 0.001*	*0.148*	*-1% ± 0.77*	*0.20 ± 0.001*	*0.20 ± 0.001*	0.064	-1% ± 0.51
	25		*1.41 ± 0.003*	*1.42 ± 0.002*	*0.152*	*-1% ± 0.77*	*0.27 ± 0.0005*	*0.27 ± 0.002*	*0.303*	0% ± 0.77	*0.20 ± 0.001*	*0.21 ± 0.002*	*0.379*	*0% ± 1.02*
Political	**100**		*1.43 ± 0.006*	*01.43 ±.01*	*0.373*	*0% ± 0.77*	*0.27 ± 0.0003*	*0.27 ± 0.002*	*0.301*	*1% ± 1.28*	*0.21 ± 0.0002*	*0.21 ± 0.002*	0.085	*2% ± 1.28*
	25		*1.42 ± 0.02*	*1.43 ± 0.03*	*0.178*	*-1 ± 1.02*	***0.28 ± 0.001***	***0.27 ± 0.002***	***0.019***	***3% ± 1.53***	*0.21 ± 0.001*	*0.21 ± 0.004*	*0.137*	*2% ± 2.6*
Non-Political	**100**		*1.23 ± 0.005*	*1.24 ± 0.007*	*0.329*	*0% ± 0.77*	*0.23 ± 0.003*	*0.23 ± 0.002*	*0.378*	*0% ± 1.28*	0.13 ± 0.0002	0.13 ± 0.002	0.482	*0% ± 1.53*
	25		***1.26 ± 0.007***	***1.23 ± 0.002***	***0.005***	***2% ± 0.77***	*0.22 ± 0.001*	*0.23 ± 0.004*	*0.118*	*-2% ± 1.80*	***0.14 ± 0.001***	***0.13 ± 0.002***	***0.001***	***10% ± 2.04***

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Measuring the effect of Russian Internet Research Agency information operations in online conversations
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Gallacher, J. D. & Heerdink, M. W. (2019). Measuring the effect of Russian Internet Research Agency information operations in online conversations. *Defence Strategic Communications.* 6, 155-198

Student Confirmation

Student Name:	John Gallacher		
Contribution to the Paper	John Gallacher (J.G) was the primary author on this paper, receiving guidance and supervision from Marc Willem Heerdink (M.W.H). The NATO StratCom COE, and in particular Rolf Fredheim, also provided guidance in conceiving and motivating this study, although their contribution did not warrant co-authorship. J.G. conceived the study, collected the data, applied the language models and implemented the causal impact modelling. M.W.H developed the approach for modelling polarization within Twitter network, which J.G then refined and applied to the Twitter networks for the BLM conversation. J.G. designed and implemented the statistical analysis, and created the data visualizations. J.G wrote the manuscript, with M.W.H providing valuable feedback and comments on early versions of the manuscript and providing assistance implementing reviewers' feedback within the peer-review process.		
Signature		Date	11/01/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Marc W. Heerdink, Ph.D.		
Supervisor comments I hereby certify that the description of my role in this paper is accurate. John Gallacher led all stages of the project, and only requested input about specific things. The paper is therefore primarily his work.		
Signature		Date 14/01/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

General Discussion

This thesis has examined the role that social media plays in fostering and exacerbating intergroup conflict. As the prevalence of online communication increases, so does the importance of understanding its impact on intergroup relations both online and offline. Over five separate chapters this work has investigated the prevalence and impact of hostile intergroup contact on offline intergroup relations, the influence of extreme outgroup denigration in ingroup discussions, and the effect of hostile manipulation on the dynamics of online conversations. Together, these results provide novel insight into how the Internet may be driving increases in intergroup conflict, extremism, and violence.

In the following sections we discuss the main conclusions from each chapter, how these connect to describe the role of the Internet on intergroup relations more broadly, and the remaining research questions this work has highlighted. We then consider what these findings tell us about how the nature of online discussions could be adjusted to promote more positive intergroup relations and mitigate some of the negative consequences observed thus far.

The nature of intergroup interactions online

Our first results demonstrate that communication between opposing groups does indeed take place online, challenging the now traditional 'echo-chamber' proposal of online communication (Pariser, 2011; Sunstein, 2017). The results in Chapter 1 suggest that online social media users are exposed to opinions that challenge their own and communicate directly with outgroup members. However, this does not result in improvements in intergroup relations. Instead, we find that within intergroup conflict situations, the level of intergroup communication is predictive of future offline violence between these groups. Inspecting the nature of this intergroup communication sheds light onto this effect, and reveals that typically this communication is negative, one-sided, short lived, and highly confrontational. These findings suggest that opposing groups may use unstructured online environments to engage with one another in hostile ways, and this may reflect a worsening of relationships, in turn explaining the observed increases in physical violence offline.

These intergroup communications therefore fail to meet the criteria for positive intergroup contact (Allport, 1954), and instead more closely resemble negative contact (Brown & Hewstone, 2005; Graf, Paolini, & Rubin, 2014), or simply intergroup 'interaction' (MacInnis & Page-Gould, 2015). Indeed, this latter term is the one used in Chapter 1 to describe online intergroup communications that are

short-lived and transient. Prior work has shown that similar short intergroup interactions offline can produce negative outcomes for individuals involved, such as anxiety and discomfort (Hyers & Swim, 1998; Littleford, Wright, & Sayoc-Parial, 2005; Shelton, 2003), and can threaten their sense of social identity (Shelton, Richeson, & Vorauer, 2006). This may better reflect our observations in Chapter 1 than the literature on intergroup contact which predicts positive outcomes. These need not be mutually exclusive views however, and it has been suggested that these short and temporary intergroup interactions should be considered individual units of intergroup contact (MacInnis & Page-Gould, 2015). In other words, there may be a threshold for intergroup 'contact' which occurs after a certain critical number of intergroup interactions, and once this threshold is reached future interactions are more likely to be positive and improve intergroup attitudes (Blascovich, Mendes, Hunter, Lickel, & Kowai-Bell, 2001; Page-Gould, Mendoza-Denton, & Tropp, 2008; Paolini, Hewstone, Voci, Harwood, & Cairns, 2006). In this way, there may be short-term costs to intergroup interactions, (including heightened intergroup anxiety, discrimination or prejudice) but long-term gains from intergroup contact (lower prejudice, discrimination, and improved intergroup relations). Our results from Chapter 1 contribute to this debate by providing partial support this hypothesis, and suggest that in unstructured online spaces these short-term costs are occurring, but without the longer-term gains, at least for the groups and conflict situations which we observed. Whether more prolonged and repeated intergroup interactions occur elsewhere online, and whether they reach the theoretical tipping point for positive intergroup contact, remains to be explored.

A further key result from Chapter 1 is to provide evidence that ingroup conversations are also predictive of future offline violence. In particular, lower integrative complexity of ingroup conversations (i.e. more one-sided and less nuanced conversations) was associated with future violence. This aligns with previous findings that decreases in integrative complexity are linked with the deterioration of group relations (Park & DeShon, 2018; Suedfeld & Bluck, 1988), and that individuals who display higher integrative complexity are less prejudiced and better able to resolve conflicts with outgroup members (Tetlock, Peterson, & Berry, 1993). The relationship we found also supports evidence that higher certainty of one's opinions leads to a greater intolerance of those groups who hold different views (Garrett, Weeks, & Neo, 2016). Conversely it challenges recent research which suggests than in some extreme online spaces, it is the subtle undermining and marginalising of opposing viewpoints which keeps users engaged. On the Facebook pages we study, this nuanced undermining of opposing viewpoints was less common, and it is the low-complexity conversations which predicted future offline violence. Anecdotally, investigation of these ingroup discussions

indicated that they instead contained a high degree of denigration of outgroups – suggesting that this hate may also play a role in reducing intergroup relations. This observation warranted further investigation and motivated the work presented in Chapters 2, 3 and 4.

The importance of both ingroup and intergroup communication on intergroup relations highlighted by our findings is mirrored in wider observations of extremist groups behaviour online. Ebner, (2017) suggests that the nature of the online world plays a role in limiting extremist groups overall exposure to cross cutting content, however this effect is interrupted by irregular spikes whereby intense intergroup interaction does take place due to external prompts. Our results may support this theory, with the observed intergroup interactions occurring prior to offline events, and likely motivated by these future events. Linked to this idea are suggestions that *'trench warfare'* may be a more appropriate term for the nature on intergroup communications online (Karlsen, Steen-Johnsen, Wollebæk, & Enjolras, 2017) than echo-chambers. This idea builds on evidence that social media provides the opportunity for users to interact with like-minded people and those with opposing views at the same time. In this way, attitudes are reinforced through a combination of contradiction as well as confirmation, with prior ingroup interactions reinforcing the positions which are then confrontationally debated in intergroup interactions. Overall, our work supports this theory by providing empirical evidence backing this hypothesis, and highlights the need to think more broadly beyond theories of echo-chambers driving online polarisation, and instead consider the interplay between both intragroup and intergroup communications on intergroup relations and conflict.

Detecting outgroup denigration and hate speech online

A logical progression from this finding that the nature of ingroup communication is associated with future offline intergroup violence (Chapter 1) is the exploration of the impact of outgroup derogation in ingroup discussions. This is mirrored in Berger's (2019) suggestion that *"extremism [is an] inherently social activity, usually carried out by individuals because they dramatically overvalue their membership in a particular social grouping"* and that extremist positions are defined by the belief that survival of the ingroup is inseparable from direct action, such as derogation, against the outgroup (Berger, 2018). This outgroup derogation often takes the forms of hate speech. Exactly defining the term "hate speech" is contested (Sellars, 2016) and in this thesis we took a group-level approach to hate speech definition, allowing us to explore the role of hate speech in a range of intergroup conflict conditions. Detecting online hate is an ongoing challenge however, and in Chapter 2 we present an

innovative approach that could help to improve this, which combines the latest advances in natural language processing and machine learning, as well as datasets from four different social media platforms.

We show that our cross-platform approach for hate speech detection outperformed existing models, both on data similar to that which these models are trained on, and for unseen data from a novel social media platform not included in the original training set. This highlights the opportunities of cross-platform approaches to improve automatic detection of online hate speech. In addition, we found that performance on novel platform data was easily improved by adding a new model trained on a small dataset from this platform, demonstrating the flexibility and ease of updating provided by this method. This approach helps to solve the challenge of sparsity in hate speech detection training data (Schmidt & Wiegand, 2017) by leveraging datasets from across multiple social media platforms. This cross-platform data provides both a higher quantity of data and greater variability. By developing a hate speech detection model which works for conversations on fringe platforms it lays the groundwork for the subsequent Chapters which investigate the nature and impact of these conversations on users.

The impact of online hate in ingroup discussions

Building on our results that ingroup discussion are linked to offline violence (Chapter 1), we investigated the drivers of hate speech expression within ingroup discussions online (Chapter 3). For these conversations we focused on a fringe social media platform, Gab, popular with the far-right and infamous for hosting a large degree of hate speech and outgroup denigration. Our results demonstrate that by participating in the conversations on Gab, individuals become increasingly likely to express hate themselves, and this effect appears driven at least in part by patterns of exposure to hate from other ingroup members. These results shed light on the role of group socialisation and social contagion in expressions of outgroup hate, with users adopting the negative outgroup attitudes of those they interact with. These results therefore support theoretical social models of group extremism where discussions with the ingroup are key (Brady, Crockett, & Van Bavel, 2020; Reicher, Haslam, & Rath, 2008; Smith, Blackwood, & Thomas, 2019), and highlights the importance of group norms on driving outgroup prejudice online, reflecting offline evidence that individuals closely adhere to social norms when expressing prejudice (Crandall, Eshleman, & O'Brien, 2002). Supporting our earlier results that ingroup discussion are linked to offline violence (Chapter 1), we find that ingroup discussions are important in driving intergroup conflict, and that influential ingroup members are central as they

drive the patterns of exposure of other ingroup members to hate. This supports long-standing theories of the importance of social group membership on driving individual social traits (Simmel, 1972), but challenges the concept that this intergroup friction is a force for social stability (Coser, 1956). Recent qualitative evidence from interviews with former members of far-right extremist groups supports our results (Google Jigsaw, 2020). These individuals cited the strength of the shared community provided by the online world and the sense of belonging these groups provided as the key drivers of their radicalisation. Interacting with other ingroup members was highlighted as a vital step in learning the specific group norms, coded language, violent ideals, and to reinforce derogatory positions against perceived out-groups (Lewis & Marwick, 2017). Our work in Chapter 3 provides quantitative evidence for these conclusions.

This relationship between online activity and hate expression was not linear however, nor did all users display an upwards trajectory of hate posting, challenging linear process models of extremism (McCauley & Moskalenko, 2008; Moghaddam, 2005) or that there is a single pathway for radicalisation (Kruglanski et al., 2014). Instead, some users appeared to arrive onto the fringe platforms with pre-existing hate positions which they expressed, and then the nature of this hate changed over time as they also adopted the hateful positions of the group. This was further demonstrated by evidence that users increased the breadth of their targets of hate, and that effects of exposure to hate speech are not limited to one type of hate and appears to transfer across hate targets. Together, our results supports wider evidence that some online users 'search out' online extremist material and hateful spaces (Hosseinmardi et al., 2020; Ribeiro, Blackburn, et al., 2020) and that searching out this material can have a stronger effect on attitudes than passive encounters with the same material (Pauwels & Schils, 2016). There are also parallels with traditional ideas of generalised prejudice, where prejudice against one group can easily lead to prejudice against others (Crandall et al., 2002). Sherif (1948) argued that '*there cannot be separate psychologies of prejudice in relation to this or that group, but that they are specific cases of the general picture of prejudice'.* Recent evidence supports this and suggests that approximately half of individuals' variability in negative attitudes towards multiple outgroups (e.g., against various ethnic, age, or religious groups) can be traced to the same underlying factor of generalised prejudice (Bergh & Akrami, 2016; Bergh, Akrami, & Ekehammar, 2012; Ekehammar & Akrami, 2003), while social factors can explain the remaining variance (Akrami, Ekehammar, & Bergh, 2011). Our results suggest that by giving easy access to hateful material, fringe social media platform may therefore be promoting this process of

generalised prejudice by allowing those predisposed to target multiple groups, and to influence those susceptible to social influence.

In addition to the role of activity on the platforms and interactions with ingroup members, we found that the time users spent online also drives increased hate speech expression. This suggests that passively viewing hateful conversations, without actively participating, can also increase users' own hateful attitudes. This supports prior evidence that social media users are influenced as much by the content that they passively consume, or 'listen to', as that which they interact with (Settle, 2018), and highlights how important it is to consider the wider impact of online hate beyond those taking part in the conversations, but on the group as a whole. These results are likely applicable to other platforms and types of extremism, and have important implications for our understanding of hostile intergroup relations and the role of the Internet in intergroup conflict and hate. As fringe social media platforms become more popular with far-right extremist groups, these patterns of hate exposure will increase, potentially driving dangerous trends in hate towards outgroups.

The connection between online and offline environments

Across multiple chapters, we find evidence of a close connection between the nature of the online world and offline violence. We find that both intergroup and intragroup communications on Facebook event pages are predictive of both the presence and degree of violence at these future events (Chapter 1). Subsequently we explored the relationship between online hateful discussions and offline hate crimes from far-right groups, and found that the degree of hate speech on Gab was predictive of future offline hate crimes (Chapter 4). This effect was largest within specific hate types, and in areas with higher levels of interest in fringe social media platforms. Together these results support a growing body of literature which shows that online conflict precedes offline conflict (Alsaedi, Burnap, & Rana, 2017; Mooijman, Hoover, Lin, Ji, & Dehghani, 2018; Müller & Schwarz, 2020a, 2020b; Williams, Burnap, Javed, Liu, & Ozalp, 2019). Further work is still needed to unpack the explanations for this relationship, but this finding is becoming increasingly robust, and exists across multiple Western locations and intergroup conflict situations.

Conversely, we also find that offline events can impact upon online hate speech, and specifically that Islamophobic online hate speech spiked following offline Islamic terror attacks (Chapter 4). Again, these findings support a growing body of evidence that online hate speech increases following perceived offline attacks against the ingroup (Awan & Zempi, 2016; Burnap et al., 2014; Vidgen,

Yasseri, & Margetts, 2019; Williams & Burnap, 2016). Interestingly we also found links between our measure of polarisation in Chapter 5 and offline violence. The #BlackLivesMatter conversation visualised in Chapter 5 (Figure 2a & 2b) from 7th July 2016 was selected for presentation as it clearly demonstrated a bimodal conversation using our measure of polarisation. Upon later inspection, we discovered that on this day, five police officers had been killed at a shooting in Dallas (NBC, 2016), and this sparked considerable online backlash, with some users blaming the Black Lives Matter movement, even though the movement itself condemned the actions (BBC, 2016). Together, these results highlight the complex interplay between the online and offline worlds, and how developing robust metrics to measure online phenomena can be useful in understanding this dynamic.

In addition, our findings demonstrate that this link between online hate and offline violence does not only occur for specific protected groups (as defined by law), but also for groups which do not always fall under this category of protected characteristics. We find for example that immigration status, and refugees in particular, are a common target of hate speech online, and that this hate speech is predictive of offline violence against these groups in both the US and UK. This raises important questions as to whether the current online hate speech legislation in the UK is robust enough to prevent offline harm, and also how these laws should be adapted to account for changing phenomena online. Given the lack of legislation surrounding hate speech in the US, these results may also provide ground for a possible argument in favour of the implementation of such laws. A further example of how it may be necessary to widen the discussion about the relationship between online hate and offline violence beyond protected characteristics alone comes from the recent case of violence in the US against politicians in the January 6th 2021 'storming of the Capitol' riots, which were planned and amplified online prior to the event (Evans, 2021). Politicians, and particular those who are female, were targeted by far-right groups, even though neither of these characteristics (gender or profession) are traditionally considered as protected characteristics. Mitigating these harms is vital, but will likely require a broader look at how groups from across society are targeted with hate, vitriol, and aggression online, and how these translate into offline violence.

The role of the Internet in mutual radicalisation

While evidence for the connection between online and offline hate is growing, we take this evidence a step further and present a model of the role that social media may play in driving mutual radicalisation between opposing extremist groups. In addition to the online/offline effects discussed above, using data from Google Search interest we show that offline Islamophobic violence preceded

increases in online interest in Islamic extremism (Chapter 4). Together, these results highlight the key role of the Internet in processes of mutual radicalisation (Moghaddam, 2018) and support a cyclical process whereby the online environment mediates both the instigation of offline violence, and the response to it.

This work once again shows that the 'echo-chamber' view of how online communication drives extremism (Pariser, 2011; Sunstein, 2017) is too simplistic, supporting our conclusions in Chapter 1. Groups are not isolated from one another online, as previous evidence has suggested (e.g. Conover, Ratkiewicz, & Francisco, 2011; Quattrociocchi, Scala, & Sunstein, 2016), but instead react dynamically with one another, internally discussing the actions of the outgroup, and responding to offline events. Opposing groups' actions fuel one another's rhetoric, which itself leads to retaliation either online or offline, and then further conflict. In this way, the dynamics of online polarisation are linked directly to processes of extremist violence. This supports Reicher & Haslam's (2016, p.1) thesis that *"the social psychology of group dynamics goes a long way towards explaining what drives ordinary people towards extremism"*. It also again highlights the importance of considering the dynamics of both ingroup and intergroup online communication when investigating the impact on the online world on intergroup relations.

Manipulation of the online environment can increase polarisation

In addition to hateful conversations naturally occurring online, manipulation of the online environment is becoming increasingly common, from both state-backed (Bradshaw & Howard, 2019) and far-right groups (Collins, Zadrozny, & Saliba, 2020; Nagle, 2017). Our results in Chapter 5 show that this malicious use of social media by inauthentic accounts can measurably affect the subsequent conversations held by genuine users.

Focusing on the activity of the Russian Internet Research Agency (Russian IRA), we find that the conversations they targeted often covered divisive topics, and their activity was effective in promoting hostile conversations among other users, likely increasing divisions among group lines. Russian IRA activity is therefore likely to be effective in increasing the distance between social groups, fuelling both ideological and affective polarisation (Mason, 2015). This in turn provides ideal circumstances for the distribution of disinformation (Del Vicario, Gaito, Quattrociocchi, Zignani, & Zollo, 2017; Garrett et al., 2016), which may exacerbate these effects further. More recent state backed information operations have looked to drive ethnic division in Eastern Europe (Gelava & Buziashvili, 2020;

Osadchuk, 2020), political division in the United Kingdom (Nimmo et al., 2019), conflict between western countries (Nimmo et al., 2020), and suppress dissent in Hong Kong (Twitter Safety, 2019). Given the increasing use of these tactics, further exploration of the effects is vital to increase social resilience.

While this research focused on state-backed groups, the results are likely also applicable to 'trolling' from extremist groups. Indeed, evidence has shown a convergence of tactics between far-right groups and Russian IRA activities (Davey, 2020). This is demonstrated in the wake of terror attacks in the UK where both alt-right groups and the Russian IRA promoted disinformation about the alleged actions of a Muslim woman in order to spread division (Gallacher & Fredheim, 2019; Innes, 2017; Wendling, 2018). Similarly, in the run up to the 2017 French presidential elections, far-right groups and Russian IRA activity focused on spreading the same divisive content gained from a leak of presidential candidate Emmanuel Macron's emails (Ferrara, 2017; Jeangene Vilmer, 2019; Polonski, 2018). This convergence of tactics shows how the online environment can be manipulated and how therefore success from one group is likely to be mirrored in success by others.

In Chapter 5 we also present novel methods of detecting online polarisation within Twitter networks, the applications of which may be useful in wider research into online intergroup dynamics.

Integration of these findings

Together, the findings in these five chapters give new insight into how the online world can affect intergroup relations and drive intergroup conflict – both online and offline. The Internet and social media have allowed groups to connect internally and externally more than ever, but this has led to undesired consequences. We demonstrate how groups interact with one another, and derogate the outgroup when communicating internally, and how both of these can play a role in reducing intergroup relations. The results are interconnected however, and by considering them together we can gain a greater understanding of the role of the Internet. The different effects studied and demonstrated in this thesis and how they might interact with one another, is summarised in Figure 1.

The interconnection between these effects is demonstrated by the dynamic between manipulating conversations and the negative impacts caused by these manipulations. For example, in Chapter 5 we demonstrate how antagonistic manipulation of the online environment can lead to increased toxicity

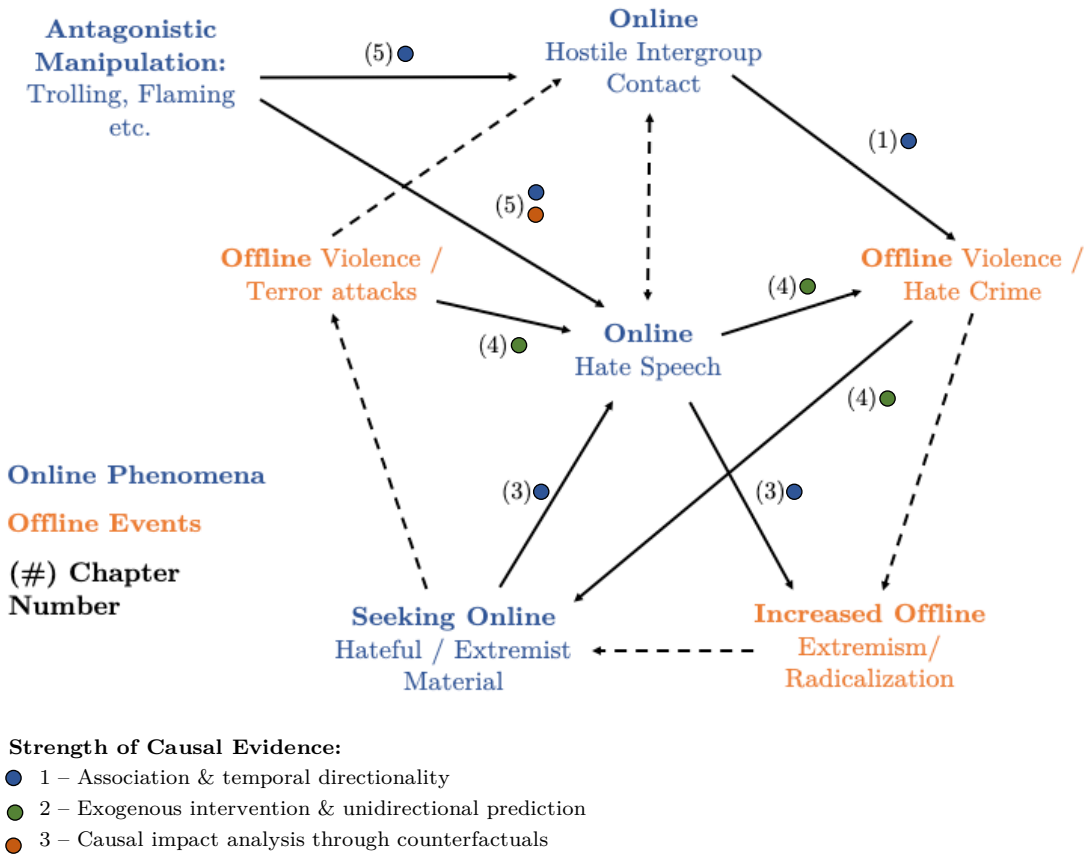


Figure 1 – How the dynamics of online communication can impact intergroup relations and intergroup conflict, both online and offline. The numbers near continuous arrows indicate in which chapter the effect is explored, dashed arrows represent relationships not investigated in this thesis, and the coloured dots relate to the strength of the causal evidence for that analysis (from lowest (blue) to highest (orange).

in these conversations. This toxicity is another form of extreme digital speech, similar to hate speech, which in Chapter 3 we show may lead to increasingly extreme and hateful positions for those exposed to it, and make them more likely to express hate speech themselves towards outgroups. In Chapter 4 we show how this increase in hate towards outgroups is linked with offline violence and hate crime, which itself is associated with the victims of the violence seeking out hateful material online themselves. Similar connections are shown between hostile intergroup contact, low integrative complexity of conversations, and future offline violence (Chapter 1). We showed this low integrative complexity can arise following antagonistic manipulation of online conversations (Chapter 5). This illustrates how these various effects can therefore become a circular and self-reinforcing process.

Additionally, in Chapter 4 we demonstrated how offline discrimination can lead individuals to search for extremist topics. In a similar way, in Chapter 3 we show that for users arriving in fringe platforms with pre-existing hateful attitudes (presumably after having searched for these spaces) their hate can

become intertwined with that of the platform as a whole, overall increasing their negative outgroup attitudes and driving extremism.

A number of aspects within this process are yet to be explored, however (dashed arrows in Figure 1). In particular, it is unclear whether hostile intergroup contact can lead to a 'backfire' effect, whereby after experiencing negative interactions with members of opposing groups users then return to their ingroup discussions and spread hate against the outgroup. If so, this would lead to a spiral of intergroup conflict, particular if this ingroup hate increased the likelihood of future hostile intergroup contact. Anecdotally, both of these effects have been observed within our datasets (Chapters 1 and 3/4), however this is yet to be tested formally. Doing so would shed further light on the interplay between online hate speech and intergroup relations. Similarly, identifying whether hostile intergroup contact increases following offline terror attacks would help identify the overall impact of these offline events on group relations.

Additionally, there is an important unresolved connection between seeking out online extremist material and whether this increases the chance of radicalisation up to the point of violence. While there is evidence that those who committed violent actions were heavily involved in extremist online communities (Evans, 2018, 2019; Gaudette, Scrivens, & Venkatesh, 2020), studying only those who commit violence does not tell us about all those (more numerous) individuals who searched for this content but did not radicalise. Exploring the nature of this relationship and the factors which increase or decrease the impact of searching for hateful content may give useful insights into how best to mitigate the impacts.

Future directions for research

Addressing these remaining unanswered questions will likely involve taking a wider look at the online ecosystem and constructing more cross-platform analyses to both understand the interplay between platforms and how users find and discover more extreme groups. Our results support suggestions that fringe platforms such as Gab may play a role in connecting more moderate right-leaning users to extreme far-right ideologies (McSwiney & Jasser, 2020) – the so called *'libertarian to alt-right pipeline'* (Hermansson, Lawrence, Mulhall, & Murdoch, 2020). However, the relationship between these spaces and more mainstream platforms where wider audiences can be reached, as well as their role as a

gateway onto more closed encrypted messaging apps or private servers (Donovan, Lewis, & Freidberg, 2018), needs to be explored further.

This challenge of platform interplay is further highlighted when considering the effect of content removal on the quantity of hateful content online. Studies have shown that when mainstream social media platforms enforce more stringent content moderation policies, this can reduce the amount of hateful content on that specific platform (Conway et al., 2019). However, it can also have the effect of shifting the hate onto other more esoteric platforms and groups swiftly recreate their prior connections and regain prominence in a different (usually less moderated) online space (Ribeiro, Jhaver, et al., 2020; Urman & Katz, 2020). In order to measure the effect of takedowns on the overall ecosystem, and the amount of hate online as a whole, it will therefore be important to concurrently measure multiple platforms in a consistent way. This highlights the importance of hate speech detection approaches which work accurately across multiple different platforms, and we hope that our advances in this area will prove useful in this future work.

Additionally, much of the work presented in this thesis has focused on hostile confrontation and outgroup denigration, however there are a range of other extremist behaviours which may contribute to an online community displaying signs of extremism and radicalisation (Grover & Mark, 2019), and may also impact on intergroup relations. Such behaviours include fixation on specific topics, identification with radical actions and radical role models, the leakage of plans to commit violence, among others (Cohen, Johansson, Kaati, & Mork, 2014). In order to build a wider picture of how groups extremity develops online, these additional factors should be considered, and their relationship with hate speech explored. For example, while the expansion of the number of groups an individual targets with hate might reflect a more generalised prejudice (Chapter 3), the fixation on a specific group as being responsible for multiple ills within society, and conspiratorial thinking around this group—as is common with Anti-Semitism for example (Baddiel, 2015)—may also be a warning sign of radicalisation and potential future violence, and this should be explored in future work.

Limitations

Given the observational nature of the research in this thesis it is likely that there are a number of unobserved factors which influence both a user's propensity to use these social media platforms, and their likelihood of expressing hate speech, to interact confrontationally with the outgroup, or to commit offline violence. These factors may include their offline social environment, socio-economic

status, employment status, along with many others (e.g., Allan, Glazzard, Jesperson, Reddy-Tumu, & Winterbotham, 2015; Knigge, 1998; Obaidi, Thompson, & Bergh, 2019). These individual factors and vulnerabilities are likely to be key determinants of susceptibility to group influence (Durodie & Ng, 2009), and future work should make attempts to account for some of these in a user's propensity to engage in intergroup conflict.

A further limitation of the work in this thesis is that it cannot provide conclusive causal evidence for the relationship between social media use and intergroup conflict, as it is observational in nature. We have made attempts to provide evidence which climbs the 'ladder of causality' (see Introduction; Pearl, 2018), and this work initially provides temporally directional associative evidence, subsequently evidence from unidirectional predictive models, and finally creates counterfactuals to assess what is likely to have occurred in the absence of observed events (Figure 1). However, while the approaches go beyond simple association, they cannot definitively show causation. Furthermore, in order to provide the strongest causal evidence, online interventions should be considered, where changes to the online environment are made and the impact on intergroup relations measured. For example, one could alter the type of content some users are exposed to and test how this affects subsequent behaviours, relative to a control group. Recent academic studies have tried this approach (e.g., Bail et al., 2018) and do indeed show measurable effects. However there are important ethical and practical considerations surrounding these methods and so their use has been limited, Social media companies have much a much larger ability to perform this type of experiment (e.g., Bond et al., 2012; Kramer, Guillory, & Hancock, 2014), however the ethical implications are often under-considered and this has led to backlash against social media platforms in the past (Selinger & Hartzog, 2016). One potential solution to this is to combine academic and industry research, with academic researchers providing oversight, transparency, and an open-source publication model, while social media companies provide the resources and platform to host the experiment. This type of partnership holds promise; however, challenges and conflicts of interest undoubtedly remain.

Policy implications

An important question arising from our results indicating an increasing cycle of intergroup conflict, both mediated by, and reliant on, social media and the Internet, is that of how to respond. Our results give evidence that radicalisation is a social process, and successful interventions to counter violent extremism will therefore need to incorporate social understanding.

The advances in detection of hostile and hateful online content have led to suggestions that this type of content should be taken down from social media sites, or blocked at the point of upload. Indeed, this approach has been promoted by Germany with their 'NetzDG' legislation passed in 2018 (The Bundestag, 2017), which requires social media platforms to remove online speech deemed illegal under domestic law within 24 hours (Tworek & Leerssen, 2019). This affirmative approach is mirrored by social media platforms which are taking an increasingly proactive stance on removing extreme and hateful material (Facebook Newsroom, 2017). This has led to worries however about the costs to freedom of speech if private enterprises are to decide the rules of permitted online communication, and potential biases which may exist within platforms' content moderation policies and automated detection approaches (Douek, 2020).

This is a difficult and challenging balance to make, which we do not attempt to definitively solve here. However, the research in this thesis does make some contributions to this debate. Evidence in Chapter 3 suggests that decreasing users' exposure to online hate speech may reduce some of the effects of socialisation around hateful norms, and therefore blocking or removing hateful content may reduce online radicalisation. This is unlikely to be the complete solution, however, as evidence also shows how users seek out hateful material online (Chapters 3 & 4), and may also seek out antagonistic discussions with outgroup members (Chapter 1). Therefore, if content is removed in one place, users may simply actively seek it out elsewhere. An undesirable consequence may be that this might detach them from more diverse settings where they may come into contact with moderate voices, for example by moving from a mainstream platform to a more fringe one. This may therefore lead to the situation whereby following removal from a mainstream platform, the most extreme subset of this community of users migrates to an alternative space, which is less moderated, and therefore this new community becomes more toxic and hateful, and this drives radicalisation (Ribeiro, Jhaver, et al., 2020). The appropriate trade-off between larger moderately hateful communities and smaller extremely hateful communities is currently unclear but warrants prompt investigation.

Additionally, our results demonstrate that even for concerted efforts into automatic hate speech detection with the latest techniques (e.g. BERT) in a comparatively constrained way, approximately 17% of hateful content will percolate through undetected (Chapter 2). Automatically removing online hate is made even harder by the rise in distributed online platforms (of with Gab is now a part) which cannot be centrally moderated (Bevensee, 2020).

Together, these challenges mean that we should use a more nuanced approach when thinking about how to deal with harmful online content, rather than focusing on a binary solution between leaving up or taking down content. This thesis gives insights into changes to the online environment which could be made to promote 'healthier' online communications and reduce or mitigate intergroup conflict. This is not to say that the most egregious content should not be removed, indeed it certainly should, but that this removal alone is not a panacea, and alternative (or complementary) approaches should also be explored. Here we discuss four potential approaches which build on the work in this thesis and may help bridge this gap: promotion of positive intergroup contact, promotion of positive ingroup voices, addition of inertia to online systems, and inoculation against hate. This not an exhaustive list, and more research is needed to identify other options.

Encourage sustained positive intergroup contact rather than short negative interactions

Chapter 1 demonstrated that in unstructured online spaces, communication between members of opposing groups better reflects hostile intergroup interactions rather than sustained positive intergroup contact. This suggests that effective policy responses will need to move beyond assuming that simply increasing the level of intergroup communication online will improve intergroup relations (Garimella, 2017; Obama, 2017). Offline evidence indicates that not all intergroup interactions have the same impacts on intergroup relations, and the success of the contact can be increased by the presence of certain facilitating conditions: equal status between groups, the sharing of common goals, intergroup co-operation, personal interaction, and support from authorities (Allport, 1954). Online intergroup contact which meets these criteria is most likely to have positive benefits on intergroup relations. These mechanisms have already been shown to help facilitate the positive effects of offline intergroup contact on reducing Islamophobia and prejudice in the immediate aftermath of Islamic terror attacks, reinforcing the potential benefits of positive intergroup contact even when intergroup tensions are heightened (Abrams, de Vyver, Houston, & Vasiljevic, 2017).

Some of the benefits of these facilitating conditions have already been shown online. For instance, equal status in one-on-one interactions between Israeli and Palestinian individuals via Facebook have been shown to improve intergroup attitudes (Schwab, Sagioglou, & Greitemeyer, 2019). Challenges to hateful positions posted by authoritative and credible sources can also be an effective measure to prevent future propagations of these harmful narratives (Ozalp, Williams, Burnap, Liu, & Mostafa, 2020). Additionally, online intergroup discussions on divisive issues which start from an interpersonal

perspective and an initial position of agreement are much more likely to lead to positive group level outcomes and a reduction in conflict than conversations which start from a confrontational initial position (Gehlback, Robinson, & Vriesema, 2018). In addition to direct contact between opposing groups, the sharing of common goals is demonstrated as a way to reduce prejudice and discrimination by evidence that when Muslim player Mohamed Salah joined Liverpool football club, it led to a 50% reduction in Islamophobic Tweets from Liverpool fans relative to comparative football clubs (Alrababah, Marble, Mousa, & Siegel, 2019). This effect was attributed to the salience of Salah's Muslim identity, with Muslims and non-Muslims sharing a common identity through the football club. Together, studies such as these show how many of the facilitatory conditions of positive offline intergroup contact may carry across to the online world.

In addition, evidence suggests that finding the 'sweet spot' in the balance of group saliency vs individual saliency in contact situations is key to extrapolating the positive impacts of intergroup contact to the groups as a whole and not just the individuals involved: the generalisation of this reduction in prejudice to the outgroup as a whole is key. Currently, diverse online conversations on non-political topics do not appear to have generalisable effects or lead to reductions in political intergroup prejudice because of the anonymity of the political group affiliation of those involved (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015) – participants are simply not aware that they have had a positive contact with a member of the opposing group. When social media users do get clues to the political identity of outgroup members however, they typically overestimate the extremity of outgroup members positions and this reduces the opportunities for personal interaction and perception of shared identities or common goals (Settle, 2018) – as also demonstrated in our results in Chapter 1. One solution to this may be to adjust the online environment so that personal and group-level identities are most salient at different times within an intergroup contact situation. Pettigrew (1998) proposed a three-stage model which looks to optimize successful offline contact and generalisation of attitude improvement and prejudice reduction to outgroups as a whole. Initially, users' personal identities should be emphasized to reduce intergroup anxiety and promote interpersonal connection (Miller & Brewer, 1984). Secondly, individuals' social categories and group-level identities should be made more salient in order to achieve generalisation of positive effects of interpersonal contact to the outgroup as a whole (Johnston & Hewstone, 1992). Finally, at a 're-categorisation' stage the group identities of both individuals are replaced with a superordinate group, encouraging all participants to think of themselves as members of the same larger community (Dovidio, Gaertner, Anastasio,

Bachman, & Rust, 1993). Future research should investigate how this type of intergroup contact could occur online.

If facilitatory mechanisms for positive intergroup contact can be achieved and sustained over the longer-term then positive effects of online intergroup contact are more likely to be achieved.

Promote more positive ingroup discussions

In addition to highlighting the importance intergroup communication, Chapters 3 and 4 revealed the role of ingroup discussions on setting hateful group norms. A corollary is that if users can learn outgroup hate, then they should also be able to learn positive outgroup associations from ingroup discussions, and these positive attitudes may also spread across outgroups.

Within hateful ingroup discussions offline, a single dissenting voice expressing anti-racist views can reduce tolerance for racist acts among the conversation participants, as well as reduce privately-held racist attitudes (Blanchard, Crandall, Brigham, & Vaughn, 1994). This demonstrates the power of ingroup members to disrupt hateful group norms. Conversely however, when the same dissenter instead passively accepted the racist acts, conversation participants also recommended acceptance of these acts. This mirrors the work on conformity from Asch's famous line judgement tasks, whereby the presence of just one confederate that goes against the majority choice can reduce conformity to group norms by as much as 80% (Asch, 1955). Equally, when social norms overtly sanction prejudice, outgroup discrimination becomes substantially less prevalent (Dovidio & Gaertner, 1991). This demonstrates the power of ingroup members who break norms around outgroup discrimination and prejudice.

This offline evidence therefore indicates that the promotion of moderate ingroup voices can play an important role in reducing group level extremity and challenging hateful group norms. This is not the standard situation online however, and currently the most extreme users are most active (Chapter 3; Barberá & Rivero, 2015; Preoţiuc-Pietro, Liu, Hopkins, & Ungar, 2017), and the most extreme content gains the most traction (Chapter 3; Brady, Wills, Jost, Tucker, & Van Bavel, 2017). This means that the group norms are typically set by the most extreme voices. Challenging these perceptions may be an effective strategy, as most users do not share these extreme views, but they are invisible due to their inactivity. Encouraging moderate users to share their ambivalence on contentious topics could help challenge this false consensus effect (Lavine & Johnston, 2012), whilst

also highlighting the true distribution of opinions on a topic, and making the online world more resistant to inflation at the extremes (Settle, 2018).

This approach is perhaps demonstrated best in the case of Reddit – where the basic responses to content are upvotes and downvotes (only positive and negative reactions are visible). The ambivalent responses, where users simply scroll past without interacting, are invisible and don't carry any weight in deciding what content rises to the top. If an element of inertia were introduced whereby for those users who simply scroll past a piece of content and do not react this were counted as a 'neutral' vote, then this would help 'dilute' the impact of the more 'extreme' positive and negative responses. In this way the proportion of active responses would need to be higher for content to rise to the top, and might help egregious content from going viral. Additionally, algorithmic adjustments could promote moderate ingroup voices. By subtly demoting hateful and inciteful content and promoting moderate content, by adding 'friction' to inciteful posts, the promotion of hateful norms could be challenged. This has been shown in other contexts. For example, Twitter has found that adding a small step asking users if they want to share a news article that they haven't read reduces the propagation of misinformation (Gadde & Beykpour, 2020). A similar tactic may reduce the traction of hateful content and make moderate voices more easily heard.

The promotion of positive ingroup voices can be taken further, and evidence for the powerful effects of ingroup members challenging hateful norms comes from the literature on online counter speech - direct responses to hate speech which are intended to influence and challenge this behaviour (Benesch, 2014). On Twitter, counter speech can reduce instances of racist speech if instigators are morally sanctioned by a high-status in-group member (Munger, 2017, 2020). Similarly, counter speech messages priming a common identity and containing endorsements from elite actors are particularly effective in decreasing users' subsequent levels of hate speech (Siegel & Badaan, 2020). Additionally, counter speech comments which receive higher engagement from the wider online community are more likely to prompt the producers of hate speech to change their behaviour (Mathew, Kumar, Ravina, Goyal, & Mukherjee, 2018; Mathew et al., 2019), while threads with a higher number of unique counter speech contributors are also more likely to be successful (Procter et al., 2019). Together these studies support the offline evidence that ingroup members challenging the norms of acceptable behaviour can be effective in reducing discrimination. There is limited investigation into the prevalence of counter speech on social media, but at times it has been shown to outnumber hateful content. For example, following the 2015 Paris terror attacks the quantity of Tweets defending

Muslims was greater than the number of Islamophobic Tweets (Magdy, Darwish, Abokhodair, Rahimi, & Baldwin, 2016).

Future work should continue to explore what kinds of counter-speech might be most effective in diverse cultural contexts and on different platforms, as well as how counter-speech can be encouraged among everyday social media users. This is likely to be particular effective when coming from influential ingroup members, whilst negative effects may occur if this counter speech is done aggressively by outgroup members (Chapter 1).

Making hate harder to find

While promoting positive ingroup voices may help prevent the socialisation of hate, it is unlikely to deter those who are actively searching for hate online, and our results from Chapter 3 show that this is a substantial community. An alternative approach which builds in inertia to online systems, making it harder to discover hateful material, will therefore likely be required for these users. One such approach has been trialled in a collaboration between Google's Project Jigsaw and counter-extremism thinktank Moonshot CVE. This approach looks to prevent users searching for hateful material from arriving on social media platforms which host this content, and instead re-direct them towards more positive material. In the case of Islamic extremism, the methodology leverages Google's search engine and YouTube to place advertisements that undermine extremist narratives at the top of search results for those interested in joining ISIS (Google Jigsaw & Moonshot CVE, 2017). More recently, this has also been rolled out for far-right extremism, in collaboration with the Anti-Defamation League (Greer & Ramalingam, 2020), while Facebook has launched a similar 'Redirect Initiative' which redirects users who search for hate-related terms towards resources, education, and counter-extremism groups (Facebook, 2020). These groups include organisations founded by former extremists themselves, therefore looking to leverage the power of positive ingroup voices as discussed above.

Early results suggest that these approaches are successful at directing a reasonable quantity of clicks away from hateful material and towards more positive content (Helmus & Klein, 2018). Given that our results in both Chapter 3 and 4 show that users appear to be searching for hateful content online and that this plays a role in process of radicalisation, this is a positive step. Whether these approaches have a significant impact over the longer-term is less clear and deserves further research.

This concept of adding inertia or friction to online systems can also be applied to platforms which are already hosting hateful material. There is evidence that algorithmic filtering and social media recommendation systems can promote further hateful content to users who consume this material on social media platforms (see Literature Review). On YouTube for example, users can be drawn to progressively extreme content over time based on YouTube recommender algorithm (O'Callaghan, Greene, Conway, Carthy, & Cunningham, 2015; Ribeiro, Ottoni, West, Almeida, & Meira, 2019), while a high degree of YouTube activity has also been identified in the browsing habits of 'lone actor' terrorists (Basra, 2020). The addition of random noise within this recommender process, or friction where users need to actively click to view the next video, could help reduce this process.

Inoculating against hate

In addition to these more 'reactive' measures to counter online conflict, a number of more pro-active measures could also be taken, which look to protect or infer resistance to users prior to exposure to certain content. These approaches look to reduce the 'demand' side of hate speech consumption rather than the supply, whilst also making it harder to manipulate the online environment, whether that is by far-right trolls or agents for foreign states.

These 'inoculation' proposals have initially been made in relation to work on online misinformation. Pre-emptive interventions designed to induce scepticism and critical thinking prior to misinformation exposure have been shown to reduce users' susceptibility to believe this false information (Maertens, Roozenbeek, Basol, & van der Linden, 2020; Roozenbeek & van der Linden, 2019; Roozenbeek & Linden, 2018; Roozenbeek & van der Linden, 2020). The proposal is therefore that increasing awareness of the prevalence of false information and its potential impacts, including on polarisation and elicitation of emotional responses, can provide a degree of 'cognitive immunity', protecting against future exposure to this content. More recently, this approach has been introduced as a strategy for building resistance against the adoption of extremist beliefs and attitudes following exposure to hateful online material (Braddock, 2019). Exposure to an 'inoculation message' consisting of a warning before reading left- or right-wing extremist propaganda reduced both the intention of individuals to support the extremist group, as well as perceptions of the extremist group's credibility.

These ideas have the potential to reduce the impact of exposure to hate speech (Chapter 3) or the impact of a hostile interaction with an outgroup member (Chapter 1) or with an inauthentic social media account (Chapter 5). These strategies are still in their infancy however, and more research is

needed to determine in which contexts they are most effective, how long the potential 'immunity' effect lasts, and discover any potentially undesirable effects, such as the loss of trust in authority figures cause by increased general scepticism.

Conclusion

The challenge posed by the Internet in driving intergroup conflict and the threat of extremism, online and offline, continues to grow. The increased opportunity for global connectivity that the Internet, and in particular social media, provides, has led people from diverse backgrounds to become more connected, overcoming existing geographical, political, or social barriers. This connectivity has not led to the desired increases in social cohesion across all walks of life, however. Instead, the results discussed in this thesis show how social media has allowed for hostile intergroup contact between opposing groups, the extreme denigration of outgroups and minorities within ingroup discussions, and the manipulation and antagonism of controversial discussions by state-backed actors. Together, these dynamics may work to further polarise already divided societies, drive intergroup conflict and extremism, with the result being seen offline through increased hate crime and intergroup violence.

The processes discussed and studied in this work have unfortunately been clearly demonstrated in the course of writing up this thesis. Most recently, in October 2020, an Islamic extremist terror attack occurred in Paris (Willsher, 2020). Footage of the attack was shared widely across far-right social media channels, and a few days later two Muslim women were attacked under the Eiffel Tower in an alleged retaliatory Islamophobic attack (Aljazeera, 2020). While it may not be possible to say that social media 'caused' either of these attacks, it does appear to have driven the intergroup conflict which underlies the violence. Subsequently, a Jihadi-inspired terror attack occurred in Nice and an attempted far-right attacker was killed in Avignon, both on the same day (Salaun & Gaillard, 2020; Tidman, 2020). In this latter attack the attacker was found to be a member of an online far-right group and attempted to inspire further division by imitating a Muslim during the attack. This example demonstrates how necessary it is to combat all forms of terrorism, and the interplay between them – catalysed by the Internet.

Overall, the work in this thesis highlights how understanding the role of the Internet and social media in fostering and exacerbating intergroup conflict is more important than ever, but by understanding these processes we can also start mitigating their negative effects and promote more positive intergroup relations and a healthier online environment.

References

Abrams, D., de Vyver, J. Van, Houston, D. M., & Vasiljevic, M. (2017). Does terror defeat contact? Intergroup contact and prejudice toward muslims before and after the london bombings. *Peace and Conflict*, *23*(3), 260–268. https://doi.org/10.1037/pac0000167

Akrami, N., Ekehammar, B., & Bergh, R. (2011). Generalized prejudice: Common and specific components. *Psychological Science*, *22*(1), 57–59. https://doi.org/10.1177/0956797610390384

Aljazeera. (2020). Two women stabbed at Eiffel Tower in apparent racist attack. *Aljazeera*. Retrieved from https://www.aljazeera.com/news/2020/10/22/two-women-stabbed-under-eiffel-tower-in-apparent-racist-attack

Allan, H., Glazzard, A., Jesperson, S., Reddy-Tumu, S., & Winterbotham, E. (2015). *Drivers of violent extremism: Hypotheses and literature review. RUSI Serial Report*. Retrieved from https://assets.publishing.service.gov.uk/media/57a0899d40f0b64974000192/Drivers_of_Radicalisation_Literature_Review.pdf

Allport, G. (1954). *The nature of prejudice*. (K. Clark & T. Pettigrew, Eds.). Addison-Wesley Publishing Company. https://doi.org/10.1002/9780470773963

Alrababah, A., Marble, W., Mousa, S., & Siegel, A. (2019). Can Exposure to Celebrities Reduce Prejudice? The Effect of Mohamed Salah on Islamophobic Behaviors and Attitudes, (19). https://doi.org/10.31235/osf.io/eq8ca

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can we predict a riot? Disruptive event detection using twitter. *ACM Transactions on Internet Technology*, *17*(2). https://doi.org/10.1145/2996183

Asch, S. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35. https://doi.org/10.1038/scientificamerican1155-31

Awan, I., & Zempi, I. (2016). The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, *27*, 1–8. https://doi.org/10.1016/j.avb.2016.02.001

Baddiel, D. (2015). Short of a conspiracy theory? You can always blame the Jews. *The Guardian*. Retrieved from https://www.theguardian.com/commentisfree/2015/jul/22/conspiracy-theory-jews-david-cameron-antisemitism-extremism

Bail, C., Argyle, L., Brown, T., Bumpus, J., Chen, H., Hunzaker, M. B., … Volfovsky, A. (2018). Exposure to opposing views can increase political polarization: Evidence from a large-scale field experiment on social media. *Proceedings of the National Academy of Sciences*, 1–6. https://doi.org/10.17605/OSF.IO/4YGUX

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, *33*(6), 712–729. https://doi.org/10.1177/0894439314558836

Basra, R. (2020). *The YouTube browsing habits of a lone-actor terrorist*. Retrieved from https://gnet-research.org/2020/06/22/the-youtube-browsing-habits-of-a-lone-actor-terrorist/

BBC. (2016). Dallas shooting suspect Micah Johnson "acted alone." *BBC*. Retrieved from https://www.bbc.co.uk/news/world-us-canada-36752603

Benesch, S. (2014). Defining and diminishing hate speech. In *State of the World's Minorities and Indigenous Peoples 2014* (pp. 18–26). Minority Rights Group International.

Berger, J. M. (2018). *Extremism*. Cambridge, Massachusetts: The MIT Press.

Bergh, R., & Akrami, N. (2016). Generalized prejudice: Old wisdom and new perspectives. *The Cambridge Handbook of the Psychology of Prejudice*, (1946), 438–460. https://doi.org/10.1017/9781316161579.019

Bergh, R., Akrami, N., & Ekehammar, B. (2012). The personality underpinnings of explicit and implicit generalized prejudice. *Social Psychological and Personality Science*, *3*(5), 614–621. https://doi.org/10.1177/1948550611432937

Bevensee, E. (2020). *The decentralized web of hate*. Retrieved from https://rebelliousdata.com/wp-content/uploads/2020/10/P2P-Hate-Report.pdf

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993–997. https://doi.org/10.1037/0021-9010.79.6.993

Blascovich, J., Mendes, W. B., Hunter, S. B., Lickel, B., & Kowai-Bell, N. (2001). Perceiver threat in social

interactions with stigmatized others. *Journal of Personality and Social Psychology*, *80*(2), 253–267. https://doi.org/10.1037/0022-3514.80.2.253

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Braddock, K. (2019). Vaccinating against hate: Using attitudinal inoculation to confer resistance to persuasion by extremist propaganda. *Terrorism and Political Violence*, *12*(3), 1–23. https://doi.org/10.1080/09546553.2019.1693370

Bradshaw, S., & Howard, P. N. (2019). *The global disinformation order: 2019 global inventory of organised social media manipulation*. Retrieved from https://comprop.oii.ox.ac.uk/research/cybertroops2019/

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brown, R., & Hewstone, M. (2005). An integrative theory of intergroup contact. *Advances in Experimental Social Psychology*, *37*, 255–343. https://doi.org/10.1016/S0065-2601(05)37005-5

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., … Voss, A. (2014). Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, *4*(1), 1–14. https://doi.org/10.1007/s13278-014-0206-4

Cohen, K., Johansson, F., Kaati, L., & Mork, J. C. (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence*, *26*(1), 246–256. https://doi.org/10.1080/09546553.2014.849948

Collins, B. Ben, Zadrozny, B., & Saliba, E. (2020). White nationalist group posing as antifa called for violence on Twitter. *NBC News*. Retrieved from https://www.nbcnews.com/tech/security/twitter-takes-down-washington-protest-disinformation-bot-behavior-n1221456

Conover, M., Ratkiewicz, J., & Francisco, M. (2011). Political polarization on Twitter. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, *133*(26), 89–96. https://doi.org/10.1021/ja202932e

Conway, M., Khawaja, M., Lakhani, S., Reffin, J., Robertson, A., & Weir, D. (2019). Disrupting Daesh: Measuring takedown of online terrorist material and its impacts. *Studies in Conflict & Terrorism*, *42*(1–2), 141–160. Retrieved from https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513984

Coser, L. A. (1956). *The functions of social conflict*. Free Press. https://doi.org/10.4324/9780203714577

Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of Personality and Social Psychology*, *82*(3), 359–378. https://doi.org/10.1037/0022-3514.82.3.359

Davey, B. J. (2020). *Infiltration operations: How 4chan sought to compromise the Black Lives Matter protests*. Retrieved from https://www.isdglobal.org/digital_dispatches/infiltration-operations-how-4chan-sought-to-compromise-the-black-lives-matter-protests/

Del Vicario, M., Gaito, S., Quattrociocchi, W., Zignani, M., & Zollo, F. (2017). Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the Italian Referendum. *ArXiv*. Retrieved from http://arxiv.org/abs/1702.06016

Donovan, J., Lewis, B., & Freidberg, B. (2018). Parallel ports: Sociotechnical change from the alt-right to alt-tech. In *Post-Digital Cultures of the Far Right* (pp. 49–66). Transcript. https://doi.org/10.14361/9783839446706-004

Douek, E. (2020). *The rise of content cartels. Knight First Amendment Institute*. https://doi.org/10.2139/ssrn.3572309

Dovidio, J F, & Gaertner, S. L. (1991). Changes in the expression and assessment of racial prejudice. In *Opening doors: Perspectives on race relations in contemporary America*. The University of Alabama Press. Retrieved from https://psycnet.apa.org/record/1991-98067-007

Dovidio, John F., Gaertner, S. L., Anastasio, P. A., Bachman, B. A., & Rust, M. C. (1993). The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European Review of Social Psychology*, *4*(1), 1–26. https://doi.org/10.1080/14792779343000004

Durodie, B., & Ng, S. C. (2009). Is internet radicalization possible? *RSIS Commentaries*, *299*(5613), 1719–1722. https://doi.org/10.1063/1.2978249

Ebner, J. (2017). *The rage: The vicious circle of Islamist and far-right extremism*. London: I.B.Tauris.

Ekehammar, B., & Akrami, N. (2003). The relation between personality and prejudice: A variable- and a person-centred approach. *European Journal of Personality*, *17*(6), 449–464. https://doi.org/10.1002/per.494

Evans, R. (2018). How the MAGAbomber and the Synagogue Shooter Were Likely Radicalized.

Evans, R. (2019). *Shitposting, inspirational terrorism, and the Christchurch mosque massacre. Bellingcat*. Retrieved from https://www.bellingcat.com/news/rest-of-world/2019/03/15/shitposting-inspirational-terrorism-and-the-christchurch-mosque-massacre/

Evans, R. (2021). How the insurgent and MAGA right are being welded together on the streets of Washinton D.C. *Bellingcat*. Retrieved from https://www.bellingcat.com/news/americas/2021/01/05/how-the-insurgent-and-maga-right-are-being-welded-together-on-the-streets-of-washington-d-c/

Facebook. (2020). Facebook Counterspeech - Redirect Initiative. Retrieved from https://counterspeech.fb.com/en/initiatives/redirect/

Facebook Newsroom. (2017). *Global Internet Forum to Counter Terrorism to hold first meeting in San Francisco*. Retrieved from https://about.fb.com/news/2017/07/global-internet-forum-to-counter-terrorism-to-hold-first-meeting-in-san-francisco/

Ferrara, E. (2017). Desinformation an bots oparations on the run up to the 2017 French presidential election. *SSRN Electronic Journal*. https://doi.org/10.5210/FM.V22I8.8005

Gadde, V., & Beykpour, K. (2020). *An update on our work around the 2020 US Elections*. Retrieved from https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html

Gallacher, J. D., & Fredheim, R. E. (2019). Division abroad, cohesion at home: How the Russian troll factory works to divide societies overseas but spread pro-regime messages at home. In *Responding to Cognitive Security Challenges* (p. 60:79). Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Garimella, K. (2017). Quantifying and bursting the online filter bubble. *WSDM 2017 - Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, *60*(4), 837. https://doi.org/10.1145/3018661.3024933

Garrett, R. K., Weeks, B. E., & Neo, R. L. (2016). Driving a wedge between evidence and beliefs: How online ideological news exposure promotes political misperceptions. *Journal of Computer-Mediated Communication*, *21*(5), 331–348. https://doi.org/10.1111/jcc4.12164

Gaudette, T., Scrivens, R., & Venkatesh, V. (2020). The role of the Internet in facilitating violent extremism: Insights from former right-wing extremists. *Terrorism and Political Violence*, 1–18. https://doi.org/10.1080/09546553.2020.1784147

Gehlback, H., Robinson, C, D., & Vriesema, C, C. (2018). Climate conversations: Seeking a common starting point. *PsyArXiv*. Retrieved from doi 10.31234/osf.io/s8a7z

Gelava, S., & Buziashvili, E. (2020). *Georgian far-right and pro- government actors collaborate in inauthentic Facebook network*. Retrieved from https://medium.com/dfrlab/georgian-far-right-and-pro-government-actors-collaborate-in-inauthentic-facebook-network-730b9593a729

Google Jigsaw. (2020). The violent white supremacy issue. *The Current*, (002). Retrieved from https://jigsaw.google.com/the-current/white-supremacy/

Google Project Jigsaw, & Moonshot CVE. (2019). *The redirect method: A blueprint for bypassing extremism*. Retrieved from https://redirectmethod.org/downloads/RedirectMethod-FullMethod-PDF.pdf

Graf, S., Paolini, S., & Rubin, M. (2014). Negative intergroup contact is more influential, but positive intergroup contact is more common: Assessing contact prominence and contact prevalence in five Central European countries. *European Journal of Social Psychology*, *44*(6), 536–547. https://doi.org/10.1002/ejsp.2052

Greer, R., & Ramalingam, V. (2020). *The search for extremism: Deploying the redirect method*. Retrieved from https://www.washingtoninstitute.org/policy-analysis/view/the-search-for-extremism-deploying-the-redirect-method

Grover, T., & Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*, 193–204.

Helmus, T. C., & Klein, K. (2018). *Assessing outcomes of online campaigns countering violent extremism: A case study of the redirect method*. Retrieved from https://www.rand.org/pubs/research_reports/RR2813.html

Hermansson, P., Lawrence, D., Mulhall, J., & Murdoch, S. (2020). *The international alt-right: Fascism for the*

21st century? Routledge. https://doi.org/10.4324/9780429032486

Hosseinmardi, H., Ghasemian, A., Clauset, A., Rothschild, D. M., Mobius, M., & Watts, D. J. (2020). Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube. *ArXiv*. Retrieved from http://arxiv.org/abs/2011.12843

Hyers, L. L., & Swim, J. K. (1998). A comparison of the experiences of dominant and minority group members during an intergroup encounter. *Group Processes and Intergroup Relations*, *1*(2), 143–163. https://doi.org/10.1177/1368430298012003

Innes, M. (2017). *Russian influence and interference measures following the 2017 UK terrorist attacks. Cardiff University Crime and Security Research Institute*. Retrieved from https://crestresearch.ac.uk/resources/russian-influence-uk-terrorist-attacks/

J.M. Berger. (2019). The new strategy of violent white supremacy. *The Atlantic*, 1–12. Retrieved from https://www.theatlantic.com/ideas/archive/2019/08/the-new-strategy-of-violent-white-supremacy/595648/

Jeangene Vilmer, J. B. (2019). *The "Macron leaks" operation: A post-mortem*. Retrieved from https://www.atlanticcouncil.org/wp-content/uploads/2019/06/The_Macron_Leaks_Operation-A_Post-Mortem.pdf

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*(4), 360–386. https://doi.org/10.1016/0022-1031(92)90051-K

Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication*, *32*(3), 257–273. https://doi.org/10.1177/0267323117695734

Knigge, P. (1998). The ecological correlates of right-wing extremism in Western Europe. *European Journal of Political Research*, *34*(2), 249–279. https://doi.org/10.1111/1475-6765.00407

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, *35*(SUPPL.1), 69–93. https://doi.org/10.1111/pops.12163

Lavine, H. G., & Johnston, C. D. (2012). *The ambivalent partisan: How critical loyalty promotes democracy. The ambivalent partisan: How critical loyalty promotes democracy*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199772759.001.0001

Lewis, A., & Marwick, A. (2017). Taking the red pill: Ideological motivations for spreading online disinformation. In *Understanding and Addressing the Disinformation Ecosystem*. University of Pennsylvania Annenberg School for Communication. https://doi.org/DOI:

Littleford, L. N., Wright, M. O. D., & Sayoc-Parial, M. (2005). White students' intergroup anxiety during same-race and interracial interactions: A multimethod approach. *Basic and Applied Social Psychology*, *27*(1), 85–94. https://doi.org/10.1207/s15324834basp2701_9

MacInnis, C. C., & Page-Gould, E. (2015). How can intergroup interaction be bad if intergroup contact is good? Exploring and reconciling an apparent paradox in the science of intergroup relations. *Perspectives on Psychological Science*, *10*(3), 307–327. https://doi.org/10.1177/1745691614568482

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000315

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A., & Baldwin, T. (2016). #ISISisNotIslam or #DeportAllMuslims? Predicting unspoken views. *WebSci '16: Proceedings of the 8th ACM Conference on Web Science*, 95–106. https://doi.org/10.1145/2908131.2908150

Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American Journal of Political Science*, *59*(1), 128–145. https://doi.org/10.1111/ajps.12089

Mathew, B., Kumar, N., Ravina, Goyal, P., & Mukherjee, A. (2018). Analyzing the hate and counter speech accounts on Twitter. *ArXiv*. Retrieved from https://arxiv.org/pdf/1812.02712.pdf

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., … Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the 13th International Conference on Web and Social Media*, 369–380. Retrieved from https://arxiv.org/abs/1808.04409v1

McCauley, C., & Moskalenko, S. (2008). Mechanisms of political radicalization: Pathways toward terrorism. *Terrorism and Political Violence*, *20*(3), 415–433. https://doi.org/10.1080/09546550802073367

McSwiney, J., & Jasser, G. (2020). *Gab.com: The pro-Trump alternive social media. VoxPol*. Retrieved from https://www.voxpol.eu/gab-com-the-pro-trump-alternative-to-social-media/

Miller, N., & Brewer, M. B. (1984). The social psychology of desegregation. In *Groups in contact* (pp. 1–8). Academic Press. https://doi.org/10.1016/b978-0-12-497780-8.50007-3

Moghaddam, F. M. (2005). The staircase to terrorism a psychological exploration. *American Psychologist*, *60*(2), 161–169. https://doi.org/10.1037/0003-066X.60.2.161

Moghaddam, F. M. (2018). *Mutual radicalization: How groups and nations drive each other to extremes*. Washington: American Psychological Association. https://doi.org/10.1037/0000089-000

Mooijman, M., Hoover, J., Lin, Y., Ji, H., & Dehghani, M. (2018). Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 1. https://doi.org/10.1038/s41562-018-0353-0

Müller, K., & Schwarz, C. (2020a). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, *00*(0), 1–37. https://doi.org/10.1093/jeea/jvaa045

Müller, K., & Schwarz, C. (2020b). From hashtag to hate crime: Twitter and anti-minority sentiment. *SSRN Electronic Journal*, 1–47. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3149103

Munger, K. (2017). Tweetment effects on the Tweeted: Experimentally reducing racist harassment. *Political Behavior*. https://doi.org/10.1007/s11109-016-9373-5

Munger, K. (2020). Don't @ Me: Experimentally Reducing Partisan Incivility on Twitter. *Journal of Experimental Political Science*, 1–15. https://doi.org/10.1017/XPS.2020.14

Nagle, A. (2017). *Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and The Alt-Right*. Zero Books.

NBC. (2016). Sniper ambush kills 5 officers, injures 7 in Dallas following peaceful protest. *NBC News*. Retrieved from https://www.nbcdfw.com/news/local/protests-in-dallas-over-alton-sterling-death/88950/

Nimmo, B., Buziashvili, E., Sheldon, M., Karan, K., Aleksejeva, N., Bandeira, L., … Hibrawi, R. (2019). *Top takes: Suspected Russian intelligence operation*. Retrieved from https://medium.com/dfrlab/top-takes-suspected-russian-intelligence-operation-39212367d2f0

Nimmo, B., Francois, C. C., Eib, S., Ronzaud, L., Ferreira, R., Hernon, C., & Kostelancik, T. (2020). *Secondary infektion*. Retrieved from https://secondaryinfektion.org

O'Callaghan, D., Greene, D., Conway, M., Carthy, J., & Cunningham, P. (2015). Down the (white) rabbit hole: The extreme right and online recommender systems. *Social Science Computer Review*, *33*(4), 459–478. https://doi.org/10.1177/0894439314555329

Obaidi, M., Thompson, L., & Bergh, R. (2019). "They think we are a threat to their culture": Meta-cultural threat fuels willingness and endorsement of extremist violence against the cultural outgroup. *International Journal of Conflict and Violence (IJCV)*, *12*. https://doi.org/10.4119/UNIBI/ijcv.647

Obama, B. H. (2017). President Obama's farewell address. Retrieved from https://obamawhitehouse.archives.gov/farewell

Osadchuk, R. (2020). *Sockpuppet accounts on fringe websites spread pro-separatist narratives in Luhansk*. Retrieved from https://medium.com/dfrlab/sockpuppet-accounts-on-fringe-websites-spread-pro-separatist-narratives-in-luhansk-6d4505428f72

Ozalp, S., Williams, M. L., Burnap, P., Liu, H., & Mostafa, M. (2020). Antisemitism on Twitter: Collective efficacy and the role of community organisations in challenging online hate speech. *Social Media and Society*, *6*(2). https://doi.org/10.1177/2056305120916850

Page-Gould, E., Mendoza-Denton, R., & Tropp, L. R. (2008). With a little help from my cross-group friend: Reducing anxiety in intergroup contexts through cross-group friendship. *Journal of Personality and Social Psychology*, *95*(5), 1080–1094. https://doi.org/10.1037/0022-3514.95.5.1080

Paolini, S., Hewstone, M., Voci, A., Harwood, J., & Cairns, E. (2006). Intergroup contact and the promotion of intergroup harmony: The influence of intergroup emotions. In *Social Identities* (pp. 209–238). Routledge. https://doi.org/10.4324/9780203002971-11

Pariser, E. (2011). *The Filtter Bubble: What the internet is hiding from you*. New York, New York, USA: The Penguin Press.

Park, G., & DeShon, R. P. (2018). Effects of group-discussion integrative complexity on intergroup relations in a social dilemma. *Organizational Behavior and Human Decision Processes*, *146*(March), 62–75.

https://doi.org/10.1016/j.obhdp.2018.04.001

Pauwels, L., & Schils, N. (2016). Differential online exposure to extremist content and political Violence: Testing the relative strength of social learning and competing perspectives. *Terrorism and Political Violence*, *28*(1), 1–29. https://doi.org/10.1080/09546553.2013.876414

Pearl, J. (2018). *The book of why: The new science of cause and effect*. Allen Lane.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology*, *49*, 65–85. https://doi.org/10.1146/annurev.psych.49.1.65

Polonski, V. (2018). *#MacronLeaks changed political campaigning: Why Macron succeeded and Clinton failed* (World Economic Forum Blog). Retrieved from https://medium.com/world-economic-forum/macronleaks-changed-political-campaigning-why-macron-succeeded-and-clinton-failed-7d66cc77749e

Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond binary labels: Political ideology prediction of Twitter users. *Proceedings Ofthe 55th Annual Meeting Ofthe Association for Computational Linguistics*, 729–740. https://doi.org/10.18653/v1/p17-1068

Procter, R., Webb, H., Jirotka, M., Burnap, P., Housley, W., Edwards, A., & Williams, M. (2019). A study of cyber hate on twitter with implications for social media governance strategies. *ArXiv*. https://doi.org/10.36370/tto.2019.20

Quattrociocchi, W., Scala, A., & Sunstein, C. R. (2016). Echo chambers on Facebook. *ArXiv:1411.2893*, *51*(2015), 1–12. https://doi.org/10.1145/2740908.2745939

Reicher, S. D., & Haslam, S. A. (2016). Fueling Extremes. *Scientific American Mind*, *27*(3), 34–39. https://doi.org/10.1038/scientificamericanmind0516-34

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass*, *2*(3), 1313–1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., … Zannettou, S. (2020). From pick-up artists to incels: A data-driven sketch of the manosphere. *ArXiv*. Retrieved from http://arxiv.org/abs/2001.07600

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., & West, R. (2020). Does platform migration compromise content moderation? Evidence from r/The_Donald and r/Incels. *ArXiv*. Retrieved from http://arxiv.org/abs/2010.10397

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2019). Auditing radicalization pathways on YouTube. Retrieved from http://arxiv.org/abs/1908.08313

Roozenbeek, J, & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, 1–10. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, Jon, & Linden, S. Van Der. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *9877*, 1–11. https://doi.org/10.1080/13669877.2018.1443491

Roozenbeek, Jon, & van der Linden, S. (2020). Breaking Harmony Square: A game that "inoculates" against political misinformation. *Harvard Kennedy School Misinformation Review*, *1*(8), 1–26. https://doi.org/10.37016/mr-2020-47

Salaun, T., & Gaillard, E. (2020). Tunisian man beheads woman, kills two more people in Nice church. *Reuters*. Retrieved from https://uk.reuters.com/article/uk-france-security-nice/tunisian-man-beheads-woman-kills-two-more-in-nice-church-idUKKBN27E177

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings Ofthe Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/w17-1101

Schwab, A. K., Sagioglou, C., & Greitemeyer, T. (2019). Getting connected: Intergroup contact on Facebook. *Journal of Social Psychology*, *159*(3), 344–348. https://doi.org/10.1080/00224545.2018.1489367

Selinger, E., & Hartzog, W. (2016). Facebook's emotional contagion study and the ethical problem of co-opted identity in mediated environments where users lack control. *Research Ethics*, *12*(1), 35–43. https://doi.org/10.1177/1747016115579531

Sellars, A. F. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, *20*. https://doi.org/10.1093/jicj/mqaa023

Settle, J. E. (2018). *Frenemies*. Cambridge University Press. https://doi.org/10.1017/9781108560573

Shelton, J. N. (2003). Interpersonal concerns in social encounters between majority and minority group members. *Group Processes & Intergroup Relations*, *6*(2), 171–185. https://doi.org/10.1177/1368430203006002003

Shelton, J. N., Richeson, J. A., & Vorauer, J. D. (2006). Threatened identities and interethnic interactions. *European Review of Social Psychology*, *17*(1), 321–358. https://doi.org/10.1080/10463280601095240

Sherif, M. (1948). *An outline of social psychology*. New York: Harper & Brothers.

Siegel, A. A., & Badaan, V. (2020). #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, *114*(3), 837–855. https://doi.org/10.1017/S0003055420000283

Simmel, G. (1972). *On individuality and social forms. University of Chicago Press*. University of Chicago Press.

Smith, L. G. E., Blackwood, L., & Thomas, E. F. (2019). The need to refocus on the group as the site of radicalization. *Perspectives on Psychological Science*, *15*(2), 327–352. https://doi.org/10.1177/1745691619885870

Suedfeld, P., & Bluck, S. (1988). Changes in Integrative Complexity prior to surprise attacks. *Journal of Conflict Resolution*, *32*(4), 626–635. https://doi.org/10.1177/0022002788032004002

Sunstein, C. R. (2017). *#Republic: divided democracy in the age of social media*. Princeton, New Jersey, United States: Princeton University Press.

Tetlock, P. E., Peterson, R. S., & Berry, J. M. (1993). Flattering and unflattering personality portraits of integratively simple and complex managers. *Journal of Personality and Social Psychology*, *64*(3), 500–511. https://doi.org/10.1037/0022-3514.64.3.500

The Bundestag. Act to Improve Enforcement of the Law in Social Networks, Network Enforcement Act § (2017). Retrieved from https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf;jsessionid=7ED3746864613FBE240CCAD5BA61DF86.2_cid334?__blob=publicationFile&v=2

Tidman, Z. (2020). Man shot dead in Avignon after threatening public with weapon. *The Independent*. Retrieved from https://www.independent.co.uk/news/world/europe/france-terror-avignon-attack-stabbing-police-shooting-nice-today-b1420399.html

Twitter Safety. (2019). Information operations directed at Hong Kong. Retrieved from https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html

Tworek, H., & Leerssen, P. (2019). *An analysis of Germany's NetzDG law. Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression*. Retrieved from https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf

Urman, A., & Katz, S. (2020). What they do in the shadows: examining the far-right networks on Telegram. *Information Communication and Society*, *0*(0), 1–20. https://doi.org/10.1080/1369118X.2020.1803946

Vidgen, B., Yasseri, T., & Margetts, H. (2019). Trajectories of Islamophobic hate amongst far right actors on Twitter. *ArXiv*, 1–20. Retrieved from https://arxiv.org/pdf/1910.05794

Wendling, M. (2018). *Alt-Right: From 4Chan to the White House* . Pluto Press.

Williams, M. L., & Burnap, P. (2016). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, *56*(2), 211–238. https://doi.org/10.1093/bjc/azv059

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2019). Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, *60*(1), 1–25. https://doi.org/10.1093/bjc/azz049

Willsher, K. (2020). Teacher decapitated in Paris named as Samuel Paty, 47. *The Guardian*. Retrieved from https://www.theguardian.com/world/2020/oct/17/teacher-decapitated-in-paris-named-as-samuel-paty-47