Towards Analyzing Online Communities of Problematic Information: A Computational Approach

Shruti Phadke

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Tanushree Mitra, Chair

Emma Spiro

Kate Starbird

Program Authorized to Offer Degree:

Information Science

University of Washington

Abstract

Towards Analyzing Online Communities of Problematic Information: A Computational Approach

Shruti Phadke

Chair of the Supervisory Committee:

Tanushree Mitra

Information Science

Problematic information - information that is inaccurate, misleading, inappropriately attributed, or altogether fabricated - prevails in digital societies and spaces. Communities formed around sharing, theorizing, or mobilizing problematic information can lead online users through the path of social distrust, paranoia, or radicalization. Not only that but ideas propagating through such communities spreading conspiracy theories or hateful ideologies can tear at the social fabric and threaten civility in online spaces.

Despite the obvious disruptive consequences of online communities of problematic information, we don't have a large-scale, data-driven understanding of deeper social processes that are underway. In this dissertation, I contribute empirical insights into engagement, mobilization, and disengagement from communities of problematic information using theory-guided quantitative methods. My analysis of problematic online communities takes me across multiple social media platforms such as Reddit, Facebook, Twitter, and 4chan and various methodologies ranging from machine learning, and natural language processing to qualitative interviews. Equipped with these theories and methods, I analyze various instances of problematic information, such as conspiracy theories and hate movements in the West, and

coordinated political amplification in India. Specifically, I investigate what makes people engage in conspiracy theory discussions, what are the mechanisms of information mobilization and content framing in online hate movements and political campaigns, and what are the ways in which people may leave online conspiracy theory discussions.

From a data science perspective, my work uncovers how social media users engage with problematic online content and leverage technologies across platforms. From a social-psychological perspective, my research contributes to the literature on how various social, psychological, and cognitive processes motivate the users' journey through problematic information online.

I showcase that studying large-scale digital traces contributes to the discussion on how thousands of online users self-select themselves into conspiracy theory discussion communities, how they take various pathways of engagement inside online conspiracy theory discussions, and how early signs of fracture in conspiracy worldviews result in eventual disengagement from conspiracy theory discussion communities. Moreover, looking into thousands of Facebook groups discussing white supremacy, and anti-LGBTQ ideas, this research also reveals how problematic information is mobilized through accounts playing various social roles within the hate movement.

In my ongoing work, I leverage the multidisciplinary understanding from my doctoral research to design online prompts for intervening in online interactions surrounding climate change denial conspiracy theories. More broadly, I am excited about continuing this work by exploring ethical, fair, and effective ways of intervening in open online discussions to reduce problematic content online.

Not unnaturally, many elevators imbued with intelligence and precognition became terribly frustrated with the mindless business of going up and down, up and down, experimented briefly with the notion of going sideways, as a sort of existential protest, demanded participation in the decision-making process and finally took to squatting in basements sulking.

— *The Hitchhiker's Guide to the Galaxy, Douglas Adams*

Dedication

To my grandparents

Acknowledgments

My time as a doctoral student has been transformative, both personally and professionally. The last few years have shown me what it means to really love your work and how to creatively channel scientific curiosity. Completing this thesis has been a significant milestone in my academic career, and I would like to express my heartfelt appreciation to all those who have contributed to its realization.

First and foremost, I am most thankful to my mentors and senior colleagues who have had a profound impact on shaping me as a researcher. I am indebted to my Ph.D. advisor, Dr. Tanu Mitra, for her unwavering support and encouragement throughout the ups and downs of my Ph.D. In addition to developing concrete research practices and ethics, the research management and writing skills she has imparted will stay with me for the rest of my professional life. My research philosophy and rigor have been significantly influenced by Dr. Mattia Samory who helped me tap into my curiosity for understanding people through a series of exceptional and inspirational collaborations. I am deeply thankful for Dr. James Pennebaker very generously took me on many thought adventures, teaching me how to think about research beyond the constraints of methods. Simply put, these researchers continue to enrich me as a person, and I feel truly grateful for their mentorship.

Having continued my Ph.D. from a Masters's degree, I faced significantly fewer academic burdens as a young Ph.D. student. My early Ph.D. years gave me the time and space to work with wonderful colleagues and mentors at Virginia Tech such as Dr. James Hawdon and Jonathan Lloyd who introduced to me collaborative research. I am also thankful for my new UW community, especially my committee members Dr. Emma Spiro, Dr. Kate Starbird, and Dr. David Ribes who showed genuine curiosity about my work and supported me in my journey through Ph.D. milestones. I also want to thank Dr. Gary Hsieh for generously stepping in as a GSR and being patient with the scheduling conflicts.

In the early years of Ph.D., I was blessed to have a vibrant group of friends who gave me a sense of community. Late work nights as a struggling Ph.D. student would

not have been half as productive or exciting without the company of Momen Bhuiyan, Vartan Kesiz Abnousi, Alexander Rodriguez, and Mohannad Elhamod. Especially in the early phases of paper rejections and self-doubt, Momen and Vartan kept me laughing and made me feel like a part of a team. I am blessed to have an equally supportive and wonderful set of colleagues at UW. I am thankful to have Prerna Juneja as a close friend and a labmate who was a constant companion throughout my Ph.D. and made Seattle a bit sunnier for me. I have genuinely benefited by working with Kristen Engel who taught me valuable lessons about being a responsible researcher and validated my guilty pleasures in entertainment. I am also grateful to get a chance to work with Neelesh Agrawal and Saloni Dash who bring invaluable excitement to my current research.

Deciding to pursue a Ph.D. has been one of the best decisions I have made so far. Along with shaping my professional identity, my doctoral journey has given me valuable, and hopefully lifelong, connections with these truly extraordinary people.

List of Publications

Related to the Thesis :

1. **Shruti Phadke**, et al. "Framing hate with hate frames: Designing the codebook." Companion of the 2018 ACM conference on computer-supported cooperative work and social computing. 2018.

2. **Shruti Phadke**, and Tanushree Mitra. "Many faced hate: A cross-platform study of content framing and information sharing by online hate groups." Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.

3. **Shruti Phadke**, Mattia Samory, and Tanushree Mitra. "What makes people join conspiracy communities? Role of social factors in conspiracy engagement." Proceedings of the ACM on Human-Computer Interaction 4.CSCW3 (2021): 1-30.

4. **Shruti Phadke**, Mattia Samory, and Tanushree Mitra. "Characterizing social imaginaries and self-disclosures of dissonance in online conspiracy discussion communities." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-35.

5. **Shruti Phadke**, and Tanushree Mitra. "Educators, Solicitors, Flamers, Motivators, Sympathizers: Characterizing Roles in Online Extremist Movements." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-35.

6. **Shruti Phadke**, Mattia Samory, and Tanushree Mitra. "Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions." Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2022.

7. **Shruti Phadke**, and Tanushree Mitra. "Characterizing Political Campaigning with Lexical Mutants on Indian Social Media". (In submission)

8. Kristen Engel, **Shruti Phadke**, and Tanushree Mitra. "Learning from the Ex-Believers]Learning from the Ex-Believers: Individuals' Journeys In and Out of Conspiracy Theories Online". Accepted March 2023 at CSCW 2023.

OTHERS :

4. **Phadke, Shruti**, et al. "Addressing Challenges and Opportunities in Online Extremism Research: An Interdisciplinary Perspective." Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing. 2021.

Table of Contents

List of Publications **7**

1 Introduction **13**
1.1 Definitions . . . 15
1.1.1 Problematic Information . . . 15
1.1.2 Conspiracy Theories . . . 16
1.1.3 Hate and Domestic Extremism Movements . . . 16
1.1.4 Coordinated Influence Operations . . . 17
1.2 Research Arcs . . . 17
1.2.1 Engagement in communities of problematic information . . . 18
1.2.2 Practices and Mobilization in Communities of Problematic Information . . . 19
1.2.3 Disengagement from Communities of Problematic Information 20
1.3 Contributions in the Completed Work . . . 21
1.4 Notes on Research Ethics . . . 22

2 Engagement in online communities of problematic information **25**
2.1 Background . . . 26
2.1.1 Individual Factors in Conspiracy Theorizing . . . 26
2.1.2 Social Aspects of Conspiracy Theory Adoption . . . 26
2.1.3 Sunstein's Framework of Social Factors in Conspiracy Theorizing 27
2.2 Research Questions . . . 28
2.3 RQ: Understanding importance of social factors in conspiracy theory engagement . . . 28
2.3.1 Finding the Conspiracy Theory Discussion Communities on Reddit 28
2.3.2 Data and Subjects . . . 35
2.3.3 Factors in Conspiratorial Engagement . . . 41
2.3.4 Understanding the Importance of Features . . . 48

2.4 Discussion and Implications 51
2.5 Limitations 55
2.6 Conclusions 55

3 Communication Practices in Communities of Problematic Information 59
3.1 Background 60
3.1.1 Hate groups as social movement organizations 60
3.1.2 Hate movements and content framing on social media 61
3.1.3 Framing theory 62
3.1.4 Hate movements and information mobilization 64
3.1.5 Participatory activism in information mobilization 65
3.1.6 Success of social movements through information mobilization 66
3.1.7 Information Mobilization outside of West: Political Amplification in India 69
3.2 Research Questions 70
3.3 RQ1: Characterizing cross-platform content framing by hate groups . . 73
3.3.1 Building a Collective Action Framework 73
3.3.2 Data Collection and Annotation 75
3.3.3 Result: Cross-Platform Content Framing 76
3.3.4 Characterizing Cross-Platform Information Sharing by Hate Groups 79
3.3.5 Information sharing practices by hate groups with different ideologies 80
3.4 RQ2: Information mobilization in hate movements through social roles 83
3.4.1 Collecting data for extremist accounts 83
3.4.2 Identifying Roles in Online Extremist Movements 88
3.4.3 Clustering Extremist Accounts Based on the Derived Features . 92
3.4.4 Results: Roles in Online Extremist Movements 94
3.4.5 Measuring Role Dynamics 96
3.4.6 Role Dynamics in Extremist Information Sharing 97
3.5 RQ3: Coordinated Political Amplification in India 99
3.5.1 Collecting cross-platform data 99
3.5.2 Identifying Political Campaigns with Lexical Mutations 101
3.5.3 Characterizing hidden amplification campaigns 105
3.6 Discussion and Implications 110

3.6.1 Political Amplification in India . . . 116
3.7 Limitations . . . 118
3.8 Conclusions . . . 119

4 Disengagement from Communities of Problematic Information 121
4.1 Background . . . 122
4.1.1 Research exploring disengagement from problematic content . . . 122
4.1.2 QAnon conspiracy theory . . . 122
4.1.3 Conspiracists and Social Imaginaries . . . 124
4.1.4 Conspiracy Theorizing and Cognitive Dissonance . . . 124
4.1.5 Cognitive dissonance and QAnon . . . 125
4.2 Research Questions . . . 126
4.3 RQ1: Contrasting pathways of engagement inside conspiracy theory discussions . . . 127
4.3.1 Data and Subjects . . . 127
4.3.2 Characterizing engagement pathways in Online conspiracy theory discussions . . . 128
4.3.3 Models and measures for contrasting users on various conspiracy theory engagement trajectories . . . 131
4.3.4 Results: Contrasting conspiracy engagement pathways . . . 136
4.4 RQ2: QAnon disengagement and cognitive dissonance . . . 138
4.4.1 Datasets . . . 138
4.4.2 Characterizing QAnon social imaginaries: The QAnon Canon . . . 139
4.4.3 Five Dimensions of QAnon Social Imaginaries . . . 140
4.4.4 Compiling Factors in Self-Disclosure of Belief and Dissonance . . . 142
4.4.5 Creating a dissonance classifier . . . 143
4.4.6 Results: Dissonance classifier . . . 148
4.4.7 User Engagement after Dissonance Self-Disclosure . . . 151
4.4.8 Results: Changes in User Contributions after Dissonance . . . 152
4.5 Discussion and Implications . . . 154
4.6 Limitations . . . 157
4.7 Conclusions . . . 158

5 Conclusions and Future Directions 159
5.1 Future Directions . . . 159

5.1.1 Multidisciplinary research foundation for countering problematic information online 159
5.1.2 Designing online prompts to reduce engagement in problematic information 160
5.2 Conclusion 161

Bibliography **163**

CHAPTER 1

INTRODUCTION

> *I was like 12 when I found 4chan and Reddit. It engaged me a lot as a kid and young teen because of the sort of edgy weird anti-establishment community culture and the very outlandish shock humor it produced...it was much more organic and more effective at making me accept extreme stuff and conspiracies because I was literally becoming a part of it*

Simon, an interview participant in our research, explained what attracted him most to pizza gate conspiracy theory[1] as a teenager was the edgy sense of humor and community culture of Reddit and 4chan's conspiracy theory discussions. Karen remembers her initial journey down the QAnon [2] rabbit hole, as one driven by a sense of solidarity with other QAnoners on Facebook. Marc, however, mentioned that after publicly letting go of his Illuminati[3] beliefs on Facebook, he was met with extreme hostility and harassment from his former friends that deeply impacted his mental

[1] a falsified rumor that Hillary Clinton and other Democrats were heading up a child sex-trafficking ring out of a Washington pizzeria. https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory

[2] A theory born out of Pizzagate conspiracy theory purporting that a cabal of Satanic, cannibalistic sexual abusers of children operating a global child sex trafficking ring conspired against former U.S. President Donald Trump during his term in office. https://en.wikipedia.org/wiki/QAnon

[3] A conspiracy theory that claims that a secret cabal of powerful members controls the happenings of the world https://www.crf-usa.org/bill-of-rights-in-action/bria-11-4-c-conspiracy-theories-attacks-on-jefferson-set-the-pattern.html

health. Going even further, Jed describes his tenure in Reddit's alt-right subreddits as a toxic, abusive relationship, filled with paranoia and anger. May it be socially divisive conspiracy theories, hateful messages or influential political propaganda, the public sphere of social media is inundated with problematic information (defined in Section 1.1) compromising the civility and mental health in online spaces. Problematic information can not only induce panic and distrust but can also facilitate disruptive offline mobilization of online participants, as is evidenced by recent riots at U.S Capitol in Washington DC [102], JFK resurrectionist cult in Dallas [108] or growing vaccine hesitancy across the globe [253].

Online social media platforms such as Reddit or Facebook allow people to make connections all around the world and find peers and communities with similar interests. Along with the unprecedented networking affordances, social media also allows free and almost unregulated exchanges of ideas. While such exchanges can contribute towards positive social change and remove social, cultural and geographical barriers between people, similar affordances can also be leveraged to spread problematic information. Online communities formed around theorizing, sharing, or mobilizing problematic information can serve as a functional infrastructure for recruitment, growth, and even financing offline movements such as LGBTQ rights opposition or QAnon.

What makes people seek out or share problematic information? Are people simply predisposed towards engaging with problematic information due to their sociocultural markup or are there other factors involved? Why are messages from white supremacy or anti-immigration groups so appealing to a certain demographic? Who or what popularizes problematic information online, bringing it to the mass? And perhaps more importantly, how can we discourage people from engaging with problematic information online? Can we learn from those who have walked the walk?

To find out answers to these questions, one naturally turns to the exhaustive research in social psychology or criminology describing peoples' journey into fringe movements. However, many foundational works describing people and communities involved in fringe social movements are based on either integrating historic cultural perspectives or ethnographic and clinical observations of people with physical association with the movement. Take for example, theories on social movement participation [139, 140] that were devised from interviewing participants at the physical protests, or the cognitive dissonance theory [88] that resulted from immersive ethnographic exploration of a UFO religion and clinical experiments. In contrast, by creating virtual social

spaces, online communities have restructured the ways in which people participate in causes, consume information, and socialize [39].

The fact that social media has become a structural phenomenon in societies worldwide, calls for the application of existing theories in a way that scales up to digital societies made up of millions of users. My research aims to achieve just that, merging the theories with the strength of social media data. In this research, using the theories from social psychology as a backbone, along with the quantitative methods such as natural language processing and machine learning, I explore how users join communities of problematic information, how participants in such communities frame events and mobilize information and how users disengage from problematic information. In the rest of the introduction, I will define and contextualize concepts relevant to this research, briefly outline each of the research arcs and discuss ethical considerations.

1.1 Definitions

My dissertation is centered around studying various instances of problematic information such as conspiracy theories or extremist narratives. What is problematic information? Before diving deeper into the research details, I feel it is necessary to state my definitions of these terms and the contexts in which I have used them in my research.

1.1.1 Problematic Information

I borrow the phrase "problematic information" from Dr. Caroline Jack, who in her work *Lexicon of Lies* [125], describes it as:

> information that is inaccurate, misleading, inappropriately attributed, or altogether fabricated.

Problematic information can mean misinformation, disinformation, propaganda, gaslighting, hoaxes, conspiracy theories, or hate speech[73, 125]. One reason for using this umbrella term is that conspiracy theory or racial extremism narratives often have overlapping components of misinformation, disinformation, or propaganda. For example, climate change denial conspiracy theories use both, misinformation and political propaganda [252].

While using this term I acknowledge that the interpretation of problematic information can be impacted by my political biases and the biases of the research community I

affiliate with. That being said, while studying individual cases of problematic information such as conspiracy theories or extremist propaganda, I have tried to rely on definitions proposed by independent researchers, or organizations such as Southern Poverty Law Center (SPLC).

1.1.2 Conspiracy Theories

Chapter 2 and 4 investigate engagement in and disengagement from conspiracy theory discussion groups online. While historically there has been great variability in interpretations of "conspiracy theory" as a concept [39], scholars consent on basic constructs of a conspiracy theory [217]. Various threads of research describe conspiracy theories as attempts to explain the occurrence of an event as a covert plot orchestrated by secret organizations [15]. It is important to note that conspiracy theory has two components—an insinuation of a "conspiracy" and an element of "theorizing". The conspiracy is theorized to be perpetrated by powerful and influential individuals or a group of individuals [54, 200] with hidden, sinister, or nefarious goals [1, 65, 259]. The theorizing, whether done independently by individuals or constructed socially in groups, provides a conceptual framework for making sense of the alleged conspiracy [39].

This work assumes a generally agreed-upon definition of conspiracy theory that purports that conspiracy theories are attempts to explain events as a covert plot by powerful organizations [15]. Also, I acknowledge that some conspiracy theories may be true or based on genuine evidence. Not all conspiracy theories have socially harmful consequences. Moreover, this research does not support stigmatizing conspiracy theory beliefs or believers.

1.1.3 Hate and Domestic Extremism Movements

In Chapter 3, this research refers to domestic hate and extremism in the context of the United States. Specifically, I refer to Southern Poverty Law Center (SPLC)'s definition that describes hate movements as:

> an organization or collection of individuals that—based on its official statements or principles, the statements of its leaders, or its activities—has beliefs or practices that attack or malign an entire class of people, typically for their immutable characteristics. An organization does not need to have engaged in criminal conduct

> or have followed their speech with actual unlawful action to be labeled a hate group[4]

This criterion is borrowed from the FBI's definition of hate crime: criminal offense against a person or property motivated in whole or in part by an offender's bias against a race, religion, disability, sexual orientation, ethnicity, gender, or gender identity[5].

Accordingly, this research primarily studies hate organizations as designated by SPLC, affiliated with white supremacy (race), anti-LGBTQ (sexual orientation, gender identity), anti-Muslim (religion), and anti-Immigration (ethnicity) movements.

1.1.4 Coordinated Influence Operations

Chapter 3 discusses lexical mutants in coordinated political amplification operations on Indian social media. Various social media platforms define coordinated influence operations as

> coordinated efforts to manipulate or corrupt public debate for strategic goal — Facebook[6]

> attempts to manipulate Twitter to influence elections and other civic conversations by foreign or domestic state-linked entities—Twitter[7]

Coordinated influence operations can leverage problematic information to influence political and civic discourse [125]. Researchers argue that in influence operations, information, especially disinformation spreads by flowing strategically or organically through various accounts on social media [237]. We adhere to the definitions provided above and study coordinated operations in India that are designed towards amplifying political messages through lexical mutations.

1.2 Research Arcs

My exploration of online communities of problematic information is centered around three primary themes: (1) engagement in communities of problematic information,

[4] https://www.splcenter.org/20200318/frequently-asked-questions-about-hate-groups#hate%20group

[5] https://www.fbi.gov/investigate/civil-rights/hate-crimes

[6] https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf

[7] https://transparency.twitter.com/en/reports/information-operations.html

	RQs	Methods	Theories	Publications
Engagement (chapter 2)	What makes people engage in conspiracy theory discussions online?	Quantitative	Sunstein and Vermule Social Factors Framework	CSCW 2020
Practices and Mobilization (chapter 3)	How hate groups share and frame information across platforms?	Mixed	Framing theory	CHI 2020
	How do online users mobilize hateful information?	Mixed	Resource Mobilization	CSCW 2021
	Coordinated political amplification on Indian social media	Quantitative		submitted
Disengagement (chpater 4)	How are online pathways of disengagement from conspiracy theories different than of increasing engagement?	Quantitative	RECRO Model for Radicalization	ICWSM 2022
	What makes people disengage from online conspiracy theory discussions?	Quantitative	Cognitive Dissonance	CSCW 2021

Figure 1.1: Figure outlining the works included in dissertation

(2) communication and information-sharing practices in communities of problematic information, and, (3) disengagement from problematic information. In the next subsections, I briefly describe various research arcs which are also illustrated in Figure 1.1.

1.2.1 Engagement in communities of problematic information

In **Chapter 2**, I investigate the importance of various individual and social factors in joining online communities of problematic information by studying Reddit's conspiracy theory discussion communities [197]. A predisposition towards paranoid thinking [65], anxiety and insecurity [1] and suspicion towards authoritative information sources [255, 264] have been popularly identified as precursors for joining conspiracy theory discussion communities. Yet, these studies investigate individuals' attitudes isolated from their social environment. Despite the social nature of such movements and the collective action resulting from them, we have surprisingly little insight into the role of social factors in joining harmful online movements. Specifically, I ask:

1. What makes people engage in conspiracy theory discussions online?

In my CSCW 2020 paper [197], I contrasted the importance of social interactions with individual psychological factors in conspiracy discussions. Specifically, this research explored how interactions with conspiracists, in comparison with individual psychological traits such as expressed anxiety, sadness, paranoia etc., lead to users' joining online conspiracy communities. Along with identifying important social factors, this work helped in uncovering the prominent role of negative content moderation in joining conspiracy theory discussion communities. For example, users who receive heavy downvotes and comment removals from mainstream Reddit communities, tend to join conspiracy theory communities where their thoughts and beliefs are accepted by the other community members and moderators. Overall, this study has implications in content moderation suggesting that excessive and mindless censorship can drive people towards conspiracy theory discussions. This framework of social and individual factors could be extended to study other instances of problematic communities such as racial supremacy groups or anti-LGBTQ movements.

1.2.2 Practices and Mobilization in Communities of Problematic Information

What happens after users join communities of problematic information? How do users share information or frame events? How do such communities mobilize influential narratives?

In **Chapter 3** I explore various communication and information-sharing practices in communities of problematic information centered around constructing narratives, information, and shared meanings. In this chapter, I focus on domestic hate, such as white supremacy, anti-LGBTQ, anti-immigration and anti-Muslim movements, in the USA. I further extend my research into analyzing information mobilization outside of the West and investigate political amplification in India.

Specifically, I present two studies set up across Facebook and Twitter investigating

1. How do hate groups share and frame information across platforms? (Facebook, Twitter)

2. How do online users mobilize hateful information? (Facebook groups)

3. What are the characteristics of political amplification in on Indian social media (Facebook and Twitter)?

This research arc primarily revealed that problematic narratives of racial supremacy or anti-LGBTQ movement are framed differently across different platforms [195] and circulated through a series of accounts playing various social roles in the community [196]. For example, Facebook groups are used to propagate radicalization narratives to like-minded followers and Twitter is used to mass educate diverse audiences [195]. Moreover, the accounts through which such narratives are propagated assume various social roles such as *solicitors*—who solicit participation and funds for the extremist movement, *educators*—accounts that share intellectual content about extremism and prominently share and like extremist content or *flamers*—accounts that express and incite anger by posting inflammatory content [196].

Both of the above-mentioned studies are focused on the mobilization of problematic information in the West. In fact, many prior works on online adversarial information mobilization primarily focus on the Western context, neglecting other growing democracies in Asia or the Global South. For example, despite having the highest number of users on WhatsApp[8], Instagram[9] and Facebook[10] and third largest user-base on Twitter[11], along with the recorded instances of online adversarial political and religious influence [126, 168], there is little research on how problematic information gets circulated in Indian subcontinent. Hence, in **Chapter 3**, I further analyze coordinated religious and political influence operations on Indian social media. Specifically, I look at political amplification through lexical mutations and reveal how going beyond simple copy-pasta can enable political groups to mobilize information in India.

1.2.3 Disengagement from Communities of Problematic Information

While the first two arcs focus on understanding factors in joining communities of problematic information, and characterizing practices and mobilization, this part of my research aims to understand what makes people leave communities of problematic information. Specifically in **Chapter 4**, I ask two questions focusing on users' exit from online conspiracy theory discussion communities:

1. How are online pathways of disengagement from conspiracy theories different than of increasing engagement?

[8] https://backlinko.com/whatsapp-users
[9] https://datareportal.com/essential-instagram-stats
[10] https://datareportal.com/essential-facebook-stats
[11] https://datareportal.com/essential-twitter-stats

2. What makes people disengage from online conspiracy theory discussions?

In my work at ICWSM 2022 [199], I investigate online trajectories of users who participate in Reddit's conspiracy theory discussions. Specifically, I compare users with increasing engagement in conspiracy discussions with those who decreased their participation over time, using various linguistic, interaction, and engagement markers. I find that users who disengage from online conspiracy theory discussions limit their participation to specialized conspiracy topics, participate in diverse discussion groups, and show reduced language conformity with conspiracy theory subreddits.

After comparing users on various conspiracy theory engagement trajectories, I further evaluate possible mechanisms that promote disengagement from conspiracy theories. In my CSCW paper [198], I explore cognitive dissonance as one possible mechanism by which users may depart from their conspiracy theory beliefs and imaginaries. Using the theory-driven markers of cognitive dissonance, I analyze users' exit from Reddit's QAnon discussion communities and find that users leave QAnon because of various fracture points in their conspiracy social imaginaries.

1.3 Contributions in the Completed Work

Here, I want to briefly highlight some of the key contributions in my doctoral research.

- **Conspiracy theory social factors framework:** The social factors framework described in **Chapter 2**, offers a systematic operationalization of theoretically-motivated individual and social factors towards conspiracy theory engagement. This framework can be exported to platforms other than Reddit that are designed around open discussions.

- **Conspiracy Scale:** In **Chapter 2**, using a data driven approach, we construct the "conspiracy scale" to identify subreddits discussing conspiracy theories. The conspiracy scale allows the characterization of subreddits according to their similarity to `r/conspiracy` and the diversity of user contributions across different subreddits.

- **Hate Frames Codebook:** In **Chapter 3** we develop a framing theory-based annotation framework to empirically measure hate group-specific framing strategies. Using this codebook, we revealed how hate groups adopt different narrative

strategies across different platforms. I hope that scholars would find our framework useful to extend work in studying influence strategies by hate groups and other extremist movements.

- **Social movement participation framework:** In **Chapter 3**, we offer a framework for systematic operationalization of theoretically motivated characteristics of social movement participation which can be used to study social movements in digital spaces.
- **QAnon Canon Lexicon:** In **Chapter 4**, we offer the QAnon Canon[12], a lexicon of over 403 phrases capturing the symbolic language and its shared meanings across QAnon social imaginaries that can serve as a toolbox for researchers to extend our study to other platforms.
- **Subreddit generality scale:** In **Chapter 4**, we offer a *generality scale*, used in characterizing conspiracy theory worldviews, that captures the generalist or specialist nature of subreddits. This scale is not specific to conspiracy theory subreddits but can also be applied to the generality or specificity of subreddits across other topics.

1.4 Notes on Research Ethics

Before I conclude this introduction, I want to briefly mention the ethical considerations taken in this work. I refer to the AAAI code of conduct and ethics guidelines[13] that mention stakeholders, harm, privacy, and confidentiality dimensions of ethical research and conduct. First, I acknowledge that all people, especially social media users, and social computing researchers are stakeholders in this research. With this work, I intend to contribute insights that can be considered while building safer online spaces for all. Furthermore, given that this study is retrospective and involves no interaction with the studied population, I do not anticipate any direct harm resulting from this research. I take proactive steps to preserve user privacy. Specifically, by presenting results aggregated over thousands of users and by intentionally not reporting any exact quotes made by social media users, I reduce the risk of re-identification. Finally, throughout this thesis, I analyze non-confidential data that is available in the public domain and collected through publicly accessible APIs. Yet, given the potential stigma

[12] https://social-comp.github.io/ConspiracyTraces/

[13] https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php

associated with participating in problematic discussions, I do not release any raw user data from this study.

CHAPTER 2

ENGAGEMENT IN ONLINE COMMUNITIES OF PROBLEMATIC INFORMATION

Published Works

Phadke, Shruti, Mattia Samory, and Tanushree Mitra. "What makes people join conspiracy communities? role of social factors in conspiracy engagement." Proceedings of the ACM on Human-Computer Interaction 4.CSCW3 (2021): 1-30. https://dl.acm.org/doi/abs/10.1145/3432922 [197]

Most people do not partake in online communities with an intention to consume or share problematic information [193]. What factors are then involved in one's engagement with problematic information and communities? Specifically, how does the "social" part of the social network affects users' involvement in problematic information? In this chapter, I outline my research published in CSCW 2020 [197] answering these questions by analyzing Reddit's conspiracy theory discussion communities. Specifically, I ask: What makes people join conspiracy theory discussions online?

Be it in vaccine skeptics or in climate change denialists, participation in conspiracy theory discussions fuels collective action and has widespread consequences for society as a whole. Once joined, conspiracy theory discussion participants may radicalize, increasingly engaging with conspiracy theories and neglecting other communities [216]. It is thus crucial to understand what makes people participate in conspiracy theory discussions. This chapter reviews existing literature describing participation in conspiracy theory discussions and presents my research on conspiracy theory participation in online communities and the role of social factors.

2.1 Background

Existing literature offers various perspectives on why people might participate in conspiracy theory discussions. This section reviews the existing literature and presents psychological and epistemological models of conspiracy theorizing.

2.1.1 Individual Factors in Conspiracy Theorizing

2.1.1.1 Psychological Predisposition and Conspiracy Theorizing

Research on conspiracy theory adoption have largely focused on individual's psychological and epistemological characteristics [243]. For example, feelings of hopelessness, insecurity, anxiety, and lack of trust are considered important towards forming conspiratorial beliefs [38, 109]. Moreover, individuals that engage in conspiratorial beliefs are reported to show characteristics of paranoia [118], suspicion towards authoritative information sources [243] and tendency to believe unsubstantiated or false claims [169]. Previous studies stress that the need for justifying or explaining events forms the very foundation for conspiratorial thinking [134, 162, 258, 266]. For example, one study [258] found that people feel the need to detect patterns or "connect the dots" in order to make sense of the physical and social environment they live in [258]. This may explain the core process in developing irrational beliefs where people attempt to detect patterns for random events.

2.1.1.2 Epistemology and Conspiracy Theorizing

Sunstein et. al. [243] suggest that it is thus important to understand how people acquire information related to conspiracies [243]. Specifically, absence of relevant and ample information can result in "crippled epistemologies." In other words, people who are exposed to very limited relevant information and if what they know is wrong, they have a high likelihood of fixating on their inaccurate beliefs [243].

2.1.2 Social Aspects of Conspiracy Theory Adoption

In social sense, the development of conspiracy theories can be described by groups of individuals jointly constructing the understandings of the world on the basis of shared identity [93]. From this socio-constructionist stance, conspiracy theories are born from the social processes of filtering available information and deliberating on whether it is true. For example, conspiracy theories prosper in the wake of dramatic

events [216, 236] when available information is insufficient to assess its truthfulness. Thus in such situations conspiracy theorizing is an attempt of collective sensemaking [142]. Studies focusing on the collective processes of consuming information in the context of fake news present crucial insight on the collective pitfalls that may lead to formulating (false) conspiracy theories [142, 150]. In this work, we explicitly abstain from assessing the truth of conspiracy theories. Our focus, instead, is on the social factors that lead users to conspiracy theory discussions in the first place. Specifically, we use the theoretical framework of social factors proposed by Sunstein [243].

2.1.3 Sunstein's Framework of Social Factors in Conspiracy Theorizing

Besides psychological and epistemological causes, [243] turn to the sociology of conspiracy theorizing, and suggest a categorization of sociological factors in conspiracy theory adoption. Besides crippled epistemology, conspiracy theories may spread according to the same diffusion mechanisms as *rumors and speculations*: on the one hand "conspiracy entrepreneurs" like authors of conspiracy books may propagate a theory out of personal interest. On the other hand, culture and morality may determine whether an individual accepts a conspiracy theory, based on whether the theory is compatible with previously held beliefs. Yet, in the absence of prior beliefs, especially when ample and incontrovertible evidence about an event is missing, people need to rely on others' information and judgement to form their opinion. Thus, conspiracies may follow *information cascade*. Similarly to information cascades, conspiracy belief may also spread through *reputation cascades*—sometimes people profess belief in a conspiracy theory, or at least suppress their doubts, in order to maintain the good opinion of others. Moreover, conspiracy theories may spread as a symbol to justify public concern, whether or not that concern is warranted. A particular event becomes available, and conspiracy theories are invoked both in explaining it and using it as a symbol for broader social forces, casting doubt on accepted wisdom in many domains. Regardless of the nature of the cascade, *emotions* serve as a catalyst of conspiracy theory spread; when rumors trigger intense feelings, they are far more likely to be circulated. In addition to emotional effects, *group polarization* also plays a role in the adoption of conspiracy theories where members of groups show strong group identity.

In this research, we rely on the above categorization of social factors in conspiracy theory adoption to extract social signals from user contributions prior to their joining

conspiracy theory discussion communities.

2.2 Research Questions

Guided by the background and theories explained above, I now outline research questions undertaken in this chapter. With each research question, I will explain the gap in the literature that motivates the question.

1. **RQ: How important are social factors, in comparison with individual factors, towards users' joining of conspiracy theory discussion communities online?** What drives users to join online conspiracy communities? Users who do so, show early on a distinctive use of language and choice of special-interest communities [141]. This is in line with ample research in social psychology on the individual factors associated with conspiratorial belief [38, 109, 118]. Yet, these studies investigate individuals' attitudes isolated from their social environment. Despite the social nature of conspiracy theorizing online [93, 238] and of the collective action it projects onto the real world [143, 169], we have surprisingly little insight about the role of social factors in joining online conspiracy communities. This research question addresses this gap by first, operationalizing various social factors in conspiracy theory engagement and then contrasting the importance of social factors with that of individual factors.

2.3 RQ: Understanding importance of social factors in conspiracy theory engagement

To understand the importance of various factors towards users' joining of conspiracy theory discussions, we first need to identify various online communities that discuss conspiracy theories. I start by explaining our methods to identify conspiracy theory discussion communities on Reddit.

2.3.1 Finding the Conspiracy Theory Discussion Communities on Reddit

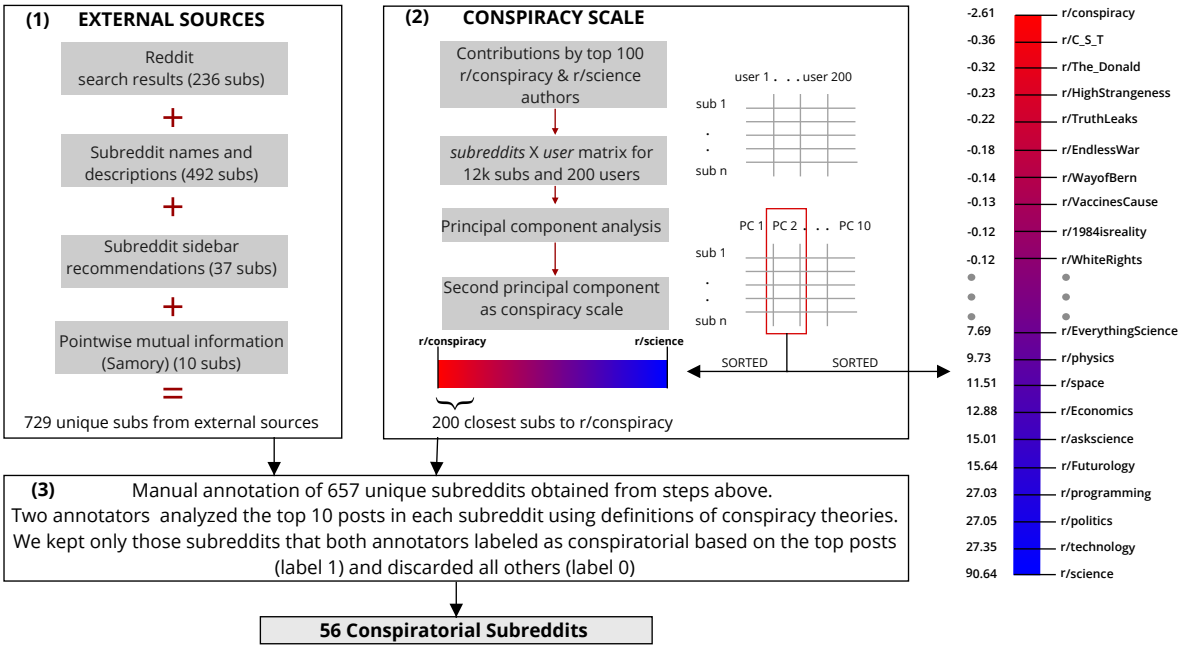
We take a socio-constructionist stance, and consider the collective of users producing conspiracy discussions as a community producing knowledge. Specifically on Reddit,

we first define a group of subreddits hosting conspiratorial discussions as the "conspiracy communities". Identifying subreddits that engage in conspiracy discussions is a challenging process for several reasons. Reddit has a total of 1.2 million subreddits with no global taxonomy that could help us easily understand the themes in different subreddits. Previous researchers have commonly focused only on `r/conspiracy`–a subreddit dedicated to discussing all types of conspiracy theories—to study conspiratorial engagement and narratives [141, 216, 217]. However, there are several other communities on Reddit that promote conspiracy theories as well. In that, some subreddits openly self-identify as conspiracy discussion communities while others host conspiratorial content without having it as their primary focus. For example, `r/ConspiracyII` invites only conspiracy theories whereas `r/ConspiracyNews` focuses on reporting news around conspiratorial topics. Moreover, even within the solely conspiratorial communities, some specialize on just one or few related conspiracy theory narratives and others welcome all types of discussions. To elaborate, there are specialized subreddits dedicated to discussing specific conspiracies, such as moon landing hoax and flat earth (`r/moonhoax` and `r/theworldisflat`, respectively) while others welcome any and all kinds of conspiratorial discussions (`r/FringeTheory` and `r/ConspiracyZone`). Given such high diversity in conspiracy discussion subreddits, it is imperative that we identify conspiracy communities with high precision. Towards this end, we employ a multi-stage, mixed-methods approach to first, mine potentially conspiratorial subreddits and then, carefully vet them using human judgement. Specifically, to find the candidates for conspiracy communities, we resort to two key steps. First we look at external sources such as Reddit recommendations and methods based on previous research. Second, we devise a conspiracy scale that weighs subreddits based on their similarity to `r/conspiracy`. Figure 2.1 displays the entire process of identifying the conspiracy communities.

2.3.1.1 External Sources:

We look at four external sources for finding conspiratorial communities. Specifically, Reddit search results, subreddit names and descriptions, subreddit sidebar recommendations and mutual information based methods used in previous research. Table 2.2 provides examples for subreddits mined from each of the four external sources alongwith the conspiracy scale described later.

1. **Reddit search results:** We look at search results provided by Reddit to emulate



(a) (b)

Figure 2.1: (a) Flowchart illustrating the process for identifying the conspiracy communities. We obtain the subreddit candidates for conspiracy communities using both, (1) external sources and (2) conspiracy scale. The conspiracy scale is generated by sorted second principal component of subreddit × user matrix of top `r/conspiracy` and `r/science` user contributions across different subreddits. We take leftmost 200 (200 subreddits closest to `r/conspiracy`) from conspiracy scale along with subreddits from external sources as candidates for conspiracy communities. We manually annotate the candidate subreddits to include 56 subreddits in conspiracy communities (b) Top 10 subreddits on both sides of the conspiracy scale alongwith their weights according to the 2nd principal component. We did not normalize the weights to preserve sparsity in values of the principal component.

how Reddit users might find conspiratorial subreddits. Reddit has a search bar at the top of any page in which users can enter the search query to find different subreddits, users and posts. We search the term 'conspiracy' on Reddit's home page and note 236 recommended subreddits.

2. **Subreddit name and description:** We consider the user's choice of knowingly participating in conspiracy discussions as an important criterion towards identifying Redditors that engage in conspiracies. Reddit users can understand the theme of a subreddit by subreddit names or the descriptions. Hence, we refer to the subreddit names and descriptions available on `files.pushshift.io/reddit` [1].

[1] `https://files.pushshift.io/reddit/` is a publicly accessible repository of Reddit datasets

Since we are interested in selecting self-identifying conspiracy subreddits, we perform regular expression match for the string "conspir" in the descriptions and the names.

3. **Subreddit sidebar recommendations:** Often, subreddit descriptions contain a sidebar in which the other related subreddits are listed. For example, `r/conspiracy` lists `r/Wikileaks` and `r/Endlesswar` as "related subreddits" it the sidebar (Table 2.2). Hence we looked at sidebar recommendations for subreddits obtained in step 2. We continued this process recursively until there were no more sidebar recommendations or the recommended subreddits were already listed in step 1-3. This process resulted in 37 new subreddits.

4. **Pointwise mutual information:** We also look at the work by other researchers characterizing conspiratorial communities. Specifically, Samory et. al. [216] find communities that share surprising number of common users to `r/conspiracy` (page 5, Table 1 in [216]). We consider top 10 subreddits listed in [216] as potentially conspiratorial subreddits.

We look at multiple sources for identifying conspiratorial subreddits to select conspiracy communities with high precision. While external sources produce useful candidates for the conspiracy communities, they are also limited in their effectiveness. Reddit recommendations produce high number of irrelevant suggestions. For example, we find several subreddits unrelated to conspiracy even within the top 10 search results on Reddit (`r/todayilearned` Table 2.2). In addition, subreddit sidebars are not always populated by the subreddit community. Moreover, pointwise mutual information approach extracts subreddits that are distinctively similar to `r/conspiracy` favouring smaller subreddits. For example, all of the top 10 subreddits closer to `r/conspiracy` listed in [216] have less number of subscribers and contribution volume (Table 2.2). Hence, we look for a data-driven, scalable approach that could capture subreddits that are generally, and not just surprisingly, similar to `r/conspiracy`. Specifically, we perform Principal Component Analysis (PCA) on contributions received in subreddits by different users to devise the "conspiracy scale" (Figure 2.1 a and b).

2.3.1.2 Conspiracy scale

We design the conspiracy scale to characterize semantic similarity between subreddits in terms of shared user base [158]. Previously, other researchers have also employed

maintained by Jason Baumgartner [20]

user participation based measures to compare subreddits [147, 158, 268]. However, such representations are not designed to specifically study the conspiratorial nature of the subreddits. Samory et. al. [216] use pointwise mutual information (PMI) to identify communities that are distinctively similar to `r/conspiracy`. While their approach successfully identifies conspiratorial subreddits, it focuses on identifying communities that share *surprisingly* common users with `r/conspiracy`. To elaborate, PMI is a co-occurrence based measure where mutual information between two subreddits is calculated based upon the number of common users in them. PMI can return *surprisingly* similar subreddits to `r/conspiracy` because it is known to be biased towards low frequency or rare items (or subreddits in this case) [31, 133]. To bridge this gap, we search for the method that would not be biased towards just surprise while measuring similarity between the subreddits. We take intuition from Samory et. al. [216] specifically that conspiratorial subreddits can be identified by contrasting the user activity in `r/conspiracy` to its polar opposite community—`r/science` [26]. However, instead of focusing on finding subreddits that are distinctively similar to `r/conspiracy`, we try to understand the similarity based on the variance in user-subreddit participation for users in `r/conspiracy` and `r/science` based on Principal Component Analysis (PCA).

2.3.1.3 Creating conspiracy scale

Figure 2.1 illustrates the process of devising the conspiracy scale. Previous scholars have shown that people who believe in one conspiracy tend to believe in others as well [26, 169, 236]. Accordingly, we presume that top contributing users—users with highest number of contributions—from `r/conspiracy` will have propensity to engage in other conspiracy related subreddits. Further juxtaposing their activity with top contributing `r/science` users can help enhance the contrast between conspiratorial and scientific subreddits. Hence, after removing bot accounts, we select the 100 top contributing users from each of the two subreddits. We extract the entire contribution timelines of these users across all subreddits using `pushshift.io` [20]. In all, this starting dataset spans over 12k subreddits. One could understand the variance in types of the 12k subreddits by analyzing the number of contributions made by the users in each of those subreddits. For example, just by sorting the raw counts of contributions within each subreddit, one could distinguish subreddits with larger subscriber counts from the smaller ones. For the task at hand, we want to extract the directionality in subreddits that places them from most similar to `r/science` to most

			Our conspiracy scale	PMI [216]	Subreddit embeddings [147]
Conspiracy ranks for conspiracy related subreddits	method		conspiracy scale sorted from r/conspiracy to r/science	ranked by closeness to r/conspiracy by PMI	ranked by cosine distance to r/conspiracy vector
	results	*rms*	**3332**	4271	17,741
		std	**4447**	4939	11,567
Science ranks for science related subreddits	method		conspiracy scale sorted from r/science to r/conspiracy	ranked by closeness to r/science by PMI	ranked by cosine distance to r/science vector
	results	*rms*	**3570**	4385	14,786
		std	**5816**	7102	21,448

Table 2.1: We validate the conspiracy scale generated from PCA (Figure 2.1 (a)(2) and (b)) by comparing the ranks generated for conspiracy and science related subreddits by our conspiracy scale and other approaches ([147, 216]). This table describes the methods used for generating ranks and the results (root mean square and standard deviation) for the ranks obtained. Our method has lowest rms and std in both, conspiracy and science rankings indicating that out scale places conspiracy related subreddits closer to r/conspiracy and science related subreddits closer to r/science.

similar to `r/conspiracy`. Principal Component Analysis (PCA) is a dimentionality reduction technique that could reduce the data along principal components that explain the maximal amount of variance. Intuitively, the first few components should give us different viewpoints to understand the variance in types of subreddits. Hence, we construct a $subreddit \times user$ matrix with values indicating contributions made by a user (column of the matrix) in a subreddit (row of the matrix) and apply PCA on it. Specifically, we extract first 10 components ranked based on the amount of variance they explain. Our underlying assumption here is that `r/conspiracy` users engage with more conspiratorial subreddits while `r/science` users engage with non conspiratorial subreddits. Hence we look for the principal component that projects subreddits in a way that places conspiratorial subreddits on one end and non-conspiratorial subreddits (subreddits similar to `r/science`) on the other end, resulting in maximal variance. The first component arranged the subreddits from smallest to largest—summarizing the general variety between subreddits. However, when sorted by *second* component of the PCA (Figure 2.1(a)(2)), `r/science` and `r/conspiracy` fall on two extreme ends indicating that the second component explains the second order variance that the first component does not capture. Since the second component identifies two poles in subreddits—`r/science` and `r/conspiracy`, we use it as the conspiracy scale. We consider top 200 subreddits from the conspiracy scale as candidates for the conspiracy communities (Figure 2.1 (b)).

External Sources					Conspiracy Scale
Reddit search results	**Subreddit names & descriptions**	**Subreddit Sidebar Recommendations**		**PMI [216]**	
r/conspiracy	r/thoseconspiracyguys	r/conspiracy	r/Wikileaks	r/CHEMPRINTS	r/conspiracy
r/insanepeoplefacebook	r/conspiratard		r/EndlessWar	r/bilderberg	r/C_S_T
r/WTF	r/muaconspiracy		r/PostCollapse	r/conspiracyhub	r/The_Donald
r/911truth	r/conspiracyundone		r/Documentaries	r/greenlight2	r/HighStrangeness
r/PanicHistory	r/ConspiracyMemes	r/conspiracytheories	r/skeptic	r/WhiteNationalism	r/TruthLeaks
r/TrueReddit	r/conspiracies		r/spacex	r/greenlight	r/EndlessWar
r/WikiLeaks	r/ConspiracyII		r/fakenews	r/HealthConspiracy	r/WayofBern
r/isconspiracyracist	r/actualconspiracies		r/bigfoot	r/OccupyLangley	r/VaccineCause
r/todayilearned	r/pokemonconspiracies	r/conspiracyII	r/OccultConspiracy	r/mysterybabylon	r/1984isreality
r/politics	r/OccultConspiracy		r/TheTranslucentSociety	r/moonhoax	r/WhiteRights

Table 2.2: Table listing the example subreddits obtained by each method. In all, we obtained a total of 657 unique subreddits from every method.

2.3.1.4 Validating conspiracy scale

How well does our scale place conspiracy related subreddit on conspiracy end and science related subreddits on the other? How does it compare with other subreddit similarity measures? We compare our conspiracy scale with two external subreddit similarity measures—pointwise mutual information by Samory et. al. [216] and community embeddings by Kumar et. al. [147]. First, we generate the list of (i) conspiracy related and (ii) science related subreddits based on the subreddit names and descriptions. Specifically, we search for the substring "conspir" and "sci" in subreddit names and terms "conspiracy". "conspiracies", "science" and "scientific" in subreddit descriptions. We curate this list to keep only relevant subreddits in both (i) and (ii). Next, we rank the subreddits using the three methods as described in Table 2.1. In all three methods, `r/conspiracy` and `r/science` have rank 1 in conspiracy and science ranks respectively. Thus it follows that the conspiracy ranks for conspiracy related subreddits should be close to one. Similarly, the science ranks for science related subreddits should be close to one. Hence, to compare the aggregate ranking of subreddits across all three methods, we calculate root mean square and standard deviation of the ranks generated. In both, conspiracy and science, our scale produces lower standard deviation and root mean square (rms) in the rankings (See Table 2.1) indicating that out scale places conspiracy related subreddits closer to `r/conspiracy` and science related subreddits closer to `r/science`. For examples, see top 10 subreddits on both sides of the conspiracy scale 2.1 (b). Moreover, unlike the similarity generated by the pointwise mutual information, our scale is not biased towards smaller or larger subreddits. For example, top 10 subreddits closes to `r/conspiracy` on the conspiracy scale (Figure 2.1 (c)) contain both, smaller and larger subreddits with respect to the subscriber count and the contribution volume.

2.3.1.5 Annotating conspiracy communities

Table 2.2 provides examples of subreddits obtained from every method discussed above. With the subreddit list obtained from the external sources and the conspiracy scale, we have 657 candidates for the conspiracy communities. For each candidate, we obtained annotations from two separate annotators who had sufficient experience and context for distinguishing conspiratorial and non-conspiratorial discussions. First, the annotators read the subreddit names and their descriptions. Then, they read at least top 10 submissions from each of the subreddits and analyzed them using the definitions of conspiracy theories aggregated in [217]. For example, one of the definitions states: "*...(conspiracies) involve multiple actors working together in secret to achieve hidden goals that are perceived to be unlawful or malevolent...*" [1, 65, 258, 283]. The annotators annotated the subreddit as 1 if either of the definitions applied in at least five posts and 0 otherwise (Figure 2.1(a)(3)). We discarded all subreddits that either of the annotators found to be irrelevant or anti-conspiracy or were about trolling conspiracists. For example, `r/ChickenApocalypse` contains jokes mentioning the conspiracies about chicken controlling the world and `r/Disinfo` is a watchdog subreddit for cataloguing misinformation and debunking conspiracy theories. After manual validation, we obtained a list of 56 subreddits that both annotators considered to host conspiracy discussions, ensuring high precision. We provide a list of subreddits in the conspiracy communities alongwith the links to the example posts containing conspiracies in the supplementary material.

2.3.2 Data and Subjects

Informed by collective action theories, we hypothesize that the participants in the conspiracy communities—**Current Conspiracists, CC**—exert some influence on people outside of the communities, via discussions and other interactions. In this light, our research question thus investigates if and how the influence of CC leads users to join the conspiracy communities. We measure CC's influence over users who will, at some point, join any one of the conspiracy communities (**Future Conspiracists, FC**). We compare and contrast CC's influence on FC against a matched cohort of control users. Using statistical matching, we find users that are comparable to FC in all respects but who never join any of the conspiracy communities—**Non Conspiracists, NC** (Figure 2.2). Below, we detail the source of our discussion data and our process for selecting our user cohorts FC, NC, and CC, and how we match FC and NC.

2.3.2.1 Discussions on Reddit

We study conspiracists on Reddit, a social media platform where users can create, share, and discuss content by participating in specific subdivisions of Reddit (or subreddits). Subreddits contain discussions around specific themes. For example, `r/Kanye` is for discussing anything related to Kanye West and `r/nintendo` is a subreddit for Nintendo news and games. Discussions in subreddits start with an opening post called submission, that sets the theme. Users can comment on the submissions and on other users' comments. For the sake of simplicity, we collectively call submissions and comments as "**contributions**", and all contributions in a discussion as "thread" (see Figure 2.2 (c)).

2.3.2.2 Finding Future Conspiracists (FC)

Figure 2.2 (b) outlines different time spans that we use throughout this work to characterize users' lifetime on Reddit. FC are Reddit users who eventually engage with any one of the conspiracy communities. We consider the time of their first contribution to any of the conspiracy subreddits as the time when a FC **joins** the community. We consider the 6 months preceding their joining as the **observation period** in which we study the individual and social factors affecting their joining. A total of 740,093 users ever contributed to the conspiracy communities; however, we impose a number of constraints to obtain a high quality sample of FC. We want to study FC who become actively engaged, and not users who post only incidentally such as spammers and trolls. Therefore, for each subreddit in the conspiracy communities, we calculate median number of contributions made by users in that community in their lifetime. We consider users that contribute more than the previously calculated median in any of the conspiracy communities as treatment candidates. To eliminate throwaway accounts we also remove users with less than 2 years of Reddit lifetime. To reliably measure signals of social factors during the observation period, we keep only users who have enough data—5 contributions—in that time. Finally, in order to reliably match FC and NC, we limit to users with at least 5 contributions and 6 months of activity prior to the observation period. Our final set of FC consists of 30,325 users.

2.3.2.3 Finding Non Conspiracists (NC)

To understand the prominence of individual and social factors towards conspiratorial engagement in FC, we need to compare such factors in normal Reddit users, the control

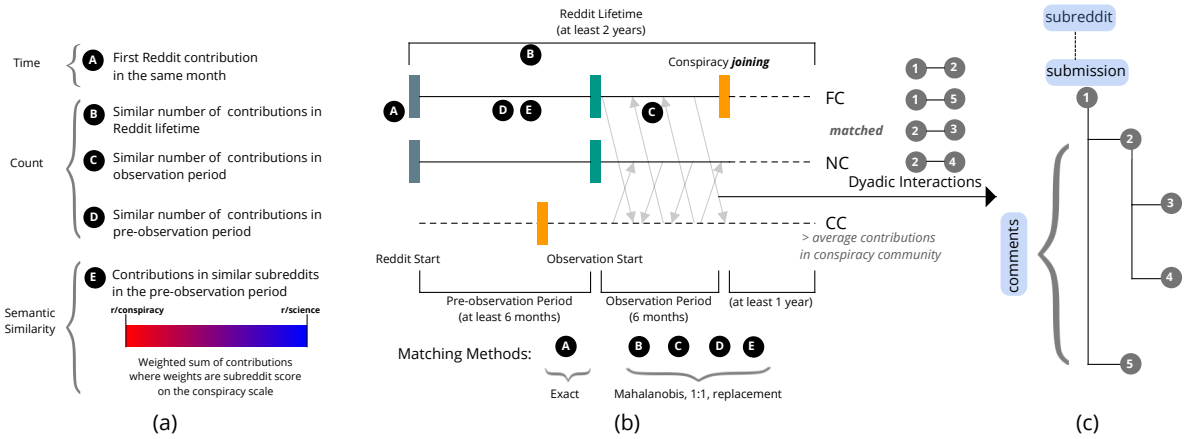


Figure 2.2: (a) Matching criteria used for finding similar FC and NC (b) Different time spans in the users' Reddit life. FC and NC are matched on Ⓐ exactly and on Ⓑ, Ⓒ, Ⓓ, Ⓔ criteria using nearest neighbor matching with Mahalanobis distance. ▌symbolizes Reddit start, ▌observation period start and ▌signifies FC's and CC's joining any one of the conspiracy communities. FC and NC have at least 2 years of Reddit life, at least 10 contributions in the observation and 10 in pre-observation period. We calculate the users dyadic interactions with CC in the observation period. (c) Elaboration of dyadic interactions (faded gray arrows in (a)) indicating replies to or from CC) with examples of dyadic interactions in a typical Reddit discussion thread.

group of NC users. Ideally, we want the FC and NC to be indistinguishable based on their Reddit contributions and tenure, but for the fact that NC never join any of the conspiracy communities. We begin with a list of 10 million Reddit users who have at least 2 years of Reddit lifetime and no contributions in the conspiracy communities. Next, we refine this list to match the group of FC users based on the following criteria.

- **A:** Reddit start month. To select users with similar Reddit tenure, we first match FC with all NC candidates that made first contribution in Reddit in the same month as FC.

Next, we want users that are similarly active on Reddit. We define different time spans over FC's life and find NC that have similar contributions in those time periods. Specifically, we match on:

- **B:** Contribution volume in the Reddit lifetime
- **C:** Contribution volume in the observation period
- **D:** Contribution volume in the pre-observation period

Finally, we also consider the similarity in contributions during the pre-observation period.

- **E:** Contributions in similar subreddits in the pre-observation period. We want to control for the types of subreddits FC and NC contribute in, prior to the observation period. Controlling for contributions in similar subreddits can give us FC and NC users who have tendencies to contribute in similar subreddits. We assess the similarity of subreddit contributions made by FC and NC using the conspiracy scale described in the previous section. The conspiracy scale gives us weights for subreddits based on their similarity to `r/conspiracy` (see Figure 2.1 b for example). Hence, understanding the users' subreddit activity using conspiracy scale can help us match FC and NC that are similarly conspiratorial or non-conspiratorial in the pre-observation period. Moreover, having FC and NC who have a history of contributions in similar subreddits can give us user cohorts with comparable chances of social interactions with other conspiracists. Thus, to compute our final matching criteria (E), we take weighted sum of normalized user contributions using the subreddit's score on the conspiracy scale. For example, if a user has 60% contributions in `r/C_S_T` (-0.36 on scale) and 40% contributions in `r/The_Donald` (-0.32) then the matching feature value is calculated as $-0.36 \times 0.6 + (-0.32) \times 0.4 = -0.344$. To validate that this feature is able to characterize the users' subreddit contributions effectively, we plot the feature values for top 100, 1000 and 10k `r/conspiracy` and `r/science` users and compare their distributions. (See Figure 2.3). In all three cases, Wilcoxon signed rank sum test revealed that the distributions for conspiracy and science users are significantly different which means our contribution similarity calculation is able to characterize different types of users accurately based on their Reddit activity.

2.3.2.4 Matching FC and NC

For each of the 30K FC, we select one NC from a pool of 3 million NC candidates using statistical matching. Since we want to find FC and NC that join Reddit in the exact same month, we perform exact matching on the Reddit start month criteria. We perform nearest neighbor matching with replacement using Mahalanobis distance on the remaining constraints. The matching procedure results in a set of 30,325 FC

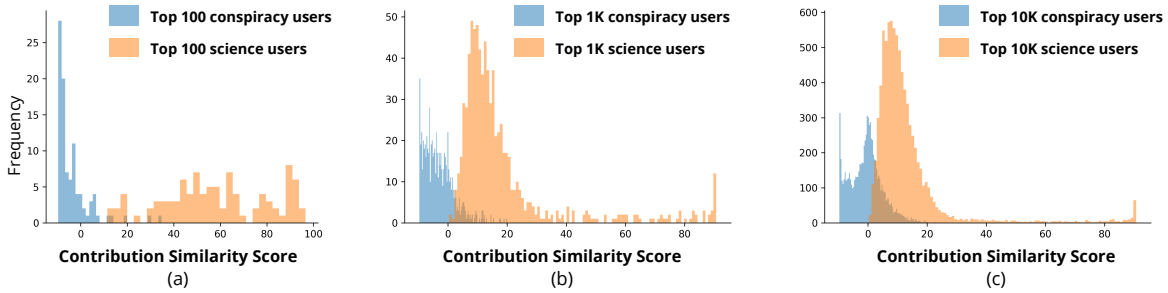


Figure 2.3: We obtained the contribution similarity scores (matching criteria ❺) for (a) top 100 (b) top 1000 and (c) top 10k `r/conspiracy` and `r/science` users. The scores were generated by taking weighted contributions by users (weights are the conspiracy scale weights for the subreddits). The Wilcoxon rank sum test between distributions returned p-values < 0.05 in all three cases. This indicates that our contribution similarity calculation is able to characterize different types of users accurately based on their Reddit activity.

users matched with 29,098 NC users [2]. Note that our matching procedure is more involved than previous empirical studies in conspiracy precursors [141] to ensure highly comparable groups of FC and NC users. We use five criteria (Figure 2.4 (a)) that find similar FC and NC users based on their Reddit joining, volume of contribution across different time periods and also the semantic similarity of subreddits they contribute in. Our intricate matching process that compares different attributes of the users' Reddit activity, enables us to confidently examine the social factors as precursors to conspiracy joining.

2.3.2.5 Quality of matching

To ensure that our matched FC and NC are statistically comparable, we check the improvement in balance across all of the matching constraints using Standardized Mean Difference (SMD)—a method commonly used by other researchers studying users on social media [213, 214]. Note that the FC and NC are matched exactly on the first criteria—❶ Reddit start month. Hence, we claculate the SMD for only the rest of the matching criteria. SMD calculates the difference in the means of distributions between the two groups as a fraction of the pooled standard deviation of the two groups. Balanced groups are considered to have SMD less than 0.2 [137]. We obtain an SMD of less than 0.08 across all of the matching constraints, suggesting high

[2] 29,098 NC users are mapped to 30,325 FC users because we allow replacements in the matching: to ensure the integrity of the results, we consider different observation periods for all FC and NC user pairs, in effect considering one NC user mapped to two FC users as two distinct NC users

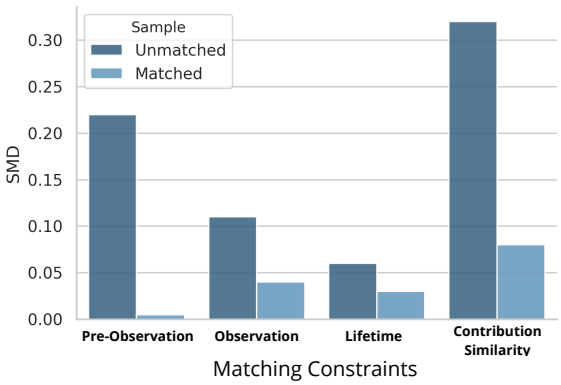


Figure 2.4: Standardized Mean Difference (SMD) for unmatched and matched users across ❷, ❸, ❹, ❺ matching constraints. All four constraints result in the SMD < 0.08 after matching indicating balanced matched groups of future conspiracists and non conspiracists

quality of matching (See Figure 2.4). Specifically, we find 77% balance increment in pre-observation contributions, 63% in observation period and 50% in Reddit lifetime contributions. We find highest balance improvement in contribution similarity scores (80%).

2.3.2.6 Finding Current Conspiracists (CC)

After matching FC and NC, we have a unique observation window for each matched FC and NC user pair (Figure 2.2(a)). Studying interactions with conspiracists in the observation period can inform about the social influence the conspiracists have on the conspiracy joining of FC. Hence, we select a group of users—*current conspiracists (CC)*—that have already joined the conspiracy communities. Each FC and NC has their own set of CC who have already made their first contribution in any of the *conspiracy communities* and who make above average contributions in aggregate in *conspiracy communities* in their lifetime. In other words, for every FC and NC pair, we select their own set of CC based on their unique observation window. In total, there are 61,073 CC involved in the interactions with FC and NC.

2.3.2.7 Characterizing social interactions with CC

For every FC/NC, we characterize their interactions with CC in the observation period. Specifically, we look at publicly available interactions between users in Reddit discussion threads. Figure 2.2(c) demonstrates a discussion thread. We consider "dyadic interaction" as a communication between two users with direct reply to a submission

	Feature group (# features)		Interpretation of feature values
Individual Factors	Psychological Factors (5)	LIWC categories of **anger**, **sadness**, **anxiety** VADER **positive** and **negative** sentiment	↑ values ↑ psychological factors
	Crippled Epistemology (3)	**Exclusivity in contributions:** disproportion of contributions across subreddits measured through Gini coefficient **Subreddit similarity to conspiracy communities:** essentially the matching criteria measured in the observation period **Content similarity to conspiracy discussions:** bag of word vectors similarity to top-scored discussions within the conspiracy community subreddits	↑ value ↑ exclusivity ↓ value ↑ similarity to conspiracy communities ↑ value ↑ similarity to conspiracy content
Social Factors	Availability (3)	**Ratio of dyadic interactions with CC:** number of dyadic interactions with CC normalized by total dyadic interactions **Ratio of CC in dyadic interactions:** number of CC normalized by the number of all Redditors users interact with, through dyadic interactions **Ratio of threads with CC:** number of discussion threads in common with CC normalized by the number of all discussion threads	↑ value ↑ availability
	Information (2)	**Contribution order in dyadic interactions:** reply sent to CC is encoded as 1 and reply received from CC is encoded as -1. The feature represents the sum of all such interactions converted to 1 (is sum is positive) or -1 (is sum is negative) **Time lapse in dyadic interactions:** average of absolute value of time difference between dyadic interactions measures in seconds	↓ value ↑ replies received from CC ↑ value ↑ time lapse
	Reputation (2)	**Age reputation:** average of the Reddit age of the CCs users interact with **Karma reputation:** average of the karma of the CCs users interact with	↑ value ↑ reputation
	Emotion (4)	LIWC **positive** and **negative** affect in user contributions in the dyadic interactions with CC **Coordination in positive and negative affect (2)**: the absolute difference between the user's and the CC's positive and negative affect in dyadic interactions	↑ value ↑ emotion in users ↓ value ↑ coordination in emotion with CC
	Group Polarization (8)	Use of **first-person singular**, **first-person plural**, the **second person** and **third-person** pronouns by users in dyadic interactions with CC **Coordination in the pronoun use (4)** : the absolute difference between the user's and the CC's pronoun use in dyadic interactions	↑ value ↑ use of pronouns ↓ value ↑ coordination in use of pronouns with CC
	Self-selection (3)	**Moderated contributions:** number of moderated contributions normalized by total number of contributions **Negatively scoring contributions:** Total contributions with negative score normalized by total contributions **Contribution trend:** trend of the line fitted on number of contributions in each of the six months in the observation period	↑ value ↑ self-selection

Table 2.3: Table summarizing individual and social factors used in this work. All features are written in **bold** with a concise description. A more detailed intuition behind the features is discussed in Section 2.3.3. The directionality high (↑) or low (↓) indicates how we should interpret the feature values and their corresponding regression coefficient signs in the logistic regression analysis. For example, for the emotion coordination feature, low (↓) value indicates high (↑) coordination between users and CC, i.e., if FC (label 1 in regression) have high coordination with CC, the sign of beta coefficients will be negative for the emotion coordination features.

or a comment. For example in Figure 2.2(c), authors of comment 1 and comment 2 are involved in a dyadic interaction. Figure 2.2(c) also shows examples of other dyadic interactions in the discussion thread.

2.3.3 Factors in Conspiratorial Engagement

Table 2.3 presents a concise summary of individual and social factors that are described below. We look at two main categories of precursors towards conspiratorial engagement—individual factors and social factors. While individual factors are designed to reflect the users' predisposition towards conspiracies, social factors capture

their engagement with the members of the conspiracy communities prior to joining those communities.

2.3.3.1 Individual Factors

Why do people believe in conspiracies even when there is a lack of well reasoned evidence? This question taps into a popular debate of whether conspiratorial belief emerges from psychological predisposition, or from other aspects such as an individual's exposure to biased information and to triggering events. We attempt to capture both arguments while measuring individual predisposition.
Psychological Factors: We explore the presence of psychological factors through analyzing sentiment and affective words in the contributions made by FC and NC in the observation period. Specifically, based on previous research associating conspiratorial belief with anxiety, paranoia and insecurity [38, 109], we measure users' proclivity to such psychological factors as follows.

- ***Cognitive and affective processes***:Researchers have argued that words and language reflect psychological states [245]. We measure Linguistic Inquiry and Word Count (LIWC) [192] categories of anger, sadness, and anxiety in the contributions made by users in the observation period, normalized by total number of contributions.

- ***Sentiment***: We calculate average VADER sentiment scores for positive and negative sentiment [107] in the user's contributions during the observation period.

Crippled epistemologies: Sunstein et. al. coined the term to refer to a scenario when an individual's tendency to adhere to limited information sources results in their epistemological isolation [243]. Thus, a conspiracy theory, which is otherwise unjustified relative to all the information available to the wider society, might be perfectly justified to someone whose worldview is already distorted due to the absence of relevant and ample information. The tendency to adhere to epistemologically isolated information sources increases the likelihood to accept conspiracy theories. On Reddit, users can exhibit crippled epistemologies by refraining from participating in diverse communities, participating in communities that might foster a conspiratorial worldview, and contributing content similar to the conspiratorial themes.

- ***Exclusivity in contributions***: Do the FC and NC exclusively contribute in fewer subreddits or do they spread their Reddit activity evenly over multiple subred-

dits? We characterize exclusivity by calculating Gini coefficient of disproportion on the subreddit contributions made by the users. The feature value varies between 0 to 1 with higher values indicating high exclusivity in subreddit contributions.

- ***Subreddit Similarity to conspiracy communities***: Apart from subreddits in our carefully compiled list of 56 conspiracy communities, Reddit has other subreddits that even though, not dedicated to conspiracy theories, occasionally host conspiracy related discussions (for example, `r/The_Donald`). Higher engagement in such subreddits might indicate that users are being exposed to conspiratorial themes. The conspiracy scale introduced in section 3.2 characterizes subreddits based on how similar they are to `r/conspiracy` compared to `r/science`. Thus, for every user, we weigh the contributions made in each subreddit by the subreddit's score on the conspiracy scale. We consider the sum of all weighted contributions as a subreddit similarity feature. Remember that we used similar computations to match the FC and NC based on their contributions in the pre-observation period. Hence, we do not expect this feature to have significantly different values for FC and NC. However, measuring the significance of this feature in the observation period can contextualize the observations about other individual and social factors.

- ***Content similarity to Conspiracies***: Another way of measuring exposure to conspiracies is to compare the actual content produced by FC and NC with the discussions inside the conspiracy communities. Top ranking posts can distinguish subreddits along the dimensions of topics, style, audience and moderation [120]. Hence we compile a list of top 10 scored submissions from every subreddit in the conspiracy communities as a representative corpus of conspiratorial discussions. Further, we also create a corpus for every FC and NC by combining their contributions in the observation period. Next, we create Bag of Words (BoW) representations for every corpus after cleaning the text data and removing stop words. As previously discussed, subreddits within the conspiracy communities vary in their interests (general conspiratorial discussion vs. specific conspiracies). In order to capture this variance in conspiratorial discussions, we calculate the cosine similarity scores for the user's BoW vector with all subreddits in the conspiracy communities. Finally, we take the maximum cosine similarity score as the user's content similarity feature.

2.3.3.2 Social Factors

How does socializing with members of the conspiracy communities affect FC's joining behavior? We quantify the social factors by analyzing users' online interactions with *current conspiracists (CC)*. Based on Sunstein et. al's framework [243], we study various statistical, temporal and linguistic aspects of the interactions between the users and CC. Below I outline the characterization of social features across various dimensions.
Availability: Conspiratorial beliefs may flourish upon *availability* of conspiratorial materials [243]. On Reddit, interactions with other conspiracists is what makes conspiratorial content available to users who are yet to join these communities. Thus, to understand the prominence of such interactions in our two user cohorts, we introduce three features.

- ***Ratio of dyadic interactions with CC:*** Dyadic interactions are pairwise interactions (Figure 2.2 (c)) where either user replies to CC or vice-a-versa and can provide venues where conspiratorial content is available to users through other conspiracists. We count the proportion of such dyadic interactions with CC normalized by all dyadic interactions the user has on Reddit in the observation period.

- ***Ratio of CCs in dyadic interactions:*** In addition to dyadic interactions, the amount of conspiracists engaged with users can also signal the exposure to available conspiratorial content. Do FC or NC engage with just one or multiple CC? This feature captures the number of CC that users engage with through dyadic interactions normalized by number of all Reddit users they interact with via dyadic interactions.

- ***Ratio of threads with CC***: While dyadic interactions are strong indicators of information exchange, users are also exposed to the contributions made by CC in the overall thread. For example, in Figure 2.2, it is possible that the author of comment 4 has read comment 1 even without a direct interaction. To understand if users passively consume the content written by CC without directly engaging with them, we also consider the number of threads on which the user and CC appear together. Specifically, we calculate the user's co-presence with CC in threads by counting the total number of threads with CC normalized by the number of all threads the user participates in during the observation period.

Information: What role does information play in the conspiratorial engagement? Researchers argue that conspiratorial beliefs can be a product of informational pressure built through social interactions [243]. For example, conspiracy theories are often initially accepted by people with low thresholds of acceptance. Informational pressure builds through social interactions with such people to the point where others even with higher acceptance threshold begin to accept the theory [243]. We consider CC as Redditors with lower acceptance threshold as they have already made contributions in the conspiracy communities. Towards understanding the role of information in conspiratorial engagement, we focus on two temporal characteristics of the dyadic interactions:

- ***Contribution order in dyadic interactions***: Do users reply to the contributions made by CC or do they often receive a reply from CC? If the user normally replies to CC, it can indicate that she is exposed to the opinions expressed by conspiracists. Sunstein et. al. claim that this can build informational pressure that can result in conspiratorial thinking. We encode every direct interaction as 1 if the user replies to CC and as -1 if the user receives as reply from CC. We aggregate this measure for all dyadic interactions by the user and consider the feature value as 1 if the sum is positive and -1 if the sum is negative. In other words, contribution sequence value of 1 indicates that the user more commonly replies to the CC.

- ***Time lapse in dyadic interactions***: How much time do users take to process the information they are exposed to by interacting with the CC? A small time duration between the interaction may indicate that users have less time to rationally consider all the information available and may tend to rely on other's information and judgment to form their opinion. While contribution order captures whether the users contribute before or after CC, time lapse feature measures the average absolute time differences in seconds between dyadic interaction. Smaller value of time lapse means the user contributes shortly before or after CC.

Reputation:

When users interact with conspiracists, the reputation of conspiracists can also exert additional pressure to join the conspiratorial belief system [243]. Due to the reputational pressure, people often ignore their own beliefs to avoid social sanctions. We characterize reputation on Reddit with two features—account age and karma of the (CC) that NC or FC interacts with.

- ***Age reputation***: Does seniority of conspiracists exert a reputational pressure on potential joiners? We first calculate the age of a CC at the time of his last direct interaction with NC or FC in the observation period. Next, for every user in our NC, FC cohort, we calculate the age reputation feature as the average account ages of all conspiracists they engage through dyadic interactions. We consider CC's account age at the time of the latest interaction with FC in the observation period.

- ***Karma reputation***: Redditors can accumulate *karma* through up-votes and down-votes on their contributions. We first find the aggregate karma of a CC user at the time of their latest direct interaction with NC or FC during the observation period. Next, for every user in our NC, FC cohort, we calculate the average karma accumulated by all CCs that users interacted with in the observation period.

noindent *Emotion:* Are emotions exchanged during interactions important towards conspiratorial belief? Sunstein et. al. argue that "emotional selection" could be an important aspect towards understanding the spread of conspiracies [243]; people select content that justify their emotional state. Studies have also shown that discussions involving personal accounts and rumours that elicit intense emotional response are likely to spread from one person to another [114]. Hence, we quantify this emotional snowballing by measuring the LIWC categories of positive and negative affect words in the dyadic interactions between the user cohorts (FC and NC) and CC.

- ***Affective process in dyadic interactions***: For every direct interaction between the users and CC, we calculate the presence of LIWC's positive and negative affect category words. The aggregate positive and negative affects averaged over number of dyadic interactions represent the affective processes in dyadic interactions.

- ***Coordination in affective processes***: Do the CC reflect the same affective state as the FC and NC? While the previous feature measures the affective processes in the contributions made by FC and NC, it is also important to understand how similarly or differently the CC counteract. We measure the coordination between the affective state within dyadic interactions as follows: for every interaction, we subtract the affective state values in the contribution by CC from those in the contribution by the user. The average of this difference over all user interactions represents average coordination in the user's affective state with the CC. Lower

values of the feature should indicate that users closely replicate the affective states of CC.

Group Polarization: Belief in conspiracy theories is often strengthened through strong group identity [93, 243]. Prior research have found that when group members—or, in-group—have a shared sense of identity and solidarity, they often discard the arguments by outsiders—the out-group—as non-credible. This suggests that if users from our *NC, FC* cohort relate to the identity of current conspiracy (CC) group members, then would likely also adopt the group's conspiratorial beliefs. One way to measure the sense of group identity is by analyzing how users and conspiracists use pronouns in interactions [129, 191, 245]. For example, first person singular pronouns can signal high self and group awareness while second and third person pronouns can indicate that users are socially interactive with larger Reddit audience [191].

- ***Use of pronouns in dyadic interactions*** : We count the average use of first person singular (I, me etc.) first person plural (we, us etc.), second person and third person pronouns in the contributions made by the user in dyadic interactions with the CC

- ***Coordination in the use of pronouns*** : We also measure the difference between the use of pronouns between the user and CC for all pronoun features mentioned above.

Self-selection: Other than exposure to limited relevant information, crippled epistemology can also develop from social self-selection [243]. As people start developing increasingly extreme conspiratorial views, they might suffer from social segregation from others with differing ideologies. Hence, we measure self-selection by observing the extent of social sanctions placed on a user's content contribution during their observation period. It comprises the following features.

- ***Moderated contributions:*** Users can feel ostracized on Reddit by having their contributions moderated. Most subreddits have content moderation policies. Contributions that violate the subreddit rules are often removed. We calculate the number of moderated contributions normalized by total number of contributions in the observation period to understand social sanctions placed on a user's contribution.

- ***Negatively scoring contributions:*** Apart from moderation, users can also face sanctions by receiving more negative scores. Contributions on a subreddit accumulate scores via upvotes and downvotes cast by others. Negative score indicates more downvotes than upvotes. Thus, we calculate contributions with total negative scores normalized by the total number of contributions.

- ***Contribution trend in the observation period:*** Users may join the conspiracy communities not only because they are ostracized outside of it, but also because they generally disengage from society. To measure disengagement, we compute the decrease in their participation in the observation period. We calculate the number of contributions per month, and fit a line via least squares regression. We take the trend of this line as the contribution trend in the observation period: a negative trend corresponds to a decrease in participation.

2.3.4 Understanding the Importance of Features

In order to evaluate the importance of individual and social factors towards conspiratorial engagement, we construct a series of logistic regression models (see Table 2.4). The dependent variable is binary and represents the type of user cohort, FC (1) and NC (0). We interpret the importance of features by comparing their regression coefficients (β values) in the logistic regression models. If features have multicollinearity—two or more features are highly related—it can lead to poor estimation of β coefficients. Thus, we tested for multicollinearity in features through Variance Inflation Factor (VIF). If any feature has VIF > 5.0 then the group of features is considered to have high multicollinearity [224]. We found all features to have VIF < 4.0 suggesting low multicollinearity. Additionally, all features vary in their means and standard deviations and variable types, such as counts, time in seconds and proportions. Hence we standardize the features for the regression analyses. Due to the high number of features and multiple testing, our model could have an increased risk of false significance. However, lower p-values have lesser chance of significance errors [85]. Hence, we report p-values in different thresholds. Specifically, (**p < 0.001**, $p < 0.01$, $p < 0.05$) in Table 2.4. Most of the p-values in the regression models are less than 0.001.

2.3.4.1 How important are social features?

We treat individual features' model as a baseline to compare how much value do social features add to the regression model. Below, we discuss each social feature group separately.

Availability: By adding *availability* features we observe improvement in the pseudo R^2 compared to the individual factors model (0.21 vs. 0.12). That is, the model containing both individual and availability features fits better than the model with just individual features. Specifically, in comparison to NCs, FC users have higher proportion of dyadic interactions (β = 1.58) and more threads in common (β= 0.12) with CCs. Moreover, out of all Redditors they interact with, the proportion of CCs they interact with, is larger for FC (β= 0.33) compared to NC. Upon adding the availability features, the subreddit similarity in individual factors becomes insignificant, indicating a possibility of partial mediation effect. Overall, high number of dyadic interactions with CC and co-presence with CC indicate intimacy with current conspiracists; intimacy is one of the four tie strength dimensions proposed by Granovetter [111]. This may suggest that FC form strong ties with CC in the observation period.

Information: Information features capture the order of contributions, i.e, whether users usually reply to the CC or vice a versa, and the time lapse between the dyadic interaction. Remember that negative feature value of contribution order indicates more replies received from CC as opposed to positive which indicates more replies sent to CC. The negative and significant beta coefficient for the contribution order feature (β= -0.13) thus implies that FC often receive more direct replies from the current conspiracists. Together with increased direct interactions with CC in general, this might reflect efforts on part of CC to engage with FC. We find no significant differences in the time lapse between the dyadic interactions of user cohorts and CC.

Reputation: While there is no difference between the seniority of CC that FC and NC interact with, they do differ in the average karma. On Reddit, comment and submission karma can indicate how well the user's opinions are accepted by other Reddit users. We find that FC interact with CC having lower karma (β=-0.08). It is possible that the current cospiracists who feel rejected by other Reddit users through lower karma are reaching out to a group of users predisposed towards conspiratorial thinking—FC—by direct interactions outside of the conspiracy communities.

Emotions: What role do emotions in interactions play in conspiratorial engagement? Adding *emotion* features, we observe an interesting coordination between affective state of FC and NC. Firstly, FC and NC do not differ in the positive and negative affect

word use in dyadic interactions with CC. However, CC still closely reflect the affective state of FC more compared to NC. To elaborate, coordination features are calculated by taking the absolute difference between the user's and CC's respective emotion states. Hence, lower values represent higher coordination. Thus, FC have higher coordination of negative (β=-0.02) and positive (β=-0.05) affective state with CC compared to NC.
Group polarization: For the group polarization features, we measure the use of pronouns by the users in dyadic interactions and how similar their pronounce usage is to the use of pronouns by CC in dyadic interactions. While talking with CC, FC use first person pronouns more compared to NC. Interestingly, FC also have higher coordination in the first person plural (β=-0.04) and third person (β=-0.05) use with CCs. Previous research has associated higher use of first person plural and third person pronouns with strong group identity [124, 245]. This means that even before joining conspiracy communities, FC communicate in the language of "we", "us", "ours" with CC expressing higher group identity.
Selection: How does self-selection affect conspiracy engagement? Do users get more sanctions and negative feedback from others? *Selection* features capture the amount of moderation, negative scores and users' disengagement during the observation period. We find that all *selection* features are significant. Specifically, FC have more moderated (β=0.06) and negatively scored contributions (β=0.18) compared to NC. However, despite facing more social sanctions, FC still increase their contribution rate compared to NC in the observation period (β=0.09).

2.3.4.2 How important is each of the social feature groups?

All social feature groups contain at least one significant predictor of joining conspiracy communities. In fact, some social factors are overall better predictors than individual factors. Notably, the ratio of dyadic interactions with CC (β=1.58) is the single most important feature in the regression model. Furthermore, iteratively adding each group of social features consistently increases model performance. These findings corroborate that social factors play a sizable role in predicting users joining conspiracy communities *even after controlling for individual factors*. To understand the relative importance of feature groups amongst social factors, we add each of them separately to the individual features, and compare their percent increase in explained variance (R^2). Figure 2.5 displays a bar chart indicating relative increase in the R^2 value over individual features. We find that among all the social features, *availability* features are the most informative (75% increase), followed by *selection* (19%), *reputation* (8%), *information* (7%), *emotion*

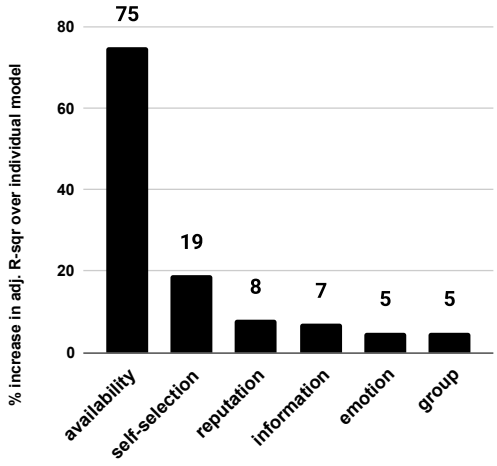


Figure 2.5: (a) Six social feature groups and their relative importance as percent increase in R^2 compared to individual features (more details in Section 2.3.4.2). Note that we add only one social feature group at a time to the individual model and calculate the percent increase in the R^2. For example, availability features improve the R^2 of individual model by 75% while self-selection feature improve it by 19%.

(5%), and *group polarization* (5%). In summary, we find evidence that the different social factors hypothesized in [243] capture specific, complementary, and relevant aspects of the joining behavior—although in varying amounts.

2.4 Discussion and Implications

Our current understanding of social factors in conspiracy adoption is assembled from mainly theoretical studies. This work calls into attention the importance of *empirically* studying how interactions with current members can influence the user's joining into the conspiracy communities. By proposing a theoretical-motivated, quantitative operationalization of social factors across six dimensions, we take a step in this direction. Specifically, we provide an empirical representation of social features proposed by Sunstein [243] in six groups: 1) importance of social availability of conspiracists, 2) informational pressure, 3) reputational pressure, 4) emotional snowballing, 5) group identity, and 6) self-selection towards conspiracy adoption. We compared the social factors with the strong baseline of individual factors from literature [38, 65, 109], and found that social factors are crucial precursors of joining conspiracy communities. Not only social factors as a whole add significant explanatory power over individual factors, but each of the six dimensions also contains significant predictors that capture

separate and complementary facets of conspiracy theory adoption. Our findings bring forth several implications for understanding how the conspiracy communities form, how they maintain their echo chamber, and how social exclusion may lead to joining [3].

2.4.0.1 Selection, Evocation, and Manipulation Among Joiners

How do conspiracy communities grow? We refer to Buss's [37] proposed causal mechanisms of individual—community correspondence to understand how future conspiracists first select, and then assimilate into conspiracy communities [37]. Buss presents three key mechanisms: *selection*, whereby individuals decide to participate in a social group based on personal preference or mere proximity; *evocation*, where individuals elicit emotional responses from the group in order to make connections; and *manipulation*, whereby they use their position in newly found environment to change it. We find that users produce content similar to discussions within the conspiracy communities (β= 0.61) and increasingly interact with conspiracy members before joining the conspiracy communities (β = 1.58). Together, these findings show the hallmarks of the *selection* process of conspiracists' future social group. Next, we observe *evocation* in how future and current conspiracists coordinate their online messages. Specifically, we find that the affective states and group identity signals of future conspiracists closely mirror those of their future social group—current conspiracists (See Table 2.4 last column). The present work studies the precursors to joining the social group, and therefore it does not directly observe future conspiracists' behavior after they become members of the community. According to Buss, in the *manipulation* phase, conspiracists would take on an active role in gatekeeping their newly found community. In particular, we see that current conspiracist may play a crucial role in recruiting new members though dyadic interactions (β = 1.58), although whether that is on purpose remains unanswered. Dyadic interactions between future and current conspiracists are at the nexus of the former's selection of a community to belong, and the latter's attempts to shape it. Studying this negotiation is essential to understand how conspiracy communities self-sustain and thrive. Our work offers crucial insight in this direction.

[3] while referencing the β coefficients downstream, we refer to the numbers only from the last column of Table 2.4

2.4.0.2 Conspiracy Communities as an Echo Chamber

Previous studies posit that consumers of conspiracy-like content are likely to aggregate in homophilic clusters—i.e. "echo chambers" [25]. In fact, conspiracy theorists are renowned for their commitment to conspiratorial attitudes, and this may come from limited access to contradicting information early on [243]. Our results empirically corroborate previous work's hypothesis that future conspiracists live in an information bubble. In fact, not only do they contribute content similar to conspiracy discussions (β = 0.61), they also engage disproportionately in subreddits similar to those in the conspiracy communities (β = 0.09). Apart from such informational isolation, echo chambers can also result from fragmentation of communities where like minded people come together to discuss ideas through a very narrow world-view. Our results indicate that along with exposure to conspiratorial material, users also directly interact with members of the conspiracy communities (β = 1.58) and current conspiracists make up a significant fraction of their social circle on Reddit (β = 0.33). While we do not claim that the information discussed in such interactions is strictly conspiracy related, the relevance of both epistemological and social isolation indicate that future conspiracists may be living in their own informational echo chamber circulating similar conspiratorial content even prior to joining conspiracy communities.

2.4.0.3 Social Stigma and Joining the Conspiracy

We uncover an important factor in joining conspiracy communities: marginalization from other communities. Through self-selection features we find that future conspiracists are ostracized from subreddits outside of the conspiracy communities through negative feedback from other members of those communities (β = 0.18) and content moderation (β =0.06), significantly more than non-conspiracists. Future conspiracists express anxiety and negative sentiment in the months leading up to their joining (psychological predisposition, Table 2.4). We give an interpretation of how this may affect the formation of conspiracy groups. While discussing deviance as a social construct, Becker proposed that groups create rules to define what they (subjectively) consider to be desirable behavior [21]. As a consequence, people who break such rules are labeled as deviants and criminalized. Despite their popularity, the public image of conspiracies is still tainted, and conspiratorial thinking bears the stigma of deviance. A two-fold process can then explain joining conspiracy communities. First, social sanctions make users feel like outsiders in mainstream subreddits. Such socially outcast users then

find home in the conspiracy communities for their rejected thoughts.

2.4.0.4 Implications for content moderation:

Researchers have found that moderators notice repeat offenders—users who have already faced sanctions before—and partially focus their moderation efforts on them [148]. We observe that future conspiracists already start facing social sanctions in terms of content moderation ($\beta = 0.06$) and negative karma ($\beta = 0.18$) prior to joining conspiracy communities. We argue that this type of ostracizing may exacerbate the segregation of future conspiracists and drive them to contribute in communities that accept their conspiratorial worldview. Therefore, community managers and social platform may play a determining role in the creation of conspiracy communities. A mindless application of norms that are too rigid may ultimately ostracize non-conforming individuals, thus running the risk of driving them into fringe groups.

2.4.0.5 Theoretical Implications:

Our study engages with methodological challenges of using observational data to implement a theoretical framework for understanding social factors in conspiratorial joining. We explore beyond the purely theoretical framework and quantitatively establish the importance of different social factors on large-scale online discussion communities. We further test the generalizability of our individual and social factors for topic-specific and general conspiracy joining. Although prior work largely framed conspiracism as an individual pursuit, focusing on psychological disposition [38, 109, 118] and epistemological characteristics [243] our results support a socio-constructionist view of conspiracy theory. In particular, this view grants drawing the parallel between discussing conspiracy theories, and entering the community that hosts those discussions. Our analysis and results focusing on social factors in conspiracy engagement provide us with a unique opportunity to consider conspiracies as social movements—"a network of interactions between groups of individuals or organizations, engaged in a political or cultural conflict, on the basis of a shared collective identity" [74]. For instance, in the case of conspiracist belief, collective identity based on political ideology can lead to upholding different types of anti-government conspiracies. Republicans are more likely to believe that a "Deep State" is colluding against President Trump [254] whereas Democrats more commonly believe that 9/11 was an inside job [230]. Characterizing conspiracies as social movements becomes more relevant when conspiracism has a potential to turn into conspiracy *activism*

towards a cause with detrimental consequences. Consider the anti-vaccination movement set in motion by anti-vaccination conspiracies which has directly resulted in lower herd immunity. Analyzing conspiracies through a social movement lens can open up further research avenues exploring how conspiracists frame their narrative, mobilize informational resources, and ultimately coordinate collective action.

2.5 Limitations

Our work has some limitations which also pave the way for promising future directions. We characterize engagement within the conspiracy communities based on the number of contributions a user makes in conspiracy subreddits; contribution based approach is a common methodological choice made when studying users in social media [112]. A more robust definition of engagement could involve analyzing the topics and synchronicity of the user's contribution content with that of the community. For example, the criteria for selecting FC could be made stricter by keeping only those who discuss topics similar to conspiracies. Additionally, similar to any observational quantitative research, we can not infer true causality. While acknowledging this, we believe that our work is an important step towards understanding the variety of statistical, temporal and linguistic social factors towards conspiracy joining in a complex, real world setting. However, we take this opportunity to invite further qualitative studies investigating conspiracy joining using insights provided in our work. While testing the robustness of social features we consider only one dichotomy—topic specific and general conspiracy discussion subreddits. Fruitful path for future exploration could be to check how social factors vary for conspiracy joining in smaller or larger subreddits or, political or non-political conspiracy subreddits. Finally, our results exemplify conspiracy joining on just one online platform—Reddit. We do not know how these results translate to other platforms such as Facebook or Gab with various levels of content moderation. We encourage future researchers to build up on our findings and to explore conspiracy joining across multiple platforms.

2.6 Conclusions

In summary, currently, our understanding of social factors in conspiracy adoption is patched together by mainly theoretical and very few empirical studies. This work calls into attention the importance of systematically studying how interactions with

conspiracists can influence the user's joining into the conspiracy communities. Using a theory driven framework of social factors across six dimensions, we perform a retrospective case control study of *future conspiracists* and compare them with *non conspiracists*. We not only find that the social factors are important but that conspiracy joining can be explained at least partially by at least one feature in each group. Given these findings, we offer a unique, empirically backed perspective on the life-cycle of conspiracists, echo chambers in conspiracy communities and the effect of social exclusion in conspiracy engagement.

	individual	+ availability	+ information	+ reputation	+emotion	+group	+selection
INDIVIDUAL FACTORS							
Psychological Predisposition							
Anger	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Anxiety	**0.06***(0.01)	**0.06***(0.01)	**0.06***(0.01)	**0.06***(0.01)	**0.07***(0.01)	**0.07***(0.01)	**0.06***(0.01)
Sadness	**-0.05***(0.01)	**-0.06***(0.01)	**-0.06***(0.01)	**-0.06***(0.01)	**-0.06***(0.01)	**-0.06***(0.01)	**-0.06***(0.01)
VADER Positive Sentiment	**0.15***(0.01)	**0.16** *(0.01)	**0.16***(0.01)	**0.16***(0.01)	**0.17***(0.01)	**0.18***(0.01)	**0.18***(0.01)
VADER Negative Sentiment	**0.78***(0.01)	**0.86***(0.01)	**0.87***(0.01)	**0.87***(0.01)	**0.87***(0.01)	**0.88***(0.01)	**0.86***(0.01)
Episteomology							
Exclusivity in Contributions	**0.23***(0.01)	**0.10***(0.01)	**0.10***(0.01)	**0.09***(0.01)	**0.10***(0.01)	**0.09***(0.01)	**0.09***(0.01)
Subreddit Similarity to conspiracy communities	**-0.07***(0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
Content Similarity to Conspiracies	**0.56***(0.01)	**0.59***(0.01)	**0.59***(0.01)	**0.59***(0.01)	**0.59***(0.01)	**0.60***(0.01)	**0.61***(0.01)
SOCIAL FACTORS							
Availability							
Ratio of Dyadic interactions with CC		**1.54***(0.02)	**1.55***(0.02)	**1.55***(0.02)	**1.55***(0.02)	**1.56***(0.02)	**1.58***(0.02)
Ratio of CC in dyadic interactions		**0.35***(0.02)	**0.34***(0.02)	**0.34***(0.02)	**0.35***(0.02)	**0.34***(0.02)	**0.33***(0.02)
Ratio of threads with CC		**0.12***(0.02)	**0.12***(0.01)	**0.12***(0.01)	**0.12***(0.01)	**0.12***(0.02)	**0.12***(0.02)
Information							
Contribution Order in Dyadic Interactions			**-0.13***(0.06)	**-0.12***(0.06)	**-0.12***(0.06)	**-0.12***(0.06)	**-0.13***(0.06)
Time Lapse in Dyadic Interactions			0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Reputation							
Age Reputation of CC				-0.02 (0.01)	-0.01 (0.01)	-0.01 (0.01)	-0.01 (0.01)
Karma Reputation of CC				**-0.08***(0.01)	**-0.08***(0.01)	**-0.08***(0.01)	**-0.08***(0.01)
Emotion							
LIWC Positive affect					0.02 (0.01)	0.01 (0.01)	0.01 (0.01)
Coordination in positive affect					**-0.05***(0.01)	**-0.05***(0.01)	**-0.05***(0.01)
LIWC Negative affect					0.03 (0.01)	0.04 (0.01)	0.03 (0.01)
Coordination in negative affect					**-0.01***(0.01)	**-0.01***(0.01)	**-0.02***(0.01)
Group Polarization							
First person singular						**0.04***(0.01)	**0.04***(0.01)
Coordination in first person singular						0.01 (0.01)	0.01 (0.01)
First person plural						**0.04***(0.01)	**0.04***(0.01)
Coordination in first person plural						**-0.04***(0.01)	**-0.04***(0.01)
Second person						0.01 (0.01)	0.01 (0.01)
Coordination in second person						0.02 (0.01)	0.01 (0.01)
Third person						0.01 (0.01)	0.01 (0.01)
Coordination in third person						**-0.05***(0.01)	**-0.05***(0.01)
Self-selection							
Moderated contributions							**0.06***(0.01)
Negatively scoring contributions							**0.18***(0.01)
Contribution trend in the observation period							**0.09***(0.01)
intercept	-0.10	0.07	0.07	0.07	0.07	0.06	0.07
Adjusted R^2	0.12	0.21	0.22	0.23	0.23	0.24	0.26
Accuracy	0.64	0.71	0.72	0.72	0.72	0.72	0.73
Precision	0.63	0.72	0.73	0.73	0.73	0.73	0.74
Recall	0.65	0.68	0.68	0.69	0.69	0.69	0.70

Table 2.4: Results of the regression analysis. In each column we successively add social feature groups to the individual features and observe β values and adjusted R^2. Significant β values are followed by *. We color code the numbers according to p-values as follows: (**p < 0.001**, $p < 0.01$, $p < 0.05$). β values stay fairly consistent throughout all models and most features do not change their significance indicating limited or partial mediation effects. Accuracy, precision and recall were calculated using five-fold cross-validation. Our R^2 values are fairly consistent with the ranges reported by other researchers studying complex social phenomena [46, 269]. Note that various social feature groups can be successively added to the individual feature model in any order. Hence, we do not use regression performance from this table to compare social feature groups with each other.

CHAPTER 3

COMMUNICATION PRACTICES IN COMMUNITIES OF PROBLEMATIC INFORMATION

Published Works

Phadke, Shruti, and Tanushree Mitra. "Many faced hate: A cross platform study of content framing and information sharing by online hate groups." Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. https://dl.acm.org/doi/10.1145/3313831.3376456

Phadke, Shruti, and Tanushree Mitra. "Educators, Solicitors, Flamers, Motivators, Sympathizers: Characterizing Roles in Online Extremist Movements." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-35. https://dl.acm.org/doi/10.1145/3476051

Submitted Work

Phadke, Shruti, and Tanushree Mitra. "Characterizing Political Campaigning with Lexical Mutants on Indian Social Media."

The last chapter discusses individual and social factors contributing towards users' joining of communities of problematic information. What happens after users join such communities? In this research arc, I will explore the communication practices in communities of problematic information with a case study of hate groups and movements within the United States. My research goals here are twofold: (1) understanding how white supremacy, anti-LGBTQ hate groups within USA leverage various social media spaces to propagate problematic narratives and information, and, (2) how problematic information related to hate movements spreads through various online groups. I will start by providing a background on hate movements within USA and how they are

tied with problematic information, along with the theories that can explain communication practices within hate groups. I argue that virtual communities formed around discussing hateful ideologies, such as racial supremacy or anti-LGBTQ activism, can be considered as social movements. For analyzing hate group communication thought the lens of social movement, this research also leverages theories of framing in social movements and resource mobilization to study the research questions.

3.1 Background

3.1.1 Hate groups as social movement organizations

Social Movement Organizations (SMOs) are purpose-driven organizations with societal reconstruction agendas [160]. Consider the "FRC": Family Research Council group on Twitter with 32,000 followers and over 30,000 tweets. FRC posts around 10 tweets per day informing their followers of the latest news and offering discussion of relevant issues. In many ways, FRC can be seen as an ideal model for social media as it brings people together bounded by one idea and creates an open platform for public discussion. The FRC is also an SPLC-identified hate group with anti-LGBT ideology. To quote from the FRC website:

> *Family Research Council believes that homosexual conduct is harmful to the persons who engage in it and to society at large, and can never be affirmed.... While the origins of same-sex attractions may be complex, there is no convincing evidence that a homosexual identity is ever something genetic or inborn. We oppose the vigorous efforts of homosexual activists to demand that homosexuality be accepted as equivalent to heterosexuality in law, in the media, and in schools. Attempts to join two men or two women in "marriage" constitute a radical redefinition and falsification of the institution, and FRC supports state and federal constitutional amendments to prevent such redefinition by courts or legislatures...*

Groups like the FRC can be classified as Social Movement Organizations (SMOs) for their aggressive attempts to challenge the mainstream political order by pushing their ideologies and actively recruiting supporters [74]. By definition, an SMO exists only when changes in a society are misaligned with the organization's goals [280]. Thus, whenever society witnesses increased racial, sexual, political or religious diversity, hate groups tend to be more active and aggressive in their efforts [280]. Moreover, their

very existence as an SMO is dependent on the readiness for mobilization of potential supporters [280]. Hence, under the SMO perspective, hate groups need to "frame" their communication to legitimatize their actions and campaign, inspire potential recruits, negotiate a shared understanding of the problematic societal condition that needs change, make attributions about who or what is to blame for the societal condition, offer alternative arrangements to promote change, and finally urge others to act so as to affect that change [23].

3.1.2 Hate movements and content framing on social media

This subsection reviews background and context of hate movements in the US. I start with discussing the existing research studying the presence of hate groups on social media and how hate groups may forge influential narratives though event framing.

3.1.2.1 Hate groups and Social Media

Charleston church shooting, Pittsburgh synagogue shooting, New Zealand mosque shooting are just a few of the many incidents that have recently reinforced concerns about organized hate on social media. In a mounting number of hate crimes against various minorities, perpetrators sought support and publicized their actions on various social media platforms. Social media can provide an efficient and fast communication for hate groups to exchange information [138] and spew radical beliefs and activism, amplifying what are otherwise fringe opinions [113].

Since its earliest days, information and communication technologies have served as an attractive conduit for hate group operations [106, 154]. In recent years, the rise of social media has opened additional avenues for hate groups to profess extreme ideologies, champion their cause, recruit members, and spread hateful content. According to the Southern Poverty Law Center (SPLC)—an organization dedicated to monitoring hate group activity in the United States—the number of active hate groups have increased more than 70% in the last few years [22]. While scholars have studied what behaviors place people at risk of viewing extremist content online [58], investigated ways of countering hate group narratives [115, 176], and more recently offered automated ways for detecting online abuse and personal attacks [45, 277, 278], few have investigated the deeper communication strategies that hate groups adopt to target a collective.

3.1.2.2 Hate groups and content framing

The appeal of hate group communication rests in part on framing the message such that it aligns with the ideology and identity of the hate group or in-group, while directing hatred or degrading attitudes toward the targeted out-group [57, 113, 233]. For example, racial extremists, anti-LGBTQ activists, or religious zealots—the in-group from their perspective—campaign against a group of people—the out-group, again from the hate group's perspective—because of their race, ethnicity, sexual orientation, gender identity, or religious beliefs, respectively. In other words, hate groups with societal reconstruction agendas utilize "framing" principles to make their social media communication more socially tolerable, effective, and forceful for the in-group, while making it more antagonizing, vilifying, and hateful for the out-group. In this research, we adopt the principles of framing from social movement theory [110] and ask: how do hate groups frame their communication to align with their identity and ideology, while directing degrading attitudes towards their targets? I explain the theoretical model behind content framing in detail, in the next subsection.

3.1.3 Framing theory

Framing is defined as a process of assigning meaning and interpreting relevant events to guide the actions of, garner support for, and weaken enemies of an organization [23, 110, 231, 233]. Specifically, hate groups "frame" social events in a way that invite supporters and motivates followers (i.e. the "in-group"), while simultaneously demobilizing the targeted population or the "out-group" [128, 232]. Framing is a wildly explored concept in social science studies [82, 231], political communication research [220], and in areas like marketing [63] which involves influence of any kind. More recently, the CSCW community has begun to explore the role of "framing" in online communication technologies. For example, Diamond et. al. used "framing" theories to analyze effects of online collective storytelling in the context of a social movement organization dedicated to end street harassment [75]. In policy studies, framing has been shown to give better understanding of how people comprehend policy issues and political events [24]. Framing a policy issue from one perspective while excluding an alternative perspective is a well studied phenomenon in political communication (see [220] for a review). It is known to influence public opinion and attitudes toward the issue [50, 174].

Most social science and political communication studies in framing involve de-

veloping an extensive codebook of frames after deep manual content analysis of a large number of documents, followed by annotating documents for the presence or absence of frames as per the codebook [19, 247]. Recent work by Boydstun et. al. have offered a "Policy Frames Codebook" as an annotation scheme for advancing content analysis on policy issues [32]. This codebook was later used by computational linguists to annotate a corpus of policy issues to better understand framing as an aspect of linguistic communication [41]. Following their footsteps, we develop our annotation scheme for characterizing hate group communication. Our aim is to do frame categorization for hate group communications. That is, we want to do for hate group communications what policy frames codebook has done for policy communications. I argue that analyzing a complex problem such as online hate calls for a holistic approach where researchers consider the issue both from the in-group perspective as well as the strategies they use. Specifically we look how the hate groups identify the problem, suggest a solution, motivate their supporters and what approach they select in their communication on social media using collective action frames.

3.1.3.1 Collective Action Frames

Collective action frames refer to the framing activity of interpreting relevant events to motivate followers and disarm the target group [23]. They are characterized by three core framing tasks: *Diagnostic, Prognostic,* and *Motivational* framing [23, 273]. *Diagnostic* framing focuses on assigning blame and responsibility for a problematic situation that the organization is trying to change. For example, the following tweet by an anti-immigration hate group identifies immigrants as a problem for U.S citizens not having high-paying job opportunities.

> *Americans were not given the opportunity to fill relatively high paying temporary jobs because of the illegal actions of a company determined to hire foreign workers.*

Prognostic framing, on the other hand, reflects on an approach for solving an identified problem. Here, solutions are proposed in order to effect desired change. To illustrate, the same anti-immigration group from the previous example suggests a solution for the immigration problem. They propose a change in the Deferred Action for Childhood Arrivals (DACA) bill to fix problems caused by immigrants.

> *All the walls in the world will not prevent illegal entry if these crucial fixes are not a part of any immigration deal over DACA*

The third core framing task, *motivational framing*, elaborates the justification of prognostic and diagnostic frames with the goal of catalyzing action. The following tweet illustrates the use of motivational framing by the same anti-immigration group:

> *If chain migration, the visa lottery, and refugee resettlement continue to dominate Sub-Saharan African immigration, the education level of immigrants will continue to decline relative to the native-born.*

In this particular tweet, the group tries to emphasize the seriousness of the immigration situation by highlighting the differences in education levels of immigrants and native born people.

Several studies benefit from the collective action perspective for analyzing social events. Moussa et. al. analyze the role of collective action framing in social movement mobilization during Gaza war. They find that collective action framing was directly associated with direct action type mobilization [173] As a specialized SMO, hate groups tend to promote active participation in problem solving. They not only offer solutions, they also provide motivations for following this solution The nature of such solutions generally depend on how the hate groups view what they defined as the problematic identity. That is, the relationship between in-group and out-group influences the prescribed solution. From this perspective, the three framing cores of collective action frames can be crucial in understanding how hate groups frame communication on social media.

3.1.4 Hate movements and information mobilization

Prior work investigating hate group activities have primarily focused on examining individual hate websites run by these groups. For example, scholars have studied how various hate groups share news, blogs, and opinion pieces on their dedicated standalone websites [47, 219, 282]. Research on link sharing across extremist blogs have reported that information communities exist across various hate ideologies [282]. Another study found that nearly 72% of the hate group websites contain links to other extremist blogs and sites to order extremist products online [219]. While these efforts provide valuable context of hate groups' operations in the pre-social media era, current trends show that hate groups extensively use social media to disseminate information and spread their hateful agenda. However, the information produced on hate group websites get circulated on internet by thousands of social media participants. Yet, we know little about the roles played by these participants in advancing hate movements.

To understand the participation in information mobilization in hate movements, I next explore the concepts of participatory activism in hate movements and how such activism can lead to the success of hate movements.

3.1.5 Participatory activism in information mobilization

Activism means taking an action to effect social change [179]. Participatory activism is a kind of activism that is grounded in communities organizing to increase popular support for sociopolitical issues by strategically engaging in both, online and offline mechanisms for participation [211]. Considering the expanse of social media, researchers have connected participatory activism to the concept of "smart mobs"—people who are able to act in concert even if they don't know each other [209, 211]. Under this new age activism, diverse set of people with overlapping interests can come together in solidarity and act without having to acknowledge who they are beyond what they support [211].

While studies demonstrate how participatory activism can empower populations in social justice causes on one hand [8, 135, 180, 181, 205**?**], on the other, anti-social movements, for example those advocating for terrorism and extremism, can also benefit from similar practices. One qualitative study emphasized the success of hashtag campaigns and information manifestos in ISIS's territorial expansion in 2014 [144]. They argued that through numerous online accounts, or "media operatives," ISIS was simultaneously able to promote its positive self-image for recruiting combatants and elicit participation from distant supporters. In other words, the "media operatives" played crucial roles in spreading diverse sets of narratives to enable recruitment into ISIS. Moreover, they did so by being a part of strategic information campaigns [144]. For example, ISIS media operatives urged their followers to download videos containing ISIS propaganda and re-upload them on various platforms so as to increase information dissemination while also evading content moderation [144].

Based on these studies, it is clear that similar networking and information sharing affordances contribute to both, positive social changes and anti-social movements. Yet, studies investigating the darker side of participatory activism are rare. In this research, we fill this gap by asking: What are the different roles played by extremist accounts in extremist social movements? How stable or transitory are these roles? And how influential are these roles in spreading mis- and disinformation? We specifically focus on U.S domestic extremism, such as white supremacy and anti-LGBT movements and

identify various roles through the lens of social movement and resource mobilization theories.

3.1.6 Success of social movements through information mobilization

Researchers attribute the success of social movements to the availability of human, material, and monetary resources [161]. Human resources include labor, experience, skills, and expertise of the members of the social movement. Material resources include property, office spaces supplies and monetary resources include funds contributed by the members. In other words, by participating in the movement, individuals make their human, material and monetary resources available to the movement [161]. These resources are are then directed towards mobilizing supporters, transforming mass public into movement sympathizers and eventually bringing about the desired social change. However, the efficiency with which the resources are translated into action depends on various actors involved, such as volunteers offering human resources or supporters offering monetary resources through fundraising. Based on how various members are involved in the movement, researchers have proposed theoretical roles in social movements participation. Next, I summarize the theoretically proposed roles in social movements and discuss the challenges in adapting them to the online setting.

Theoretically Identified Roles in Social Movement Participation: Prior scholarly work has described participants in social movements from various theoretical perspectives, such as participant's role in an organizational hierarchy (e.g., leaders and followers) [171], their involvement in resource mobilization (e.g., members who distribute resources versus members who consume resources) [161] and whether they benefit from the social movement (e.g. stakeholders in the movement) [80, 161, 186, 267]. In organizational hierarchy, *leaders* are strategic decision-makers who mobilize *followers* [171]. Considering participants' involvement in resource mobilization, *constituents* provide resources to mobilize *adherents* [161] and gain sympathy from *bystanders* [251]. Moreover, scholars dichotomize the stakeholders of social movements into *potential beneficiaries*: population that directly benefit from the goals of social movement and *conscience participants*: supporters who may not directly gain from the success of the movement [80, 161, 186, 267]. How do these theoretically identified roles describe participation on online social media? Can we directly adopt the theoretical taxonomy (e.g., *leaders*, *followers*, *constituents* etc.) to describes roles in online extremist movements?

We identified two concerns with directly adopting a theoretical taxonomy, that consequently motivate the methodologies of our first research question. First, the roles derived from theories of social movements are based on physical social movement participation, such as protest events [186, 267]. Compared to online participation, physical participation requires increased commitment by members, such as on-the-ground physical presence at the protest events and significant time investment in the movement's activities [211]. On the contrary, in online activism, individuals can become a part of the movement by simply clicking, re-posting or writing short messages on relevant links [211]. Secondly, theoretically identified roles are harder to disambiguate from each other. Consider for example *leaders* defined as per the organizational hierarchy perspective and *constituents* from the resource mobilization perspective. While *leaders* lead the followers, they can also be *constituents*—the distributors of resources. Similarly, both *adherents* and *bystanders* from resource mobilization perspective can be viewed as *followers*. Considering these two points, instead of directly adopting theoretical taxonomy of roles, we consider the underlying characteristics of participation. Based on the framework of features derived from the theories of social movement participation, we develop a new taxonomy of roles for the online setting. Next, we detail the underlying characteristics that form a background for our role identification process.

3.1.6.1 Characteristics of Social Movement Participation

The theories in social movement participation can be summarized across three dimensions: *drives for participation*, *engagement in the movement*, *strategies of mobilization*. These three dimensions, and the computational features derived from them, are at the crux of our methodology for identifying roles in the extremist movements. The first two columns in Table 3.3 summarize the models described in the next subsections.

Drives for Social Movement Participation: Theories of drives behind social movement participation can be broadly distilled in two groups—expectancy-value models and social-psychological models [87]. The expectancy-value models stress on the rationality of the participants whereby the individuals weigh costs and benefits of participation while decision making [139]. Particularly, they consider the risk-reward ratio of the involvement in the social movement before investing and mobilizing resources for their own benefit [159, 182]. Scholars pointed out one limitation of expectancy-value model that it underestimates the role of ideological drives and shared grievances in participants [139]. To fill this gap, social-psychological models attribute the movement

participation to various psychological drives. Feelings of injustice, relative deprivation and moral outrage is at the heart of movement participation [262]. While such grievances related to social movement issues are ubiquitous, not all of them turn into protests [261]. Hence, social movement scholars also consider the sense of efficacy or achievement that drives people to participate [261]. For example, people might participate because they believe that their collective action can actually achieve the social change [101]. Further, researchers also stress the importance of group identity. The more people identify with the social groups involved in the movement, the more they are inclined towards participating in the movement [227]. Emotions also play an important role in driving people into the social movements. For example, anger is considered as *the* prototypical emotion for protests [263].

Engagement Trends in Social Movements: Various types of participants could adopt different degrees of engagement and commitment to distributing social movement related resources. Specifically based on the availability of resources, *constituents* actively distribute the resources in order to proselytize the *adherents* and bystanders [161]. Moreover, actors could participate in resource mobilization for a single or multiple social movements [161]. Another important aspect of the engagement is how it varies over time. People participate in the social movements with varying degrees of continuity. Corrigall-Brown characterizes such periodic engagement across four dimensions—persistence, transfer, abeyance, disengagement [56]. This characterization of participation trajectories is especially important on social media because it accrues diverse group of individuals with varying degrees of interests and dedication.

Strategies of Information Mobilization: Social movements are goal-oriented. Participants in the social movements strategize the distribution and uses of resources to induce collective action and gain support [161]. Specifically, core members of the social movement mobilize resources in order to recruit volunteers, collect funds, spread their agenda, and hold gatherings. Other researchers also find such solicitation strategies to be crucial for financing social movements [35]. While solicitation may directly affect the progress of the movement, researchers also highlight resources that can indirectly create opportunities for collective action. Specifically in the social media setting, expressing opinions, thoughts, and beliefs around political events [257] and reporting events to increase users' knowledge of public issues, political causes, and social movements [70, 257] are considered as key strategies of online activism.

3.1.7 Information Mobilization outside of West: Political Amplification in India

3.1.7.1 Online Political Influence in India

After the 2014 general elections, social media has emerged as an important battleground in Indian politics [3, 126, 127]. In an attempt to reach out to the younger population, political parties have started deploying organized political campaigns on social media [249]. Researchers have studied social media manipulation in India during elections [66, 126, 234], and various civil conflicts such as farmers' protests, [1] COVID-19 crisis [4, 68] and protests against Citizenship Amendment Act (CAA) [79]. [2]

These studies independently observe that during elections or civil unrest events, social media in India was flooded with various influential narratives promoting political propaganda. For example, during the 2014 Indian Parliamentary elections, both contending political parties (BJP and INC) enforced political ads campaigns on YouTube. While INC appealed more to the social identity of their voters, BJP focused on highlighting the candidate profiles [234]. Moreover, the early months of COVID-19, the issues related to the pandemic were used to frame anti-Muslim disinformation and populist narratives [4]. For example, false stories about Muslim mob assaulting COVID doctors in Mumbai gained a lot of traction on Twitter [4]. These disinformation campaigns reflect how political influence narratives in India are closely related to religious fundamentalist attitudes. In fact, an in-depth interview study of supporters of different political parties in India, including the BJP, INC, and Communist party, found that Indian social media users are concerned with an increasing amount of religious fundamentalist appeals and narratives on social media [66].

3.1.7.2 Computational Research on online political amplification

While the presence of influential political rhetoric is evident in Indian social media, computational research exploring coordinated influence operations in India is limited.

[1] The 2020-2021 Indian farmers' protest was a protest against three Farm Bills that were passed by the Parliament of India in September 2020. Protesters often describe the Farm Bills as "anti-farmer laws" and politicians from the opposition say it would leave farmers at the "mercy of corporates". https://en.wikipedia.org/wiki/2020%E2%80%932021_Indian_farmers%27_protest#Farmer_unions'_demands

[2] The Citizenship Amendment Bill, 2016, was designed to amend the Citizenship Act 1955 to recognize specific types of illegal immigrants, segregated by religion and country of origin. It was enacted on December 11, 2019 as the Citizenship (Amendment) Act, 2019 (CAA).https://en.wikipedia.org/wiki/Citizenship_(Amendment)_Act,_2019

Jakesch et. al. study 75 copypasta hashtag manipulation campaigns across 600 WhatsApp groups and political Twitter accounts in India and find evidence of centrally controlled political influence operations that benefit from the voluntary participation by the party-followers [127]. Specifically, they find that the message or trend template texts are shared on WhatsApp groups and are later popularized across Twitter through different accounts. While their study brings to light an important phenomenon of influence operations in India, it also motivates further research into identifying such coordinated message templates at scale across more popular platforms such as Facebook. Moreover, while previous research has primarily focused on the ruling political party BJP, it is also important to understand the prevalence of such coordinated influence operations across multiple political parties in India. Furthermore, despite Facebook being the most popular social media platform in India [3], most prior works on influence operations in India have focused on Twitter [3, 68, 79, 127]. Given the increasing engagement seen by polarizing Indian political Facebook groups [96], it is important to investigate the coordinated influence activity on this platform.

Outside of India, researchers have studied political amplification in the West, by focusing on re-tweet or re-share or co-tweet networks [100] or by considering other posting characteristics, such as posting time, user similarity, user coordination in an ensemble [51, 121, 271]. Most similar to this work is research by Pacheco that investigates White Helmet coordinated influence networks using text similarity based on pattern recognition [187]

3.2 Research Questions

Guided by the background and theories explained above, I now outline the research questions undertaken in this chapter. With each research question, I will explain the gap in the literature that motivates the question, and describe methods and findings in short.

1. **RQ1: How do hate groups frame content across social media platforms?** Considering the availability of multiple online social media platforms, it is important to consider how hate groups might be using multiple online outlets to further their agenda. Researchers [183] found that influence activities of extreme right groups have progressed from being limited to a dedicated website to spanning

[3] https://www.statista.com/statistics/1115648/india-leading-social-media-sites-by-page-traffic/

multiple popular social media platforms, such as Twitter, Facebook, and YouTube. O'Callaghan and colleagues argue that in order to understand extreme right activities, researchers need to move beyond single platform investigation and start considering multiple online data sources. Their argument provides a motivation for this current work to investigate hate group activities across two popular social media platforms—Twitter and Facebook. Superficially, considering the importance of framing in social movement (described in Section 3.1.3) communication, we analyze how hate groups frame content across Facebook and Twitter. By qualitative content analysis we design a code book that encodes hate groups' cross-platform communication across three dimensions: prognostic, diagnostic and motivation. We show how hate groups vary their communication strategies across platforms across different dimensions and comment on how these findings might be connected to radicalization or mass education.

2. **RQ2: What are the different roles played by online accounts in mobilizing hateful information?** While known hate groups do create problematic content relevant to hateful ideologies, the spread of hateful information is not limited to only the accounts on online hate groups. As explained in the background, scholars have referred to this as participatory activism (Section 3.1.5) which is often utilized for antisocial purposes by organizations like ISIS [144]. Do hate movements enjoy the affordances of participatory activism in online spaces? Studies investigating the darker side of participatory activism, focusing on hateful ideologies and movements are rare. This RQ fills this gap by asking: What are the different roles played by online accounts in mobilizing hateful social movements? How stable or transitory are these roles? And how influential are these roles in spreading mis- and disinformation? We specifically focus on U.S domestic hate movements, such as white supremacy and anti-LGBT agenda and identify various roles through the lens of social movement (Section 3.1.1) and resource mobilization theories (Section 3.1.6). Guided by theories of participation in social movements, we explore underlying behaviors of the accounts across three dimensions: drives for participation, engagement trends, and strategies of mobilization. Using features informed from these dimensions and qualitative expert validation, we identify five roles played by extremist accounts in forwarding their social movements: *solicitors*—who solicit participation and funds for the extremist movement, *educators*—accounts that share intellectual content about extremism and prominently share and like extremist content, *flamers*—accounts

that express and incite anger by posting inflammatory content, *motivators*—who are achievement oriented and go-getters of the extremist community and who post information that portrays a positive image of their extremist agenda, and *sympathizers*—accounts that are fringe supporters of the extremist movement who sparingly engage with links from the extremist websites. These results allow us to understand how theories of social movement participation are reflected in online hate movements and where the various roles are located on the trajectories of deeper engagement into extremism.

3. **RQ3: What are the characteristics of coordinated information mobilization in Indian political campaigns?** Online political activism is rapidly becoming a weapon of mass influence on Indian social media. Starting from the Indian general elections in 2014, social media has been used as a campaigning arena in significant political events over the last several years [4, 68, 234]. Especially given the recent allegations [36, 206, 229] of platform manipulation by political parties in India, it is important to invest in computational research that studies online influence operations outside of the West. In fact, a recent study uncovered organized political influence where party supporters received tweet templates through WhatsApp and Google docs and were encouraged to create copypasta campaigns to influence public opinion [127]. One message quoted in their study also suggests that users may be instructed to tweak the messaging template without directly copy-pasting on social media [[127] Pg. 9]. We focus on two recent national events in India: the introduction of the Farm Bills and the Citizenship Amendment Act CAA (explained in detail in the Data section) and curate a cross-platform dataset of Tweets and Facebook group posts. Considering the use of multiple languages along with English on Indian social media, we use multilingual sentence embeddings with subsequent network analysis and identify over 3.8K political amplification campaigns with lexical mutations. By further establishing that nearly 34% of the messages in amplification campaigns are unique lexical mutants, our work provides an essential context into the actual expanse of political amplification beyond copypasta. After identifying the amplification campaigns, we further characterize the use of political amplification across various dimensions. Given that most of the previous studies focus only on one right-wing political party BJP or, primarily focus on only one social media platform, Twitter, we analyze amplification across multiple Indian political parties and extend our analysis to Facebook.

3.3 RQ1: Characterizing cross-platform content framing by hate groups

To characterize the cross-platform framing by hate groups, we first build a collective action framework based on Framing theory [23]. Next, we collect Facebook and Twitter data by hate groups within USA and perform qualitative coding using the framework developed before. Finally, we compare results across Facebook and Twitter and present our interpretations of frames across the platforms.

3.3.1 Building a Collective Action Framework

Stage 1: Our frame development process started in January 2018 by first collecting 65 SPLC designated hate group profiles listed in the extremist files web page (as of this writing the number of designated hate groups has gone up to 68), their corresponding Twitter handles, and a random sample of 600 public tweets from their profiles. In the first stage, we aimed to inductively develop theoretical insights about collective action framing processes undertaken by hate groups online. We started with Snow's three collective action frames and extended them through several iterative rounds of inductive and deductive testing. We brainstormed with experts in social science, criminology, and researchers who have conducted ethnographic studies of online hate groups. We adhered to an iterative, multistage process of cycling back and forth with data, framing theory, and emergent themes. Stage 1 resulted in 23 categories spread across three collective action frames.

Stage 2: Next, in order to assess the general applicability our stage 1 annotation scheme, we invited seven undergraduate students, all with background in sociology, to examine another random sample of 250 tweets. Due to the sensitive nature of online hate group material, we conducted two information sessions that involved discussing the meaning of frames along with specific examples. Following the discussions, all seven participants independently applied the initial framework to the random sample of tweets. Finally, we discussed their annotation experiences and received feedback about potentially ambiguous, misrepresented categories and possible new themes in each of the three core collective action frames.

Stage 3: Based on the feedback received in stage 2, we modified, removed, and added a few categories. Next, in order to assess the effect of the changes made, me and a sociology expert annotated another sample of 150 tweets. The disagreements in

Diagnostic	Oppression	In-group complains about being oppressed through violent or repressive action, infringement on their rights or resources, or through indictment or sanctions *"...christian school was unjustly raided...", "forced to abandon biblical principles..."*
	Failure	In-group assesses that the government, the system or other agencies such as media have failed to protect them from the problems caused by the out-group *"...government placing americans in danger..."*
	Immorality	In-group indicates that the out-group demonstrates immorality though unethical, immoral or uncivil behavior or values dissonance. *"...Islam teaches and Muslims practice deception...", "...there is no radical Islam, Islam IS radical!..."*
	Inferiority	In-group believes that the out-group is inherently inferior to them based on the political influence, genetics, or the collective failure of the out-group *"...anti-border liberals are of inferior intellect than pro-enforcement Americans..."*
Prognostic	Violence	In-group promotes violent actions towards the out-group *"...choose to be a dangerous man for Christ, wear your cross-hat..."*
	Hatred	In-group advocates protests, criticism or the show of disdain towards the out-group *"...don't take feminism or the women who support it seriously. She thinks being an obnoxious bitch with a chip on her shoulder is empowerment..."*
	Discrimination	In-group promotes avoidance, segregation, or disassociation towards the out-group *"...separation of the races is the only perfect preventive of amalgamation"*
	Policy	In-group suggests formal or hypothetical legislation, promotes political party candidates, or other legal measures that would negatively affect the out-group *"...1.Mandatory E-Verify for all the workers hired 2.No federal funding for jurisdictions/entities blocking ICE..."*
	Membership	In-group demands active association, participation in events or funds towards solving the problem *"...join us at DC rally in support and solidarity...", "...stand with us Americans!..."*
Motivation	Fear	In-group emphasizes on severity and urgency of the problem by mentioning existential or infringement threats *"...There is no way mumps is not being spread outside ICE facilities...", "...Muslim Terrorists are being released in May. Will there be risks to the public?"*
	Efficacy	In-group emphasizes the effectiveness of the action or the solution proposed at the individual or organizational level *"...Major pro-family victory!!! Washington MassResistance strategically helped to stop terrible comprehensive sex ed bill"*
	Moral	In-group discusses the moral responsibility of the audience for taking the action suggested *"...survival of people. That is the mission that matters the most..."*
	Status	In-group discusses increased privilege, social class or benefit from being associated with the in-group or by following the solution provided *"...Our people are destined to have a prosperous future, but only by bearing fruits worthy of repentance..."*

Table 3.1: Table listing the categories in diagnostic, prognostic and motivation frames.

annotations were resolved through discussion until consensus was reached. Combined efforts in the three stages helped us arrive at the point of saturation—as in, the analysis of additional data would not yield any new significant theoretical insights. Our final annotation scheme with 13 categories spread across 3 collective action frames is described next.

3.3.1.1 Hate Frames: The Collective Action for Hate Groups

Table 3.1 contains the definitions for every category alongwith the example text snippets from various tweets and Facebook posts in the dataset.

Diagnosis categories: We identify four ways in which the hate groups diagnose the situation. While assessing how the situation affects the in-group, they complain about being *oppressed* by their targets and other parts of the society, or claim that larger systems such as government and media are *failing* to either correct the problematic situation or protect the in-group from the out-group. Hate groups also diagnose the situation by assigning negative attributes to their targets. They describe the out-group as *immoral* or *inferior* based on the out-group's perceived moral, political or biological standing (e.g, referring to homosexual people as sinners or claiming that some races are genetically inferior to others).

Prognosis categories: For prognostic frames, we summarized five types of solutions proposed by hate groups towards changing the problematic situation. We categorize solutions as advocating *violence*, where hate groups promote violent actions or displays of violence against the out-group, *hatred*, where they encourage others to criticize the out-group, and *discrimination*, where they advocate for avoidance or social segregation

of the out-group. Further, hate groups also call for *policy* changes (e.g, immigration reforms) and direct associations by *membership* requests (e.g, participation in rallies, online events and meetings).

Motivation categories: Our motivation frame has four categories: *fear*, *efficacy*, *moral*, *status*. While *fear* provides a negative motivation by insinuating existential threats to the in-group, the remaining categories use positive aspects such as efficiency of the solution provided, moral high-ground, or status associated with the out-group to motivate the audience.

Next, we use this framing scheme to analyze messages posted by hate groups across Facebook and Twitter.

3.3.2 Data Collection and Annotation

3.3.2.1 Data Collection and Preparation:

We start by the hate group data published by SPLC in their *Hate Map* web page [43]. This data contains a list of 367 hate groups along with their ideologies. Next, we want to identify the social media accounts of hate groups across platforms. Previous studies have used computational matching techniques to identify cross-platform accounts of the same user. Examples include matching computationally matching user's friend network or links shared by them [225, 281]. Such methods have a high likelihood of false matches. Hence, we decided to manually identify and verify the accounts for each of the 367 organizations as follows. First, we conducted web searches with the organization's name to find their corresponding website. In most cases, the website had direct links to their social media accounts. In other cases, we searched the organization's name within the search interface of a social media platforms. We checked whether an account with similar name exists and whether the account's bio had a reference to the organization's website. For every organization, we searched for their Twitter, Facebook, YouTube, Gab, Instagram and Pinterest account profiles. Majority of the organizations had a public Facebook page and a Twitter handle—a total of 75 organizations representing five extremist ideologies with accounts. This dictated our choice of using Facebook and Twitter for our cross-platform analyses. By gathering public tweets and posts from public Facebook profile pages of these accounts between 31st March, 2019 to 1st July 2019, we obtained three months of hate group activities. While all 75 accounts had some activity on Facebook and Twitter, a handful had marginally more messages in one of the platforms. For example, one hate

group had 73 Facebook posts and 2,323 tweets. We removed three such accounts and ended up with a dataset of 16,963 tweets and 14,642 Facebook messages.

3.3.2.2 Data Annotation:

How do hate groups frame the content on Facebook and Twitter? How do they voice opinions and promote narratives in their own words? To understand this, we annotate 1,440 Facebook posts and 1,440 tweets (approximately 10% of the dataset) using categories from our developed annotation scheme (Table 3.1). With 72 accounts in each social media platform, we randomly sampled 20 messages from every account. While most of the accounts had more than 20 messages in the dataset, some had less. To make up for the deficit, we again randomly sampled remaining messages from the remaining Facebook and Twitter data. While annotating, account names were removed in order to reduce the annotator bias.

3.3.3 Result: Cross-Platform Content Framing

We summarize the annotation results in Figures 3.1 to 3.3 using Sankey diagram representation. The width of path between any ideology i and a subframe f is proportional to the number of messages from accounts belonging to i containing subframe s. For example, in Figure 3.3, the path width between "anti-LGBT FB" (Facebook anti-LGBT groups) and *membership* subframe is wider than the *policy* subframe. This suggests that anti-LGBT groups use Facebook to post higher proportion of messages containing "calls for membership" in the hate group in comparison to "demands for policy changes." Below we discuss every main frame in detail and comment about the overall differences in frames across platforms.

3.3.3.1 How do hate groups diagnose the problem?

Diagnosis categories represent how hate groups provide attribution to the problematic situation. Figure 3.1 represents how diagnostic categories are present across the two platforms. On Facebook, *oppression* and *failure* are more popularly used than in Twitter (*oppression*: 22% vs 14% and *failure*: 15% vs 8%). On the other hand *immorality* category is more commonly used on Twitter to educate the audience about negative stereotypes associated with the out-groups (27% vs 19%). Studies show that derogating the out-group via *immorality* frames can also help reinforce the hate group's identity [166].

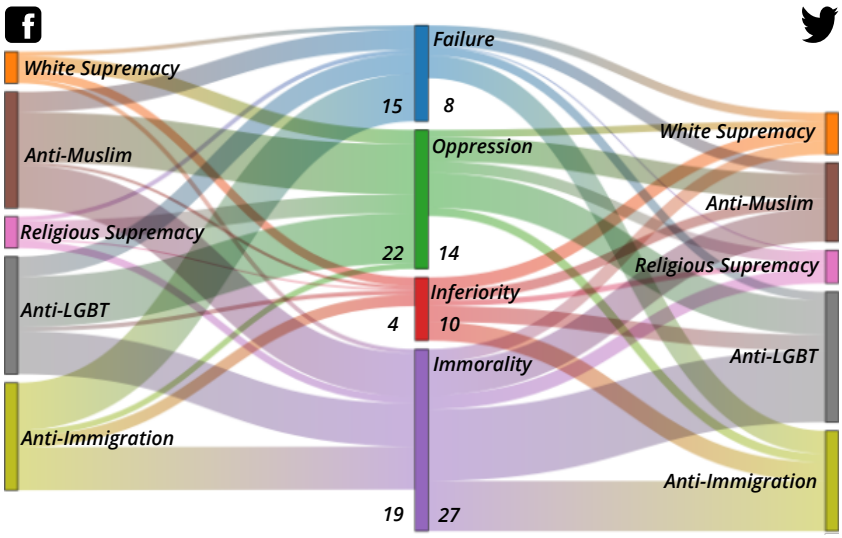


Figure 3.1: Distribution of diagnostic frame categories across Facebook (left) and Twitter (right). Overall, The categories of *oppression* and *failure* are used more on Facebook compared to Twitter whereas *immorality* and *inferiority* are more frequent on Twitter.

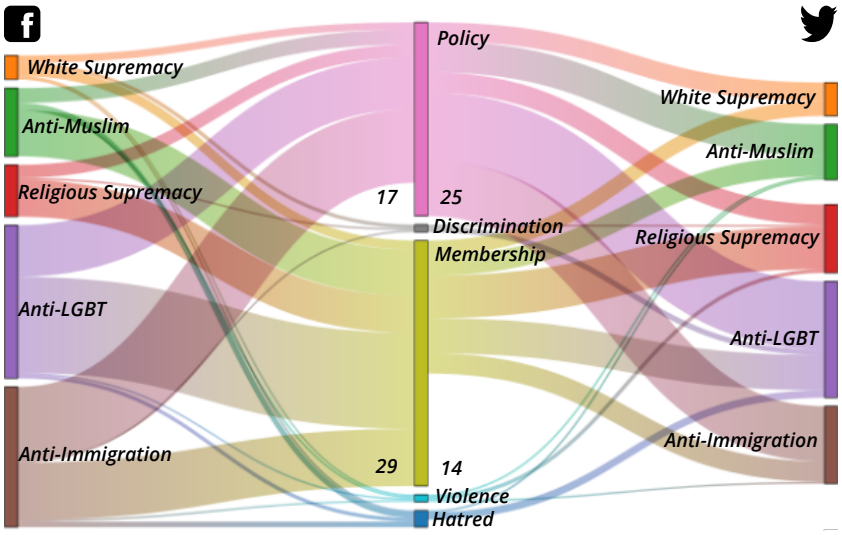


Figure 3.2: Distribution of prognostic frames. *membership* is the most prominent solution offered on Facebook (29%), while on Twitter demands for *policy* change is the predominant (25%).

3.3.3.2 What prognostic do hate group offer?

Figure 3.2 indicates how solutions of *policy* change, *membership*, *hatred*, *discrimination* and *violence* are offered across Facebook and Twitter. Advocating for *hatred*, *violence* or *discrimination* is more extreme and is more likely to get reported because it often involves the use of extreme language. Thus, it is not surprising that on both, Facebook

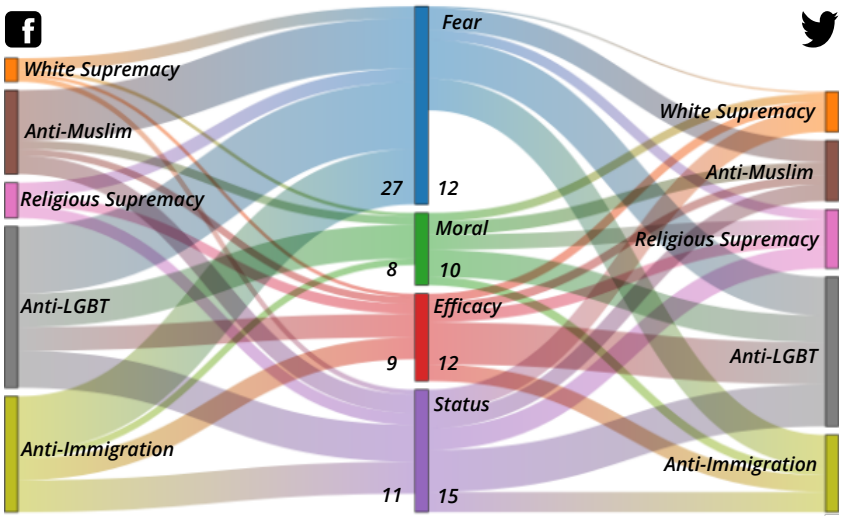


Figure 3.3: Distribution of motivation frames across Facebook and Twitter. *Fear* is a popular motivating agent on Facebook. On the other hand, on Twitter, messages contain more positive motivation such as *status* enhancement, *moral* propriety and *efficacy*.

and Twitter *hatred*, *violence* and *discrimination* subframes are less common. Looking at the frequent use of *policy* and *membership* subframes, we find that *policy* is commonly used across Twitter in comparison to Facebook (25% vs. 17%). Policies can range anywhere from demanding a general political action from the President to signing specific petitions. *Membership*, however, involves calls for direct association with the in-group. Facebook has relatively more *membership* calls compared to Twitter (29% vs. 14%), asking the audience to join events, meetings, and web conferences organized by the group.

3.3.3.3 How do hate groups motivate their audience?

Fear is the most prominent motivating agent found on Facebook (27%) (see Figure 3.3) followed by *status* enhancement (11%). *Fear* appeals are commonly used to motivate like-minded audience using existential threats [166]. *Fear* provides negative incentive to follow the solution. Whereas, *moral*, *status* enhancement, and *efficacy*, all offer positive motivation. Particularly messages with *status* enhancement and *efficacy* attempt to maintain positive self image of the in-group. We find that more messages on Twitter contain *status* enhancement category compared to Facebook (19% vs 11%). Further, other positive motivators (*efficacy* and *moral*) are also more frequent on Twitter compared to Facebook. Previous research suggests that hate groups often strategi-

cally construct messages with self-valorizing views in order to strengthen their group identity [78].

3.3.4 Characterizing Cross-Platform Information Sharing by Hate Groups

How do hate groups use social media to share links to external websites? To answer, we first extract the URL links from messages, expand any shortened URLs to obtain the full domain names, categorize domains by type and understand what types of domains are shared across-platforms. Facebook posts and tweets often contain links to other posts and tweets. Thus, we remove links containing self-referential links. We end up with 12,290 links from Twitter and 11,926 links from Facebook comprising 1,021 distinct domains.

3.3.4.1 Domain Networks

Previous studies have utilized "domain networks" to understand the ecosystem of alternative news domains on Twitter [236]. A domain network is a graph-based representation of URL domains, where every domain is a node connected based on some pre-determined criteria, such as number of common users and frequency of sharing. We leverage the concept of domain networks and modify it to fit our analysis goals.

We connect two domains (nodes in a graph) with an edge if they are shared by a hate group account. The edge weight represents the number of accounts that share both the domains connected by an edge. We remove all edges with edge weight less than two. Finally for trimming the network graph, we remove all nodes that are shared less than 5 times and those that are connected with less than two other nodes. Next, to understand cross-platform sharing behavior, we color the edges differently based on the platforms they are shared on.

Blue (Twitter) edge: If a pair of domains ($D1$,$D2$) is shared together only on Twitter, the edge between them is blue. This means that no account has co-shared ($D1$,$D2$) on Facebook.

Red (Facebook) Edge Similar to the blue edge, if a pair of domains ($D1$,$D2$) is shared together only on Facebook and never on Twitter, the edge between them is red.

Green (Both) Edge If a pair of domains ($D1$,$D2$) is shared together on both platforms, the edge between them is green. For example, if hate accounts $a1$, $a2$, and $a3$ all share domains ($D1$,$D2$) but $a1$ and $a2$ share them only on Twitter, whereas $a3$

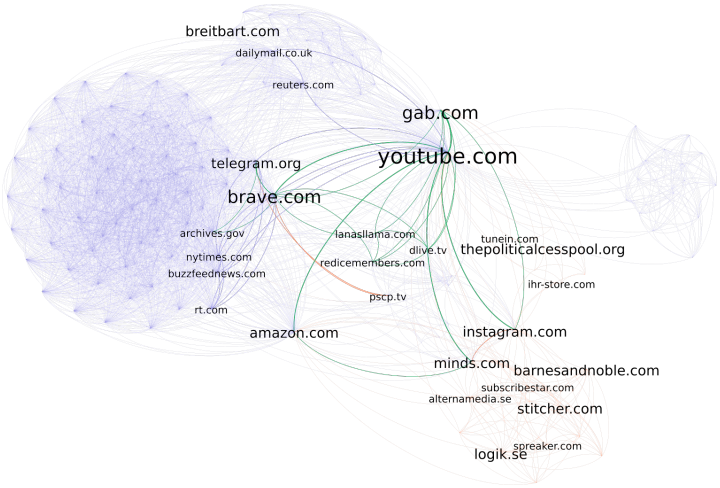


Figure 3.4: Domain co-sharing network in White Supremacy accounts. Blue links represent exclusive co-sharing on Twitter, red on Facebook and green links indicate that the pair of connected domains is shared on both platforms. Domain label size corresponds to the number of times the domain is shared.

shares them on Facebook, then the edge will be green. If the accounts share same information on both Facebook and Twitter, there should be a majority of green edges.

This type of network representation allows us to observe at once, the domains that are co-shared on each social media platform exclusively (surrounded by more blue or red edges) and the domains that are shared in common across both platforms (surrounded by more green edges). Figure 3.4-3.6 show the domain networks for various ideologies.

3.3.5 Information sharing practices by hate groups with different ideologies

3.3.5.1 White Supremacy: Information Sharing

Figure 3.4 displays the domain network for the White Supremacy accounts. Mix of alternative (`rt.com`, `breitbart.com`) and mainstream (`nytimes.com`) news sources are prominently shared on Twitter (57%). Whereas on Facebook we observe more links to promotion domains (35%). Promotion domains consist of various social platforms (`patreon.com`, `subscribestar.com`) (35%). Both Patreon and Subscriberstar have been known to house extreme right wing activists [59]. Interestingly, in promotion domains we also find references to a mix of foreign and U.S websites that host extremist books and literature (`logik.se`, `kirkusreviews.com`) and talk shows

Figure 3.5: Domain co-sharing in Anti-Muslim accounts. Anti-Muslim information sources (`jihadwatch`), blogs (`drrichswier`) and streaming services (`youtube` and `bitchute`) are most commonly shared

(`thepoliticalcesspool.org`). We also observe that domains referring to other social media (`gab`, `telegram`, `bitchute`) are shared commonly across both platforms. We suspect that by diverting followers from Facebook and Twitter to more private and less-censored platforms such as Telegram and Gab, White Supremacy groups might be diversifying their online presence. Particularly in the light of recent censorship of white nationalism on Facebook [84], hate groups might be quickly adapting and moving their online operations in alternative platforms championing free speech.

3.3.5.2 Anti-Muslim: Information Sharing

Figure 3.5 displays the domain network for the Anti-Muslim accounts. Interestingly, there seem to be two main information sources shared by the accounts for anti Islamic news. `jihadwatch.org` appears to be popularly co-shared on both platforms, while `drrichswier.com` is exclusively co-shared on Facebook. `jihadwatch` belongs to one of the Anti-Muslim groups in our dataset. Even though links containing `jihadwatch.com` domain were removed from the account belonging to that hate group, its prominence here suggests that other Anti-Muslim organizations also heavily refer to that domain. Other news websites such as `breitbart`, `foxnews` are shared commonly across both platforms, while Facebook remains a place for sharing informational (43%) websites and blogs (16%). Information sources shared on Facebook often serve as watchdogs for reporting geopolitical issues related to Islam. On the other hand, blogs promote opposition to the fundamentals of Islamic ideologies. This

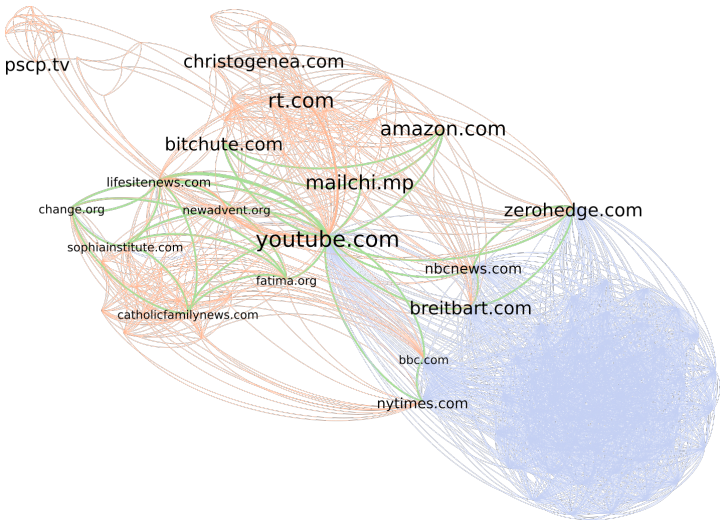


Figure 3.6: Domain co-sharing in Religious Supremacy accounts. Popularly shared domains such as `zerohedge.com` and `rt.com` are labeled as conspiracy promoting sources by `mediabiasfactcheck.com`

suggests that anti-Islamic hate groups might be promoting both religious and political aspects of anti-Islamic hate through Facebook. Further, 7% of Facebook links fall under promotion category with references to other social network domains (`blabber`, `tumblr`) and email marketing (`mailchimp`).

3.3.5.3 Religious Supremacy: Information Sharing

Figure 3.6 represents the domain network for accounts with religious supremacy ideologies. Twitter hosts more news than Facebook (63% vs 28%) (such as `dailymail.co.uk`, `usatoday.com`). Facebook contains links to *opinion* domains (`billshade.org`) (42%). Notably in *promotion* type domains, Facebook hosts a number of platforms used for petitions and donations (`change.org`, `lifepetitions.com`) (11%). While `change.org` observes more diverse user base, `lifepetitions.com` exclusively serves the pro-life and pro-family communities.

3.3.5.4 Anti-LGBT: Information Sharing

Similar to other hate ideologies, Anti-LGBT accounts also share more news on Twitter (54%) compared to Facebook (34%). However, Facebook has more links to blogs (12%) and informational forums (26%) (e.g, resources for parenting (`fatherly.com`, `dadsguidetowdw.com`, `childdevelopmentinfo.com`)) compared to Twitter. Similar

to the Religious Supremacy accounts, we find several websites in promotion category that host petitions (`endbirthdayabortion.com`, `focusonthefamily.com`).

3.3.5.5 Anti-Immigration: Information Sharing

Anti-Immigration accounts share almost 87% of domains exclusively on Twitter with only 9% shared exclusively on Facebook. Both Twitter and Facebook prominently contain news websites and immigration think tanks (`breitbart.com` , `townhall.com`, `cis.org` ,`immigrationreform.com`). In general we observe fewer difference in the types of domains shared across platforms. Anti-Immigration groups tend to dedicate their efforts into raising general public awareness of the social consequences of unauthorized immigration [104]. Previous studies show that greater number of negative immigration related news reports increase perceived level of threat from immigration [221]. Together with the news shared and *fear* appeals on both platforms we believe that Anti-Immigration groups are effectively broadcasting across both platforms to offer influential and educational narratives of hate.

3.4 RQ2: Information mobilization in hate movements through social roles

In this research question, I explore the ecosystem of online accounts through which information flows in the hate movements with the lens of participatory activism—the potential and magnitude of individuals and groups to engage in sociopolitical issues [194]. I focus on extremist movements such as white supremacy, anti-LGBTQ coalition, anti-immigration and anti-Muslim debate and investigate various roles played by online accounts in mobilizing relevant information. I start by explaining the data collection process.

3.4.1 Collecting data for extremist accounts

In this work, we focus on identifying roles in online extremist movements and assess how influential various roles are in spreading information from extremist, biased, fake news and conspiratorial information sources. Specifically, we focus on ***extremist accounts***—public Facebook pages and groups that share links from extremist websites. In this section I detail our process of identifying the extremist accounts and collecting

statistics (per account)	min	max	mean	std.
posts	71	932K	7,067	30,574
link posts	23	78,571	1,614	3,915
extremist link posts	10	5,129	207	528
engagement				
page likes	106	1.8M	7,241	36,576
group members	35	2.2M	2,827	13,603

Table 3.2: Descriptive statistics for extremist accounts in our dataset. There are a total of 4,876 extremist accounts in our dataset.

their Facebook activity data. I first explain the problem of extremism on Facebook and then describe the data collection and pre-processing steps.

3.4.1.1 Extremism on Facebook Groups and Pages

Despite the policies against extremist content, Facebook is still crowded with groups and pages circulating and discussing violent ideas [136]. Very recently, Facebook banned a number of pages and groups involved in the "boogaloo movement"—an anti-government movement by right wing extremists organizing for armed revolt [136, 204]. However, this ban was not as effective as initially perceived. The pages and groups related to the boogaloo movement simply renamed and rebranded themselves with unassuming names and continued their extremist activities [204]. According to the International Association of Chiefs of Police—a non-profit organization that is the world's largest professional association for police leaders—Facebook groups and pages are at the center of extremist recruitment, radicalization, and mobilization [123]. They found that, by just re-posting and linking information from websites hosting graphic videos and other violent content, extremism related groups and pages are able to abide by Facebook's policies against hate speech and still spread relevant information. In our first research question, we identify the roles played by various such extremist accounts in advancing the extremist movements online. In order to model these roles, we first need to identify the extremist websites, and then select the extremist accounts—Facebook pages/groups that share links from the extremist websites.

3.4.1.2 Identifying Extremist Websites

To identify extremist websites, we take help of the resources published by external organizations who are experts in social justice causes. Specifically, we refer to

Southern Poverty Law Center's (SPLC) website [4]. SPLC is a non-profit organization engaged in legal advocacy for social justice issues. Every year, SPLC releases a extremist groups' dataset [5] that records the names, locations, and ideologies of extremist groups operating in the United States. We searched through SPLC's 2018 and 2019 extremist groups datasets. For each listed extremist group, we manually searched for their official website. Note that each listed extremist group has physical headquarters across various states in the USA. By searching for their websites we identify the home for the extremist groups' content in the online world. We identified 289 websites hosted by the listed extremist groups. For each website, we also note the website domain—hereafter referred to, as extremist domains. For example, for Virginia Dare, a white supremacy group, we record `vdare.com` as a website domain and for Alliance Defending Freedom—anti-LGBTQ advocacy organization—we record `adflegal.org`.

According to the SPLC's policies, SPLC prioritizes identifying all U.S based hate groups regardless of the group's "left" or "right" political leaning [6]. For example, the 2019 SPLC dataset contains 27 groups with Black Separatist ideology which is not on the far-right of the U.S political spectrum. However, we could identify websites for only 10 out of 27 Black Separatist groups. While we did not set out to study online extremist groups only from the far-right political spectrum, our collection of extremist websites largely belong to far-right groups representing anti-Immigration, anti-Muslim, White Supremacy and anti-LGBTQ ideologies. This skew towards far-right extremism is representative of the picture of domestic extremism in the United States. According to the Center for Strategic and International Studies (CSIS), [7] far-right extremism has massively outpaced far-left and other types of extremism in the United States [130]. Another independent report on Global Terrorism Index produced by the Institute for Economics and Peace reports that the far-right attacks in In North America, Western Europe, and Oceania have increased by 250% since 2014, making it more lethal than the far-left extremism [92]. While we believe that our dataset is reflective of the ecosystem of domestic extremism in the United States, we discuss this skew further with respect to extremism in other countries in the Limitations and Future Directions section.

[4] `https://www.splcenter.org/`

[5] see "DOWNLOAD DATA" in `https://www.splcenter.org/hate-map`

[6] `https://www.splcenter.org/20200318/frequently-asked-questions-about-hate-groups`

[7] CSIS conducts policy studies and strategic analyses of political, economic and security issues throughout the world. CSIS is labeled as least biased and highly factual source of information by mediabiasfactcheck.com

3.4.1.3 Identifying Extremist Accounts

To identify extremist accounts—Facebook groups/pages that share links from extremist websites—we use the CrowdTangle Link Search API [8]. The CrowdTangle Link Search API retrieves posts by public Facebook groups and pages containing a certain link or the link domain. With the 289 identified extremist domains, we query the Link Search API separately for every domain. For extremist websites containing generic domains such as `sites.google.com`, we query the full link with the sub-domain, for example, `sites.google.com/site/newblackliberationinstitute`. For every queried domain, the API returns up to 1000 posts containing that link domain. Hence, to increase the completeness of our data, we queried every extremist domain separately for every calendar month starting from January 2018 to December 2019. For every queried domain in any calendar month, the number of returned posts was always less than 1000. This indicates that we have retrieved all public posts on Facebook available to CrowdTangle between 2018 and 2019 that share links from the identified extremist domains. Every post retrieved from CrowdTangle contains the account (page or group) name, post text, embedded links, timestamp, reactions (e.g., Like, Love, HaHa etc.), number of comments, number of public shares and 12 other fields. We aggregate all returned posts (450K posts) and find that our data contains 71,430 unique Facebook pages/groups accounts. In other words, 71,430 accounts shared at least one link from the extremist domain. User activities on social media often follow skewed, long-tailed distributions where most users contribute less frequently while fewer users are more active. We observe similar distribution with extremist links posted per account. Previously, researchers have used activity thresholds to eliminate accounts or communities that are less active [112, 147]. For example, while studying Wikipedia edits, Kumar et. al. remove the users that make less than 5 edits [147]. We decide the activity threshold by analyzing the percentile values of the extremist links per account distribution. Based on the 95^{th} percentile cut-off, we remove all accounts that share less than 10 unique links from extremist domains. The entire data collection process happened across three weeks in May 2020. This means that the extremist accounts that were active in 2018 and 2019 but got banned before May 2020 are not included in the dataset [9]. Finally, we have 4,876 Facebook pages/groups remaining in our dataset. Table 3.2 displays the descriptive statistics for the accounts in our dataset.

[8] `https://github.com/CrowdTangle/API/wiki`

[9] As of January 2021, 207 extremist accounts from our dataset (156 Facebook groups and 51 Facebook pages) have been removed from Facebook.

3.4.1.4 Qualitative Validation of the Extremist Accounts

While we know that the extremist accounts posted 10 or more unique links from the extremist websites, do they promote extremist worldviews in general? In this subsection, we present our qualitative validation of the ideologies and the views promoted by the extremist accounts. To validate, we invited two experts from the Southern Poverty Law Center specializing in white supremacy and anti-LGBTQ hate groups. We requested the experts to qualitatively analyze a random sample of extremist accounts. Specifically, we randomly sampled 20 extremist accounts that share links from the white supremacy and 20 accounts that share links from anti-LGBTQ extremist websites. Next, we asked the experts to review Facebook timelines of the extremist accounts and describe their ideologies based on the content hosted and the page/group name and description. 16 of the 20 accounts posting links from white supremacy extremist websites, generally promote either far-right conspiratorial views and racists or misinformative content. Similarly, 18 of the 20 accounts posting links from anti-LGBTQ extremist websites usually peddle anti-choice, anti same sex marriage or anti-trans views. Only 2 of the 40 groups focused exclusively on memes without actively promoting any aspect of the white nationalist or anti-LGBTQ rhetoric. The complete expert analysis is available at the link in the footnote [10].

While the experts validated the extremist views and the content hosted by extremist accounts, the 4,876 extremist accounts in our dataset consist of Facebook pages and groups that engage multiple Facebook users. We treat every page or a group as one extremist account that hosts content posted by it's page owners or the group members. Who contributes to the content on these pages or groups? CrowdTangle, or any other Facebook API does not disclose personally identifiable information even in the public posts [11] [12]. In other words, when requesting the content's of a post, the response will not include the name of the member who created the post. This poses a challenge in understanding the agency of posts shared on the extremist accounts. While this is a limitation of the Facebook dataset, we approximate the engagement with extremist accounts by reporting the distribution of page likes and group members in Table 3.2.

[10] https://www.dropbox.com/s/mf3jt9xbk9z9xfw/SPLC_annotations.pdf?dl=0

[11] https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking

[12] https://developers.facebook.com/docs/groups-api/

3.4.1.5 Extracting Information Shared by Extremist Accounts

We identified 4,876 extremist accounts. In our third research question, we assess how influential various accounts are in spreading links from biased, fake news and conspiracy domains in addition to the extremist domains. Towards this goal, we need all *link posts*—posts with embedded links—from the extremist accounts and not just the ones originating from extremist domains. Hence, next we extract all link posts made by the extremist accounts between 2018 and 2019 and then group the link posts by the link domain type. First, we use the CrowdTangle Post Search API to acquire all of the Facebook link posts made by the 4,876 extremist accounts. In total, we obtain 223K link posts made by extremist accounts across two years, 2018 and 2019. Next, we extract the links shared in the posts along with the timestamps and the link domains. In total there are 74,314 unique links in our dataset spanning over 1,236 link domains.

3.4.1.6 Data Slicing for the Downstream Analysis:

Our role identification process discussed in next sections is based on the Facebook activity of extremist accounts. Recall that we have two years (2018 and 2019) of activity for each account. Here, we determine the time unit of analysis in which we identify the roles and observe their dynamics. To study the role stability over time, we need to divide the two year timespan into smaller windows and analyze how extremist accounts transition into different roles over successive time windows. We slice the data into four windows of six months—T_1, T_2, T_3, T_4–from January 2018 to December 2019. In the previous analysis, we identify the roles played by extremist accounts in T_1 and here, we track the role stability over T_2, T_3 and T_4.

3.4.2 Identifying Roles in Online Extremist Movements

In this analysis, we identify roles played by extremist accounts in the extremist movements based on theory guided characteristics of social movement participation. Specifically, we operationalize *drives for participation*, *engagement in the movement* and *strategies of mobilization* based on the activity of extremist accounts in the six month time period T_1 (Jan 2018-Jun 2018) and build a feature set (Table 3.3) to identify roles. Next, we represent every extremist account in our dataset with the derived features. To identify roles, we cluster the accounts based on the derived features. Finally, we qualitatively analyze and label each cluster as a role in extremist social movement with the help of experts in social psychology.

Characteristics of Participation	Theoretical Models	References	Behavior	Operationalization
Drives for participation	Expectancy-value models	[139, 159, 182]	Risk	Proportion of LIWC Risk words (e.g., caution, crisis, failure)
		[139, 159, 182]	Reward	Proportion of LIWC Reward words (e.g., benefit, bonus, award)
	Social-psychology models	[262]	Injustice	Proportion of MFD Fairness words (e.g., parity, fair, justice)
		[101, 261]	Achievement	Proportion of LIWC Achievement words (e.g., accomplish, ability, attain)
		[227]	Group Identity	Proportion of LIWC we words (e.g., we, ours, us)
		[263]	Anger	Proportion of LIWC anger words (e.g., resent, argue, angry)
Engagement in the movement	Degrees of participation	[161]	Proportion of links from extremist domains	Ratio of links from extremist domains to total link posts
	Degrees of participation (popularity)	[161]	Likes	Proportion of likes on extremist links to likes on the rest of the link posts
			Shares	Proportion of shares on extremist links to likes on the shares of the link posts
			Comments	Proportion of comments on extremist links to comments on the rest of the link posts
	Trends in participation	[56]	Trend	Trend line fitted on the number of extremist links posts per month
Strategies of mobilization	Opinions	[257]	Expressions of opinions	Proportion of extremist link posts containing opinion patterns (see Table ??)
	Solicitation	[35, 161]	Expressions of solicitation	Proportion of extremist link posts containing solicitation patterns (see Table ??)

Table 3.3: Table summarizing features used to identify roles in online extremist movements on Facebook. We build the feature set based on underlying characteristics of participation and the theoretical models describing them.

3.4.2.1 Operationalizing Characteristics of Social Movement Participation

Drives for Social Movement Participation (6 features). Theoretical models exploring the drives for social movement participation consider two distinct perspectives: *expectancy-value models*, whereby participation is driven by perceived risk-reward assessments and *social psychology models*, where the psychological features are the core drivers of participation. Below we detail the computational operationalizations of *drives* informed by these theoretical models explained in Section 3.1.6.

- **Expectancy-value features:** Participation in social movement could be driven by perceived risk-reward assessment related to engagement in the movement [139, 159, 182]. Analyzing the language used by the extremist accounts while sharing links from extremist websites, especially the words related to cost-benefit, could indicate whether the accounts considered the costs and benefits of participating in extremist movements. To operationalize these features, we use risk and reward lexicons from the Linguistic Inquiry and Word Count (LIWC) 2015 [245]. LIWC is designed to record words that reflect various psychological states and perceptions [245]. Specifically, we calculate the proportion of risk related words (e.g., caution, crisis, failure) and reward related words (e.g., benefit, bonus, award) used by an extremist account while sharing links from extremist websites.

- **Social Psychology features:** Guided by the social-psychology based theoret-

ical models, here we want to measure the feelings of injustice [262] , sense of achievement [101, 261], group identity [227] and anger [263] that could serve as potential drives for participation. To identify the language related to injustice, we use Moral Foundations Dictionary that contains a systematically derived list of words pertaining to moral foundations in political ideologies [244]. Specifically, we use the "fairness" lexicon which accommodates virtue words such as rights and equality, and vice related words such as bigot, favoritism, and prejudice [98]. Next, to measure the sense of achievement, we use LIWC's [245] achievement category which contains words such as "accomplish", "ability", "attain" etc. Further, language related to group identity can be reflected by the use of third person pronouns such as "we", "us", "ours" [129, 191, 245]. Hence we measure the third-person pronoun usage by LIWC "we" category. Finally, to measure anger related words, we use LIWC anger category containing words such as "resent", "argue", "angry." For each of these lexicons, we calculate the proportion of words in each lexicon while sharing links from extremist websites.

Engagement Trends in Social Movements: (5 features). As per the details described in Section 3.1.6, here we provide our methods to characterize engagement trends in the social movements. Participants can engage with social movements in various degrees of interests and continuity [56, 161]. Hence, we calculate proportion, popularity and trends in sharing links from extremist websites.

- **Proportion of Links from Extremist Domains:** Various degrees of participation in distributing resources can reflect the involvement in social movement [161]. Hence, for every extremist account, we first calculate *proportion of links from extremist domains*—proportion of links shared from extremist domains to all links shared by that account in a given time-frame.

- **Popularity of Links from Extremist Domains:** The amount of positive reactions and interactions on the shared links could reflect how popular the posts containing extremist links are on a Facebook page/group. Hence, we calculate the average likes, shares and comments received on posts with links from extremist websites and divide it by the average likes, shares, and comments (respectively) on all links posted in a given time-frame. High values of these features indicate that the extremist content is more popular compared to the rest of the content published on that page/group.

- **Trends in Disseminating Links from Extremist Domains:** We also account for the engagement trends. For each month within the six-month period, we calculate the number of links from extremist websites posted on an extremist account and fit a line via least-square regression. Least-square regression finds optimal fit for the line by minimizing the sum of squared residuals. We calculate the trend of this fitted line to measure engagement. Specifically, positive values of trend can indicate increasing engagement and negative values can indicate disengagement in posting links from extremist websites.

Strategies of Information Mobilization: (2 features): As described in detail in the background section, core members of the social movements may strategically solicit participation through calls for donations, volunteers and invitations for social gatherings [35, 161]. Similarly, members can also strategically create opportunities for collective action by expressing opinions, thoughts and beliefs around political events [257]. Hence, we build two features that capture the expressions of personal opinions and the language of solicitation used by extremist accounts while sharing links from extremist domains. For both features, we calculate the proportion of extremist link posts containing the expressions of opinions and solicitations respectively.

- **Expressions of Opinions in Posts with Links from Extremist Websites:** By "opinions" we refer to the expression of thoughts, beliefs and personal opinions [218]. To calculate the proportion of opinions present in posts, we extract phrases that signal expressions of personal opinions or private states [272]. Previously, researchers studying emotional and informational support [27, 270] relied on emotional and informational support related nouns, verbs and adjectives to extract phrases related to their task. For example, Wang et. al. [270] used $< you + MODALVERB >$ pattern to extract phrases containing suggestions ("you should" or "you must"). We use similar methodology to extract phrases related to personal opinions. How can we identify verbs, nouns and adjectives related to personal opinions? Chen et. al. argue that individuals form their opinions via cognition and internal perceptual cues [49]. Hence, we first look at LIWC 2015 cognitive processing lexicon and its subcategories that record words related to thinking, perception and expression. To construct phrase patterns, we first split all words from LIWC cognitive processing categories into their part of speech labels (verbs, nouns, adjectives). Consider the verb *prefer* and noun *preference* from LIWC cognitive processing category. Both words, when paired

with different pronouns can form expressions of opinions. For example, *"I (first person subjective) prefer"* and *"My (first person possessive) preference"* both indicate personal opinions. Moreover, variations such as negations (*"I do not prefer"*) or adjectives (*"My strong preference"*) also signal opinions. Hence, we build our initial set of opinion patterns from LIWC's cognitive processing verbs, nouns and adjectives by pairing them with appropriate pronouns and variations. We iteratively improve upon this list of phrase patterns by first, extracting sentences containing those patterns and then, manually eliminating verbs, nouns and adjectives that do not signal opinions. Note that our opinion extraction method is based on LIWC cognitive processing lexicon that contains limited number of words. Hence, it is possible that our opinion extraction misses out of some expressions of opinions.

- **Expressions of Solicitation in Posts with Links from Extremist Websites:** In solicitation, we want to identify expressions that demand some action on the reader's part. Here, we are looking for calls for donations, invitations for events and protests and participation in policy advocacy (e.g., sign the petition, call your representative etc). To extract solicitation patterns, we follow similar procedure as opinion extraction but instead look at verbs, nouns and adjectives from LIWC's social and affiliation categories. LIWC social and affiliation categories contain words such as *sign*, *call*, *contact*. We build phrase patterns and iteratively evaluate them using methods similar to opinion extraction.

3.4.3 Clustering Extremist Accounts Based on the Derived Features

We use the features described above, to cluster the extremist accounts and label each cluster as a role in the extremist movement. We use the theory based features (described in the previous subsection) representing drives, engagement and strategies to discriminate between different roles in meaningful way. For role identification, we first need to decide on the number of roles and then use a technique that integrates structural data (feature vectors for extremist accounts) with interpretive analysis that will allow us to describe roles in a relevant way. We use K-Means clustering—an unsupervised clustering algorithm commonly used by other CSCW scholars in role identification studies [7]. For example, Arazy et. al. used K-Means to to identify emergent roles in Wikipedia contributors [7]. We use the popular K-Means method with kmeans++ initialization to cluster the extremist accounts based on their activity

in the six month time window T_1. Specifically, we represent every extremist account with a feature vector of length 13 (6 drives + 5 engagement + 2 strategies). Next, we perform a series of robustness checks to first, determine the number of clusters in order to obtain the optimal separate between different roles and then, to check the stability of clusters. All features were standardized for the downstream analysis. In K-Means algorithm, the number of clusters , K is a free parameter. We first find the best value of K with an elbow analysis that offers a natural trade-off between the best separation between the clusters and the number of clusters. Specifically, we train the K-Means algorithm for number of clusters ranging from 2 to 20 and plot distortions—sum of squared distances from each point to its assigned center. We observe the elbow at $K = 5$. Thus we assume 5 as the optimal number of roles. We repeated the elbow experiment with other scoring parameters such as silhouette distance—a measure of how similar a data point is to its own cluster compared to other clusters—with similar results. We also check for the stability of final cluster assignments with various random seeds. Moreover, to check the robustness of our method, we perform the clustering with alternate clustering methods observing similar elbow and cluster assignments. We assume that every cluster generated, represents a role in the online extremist movement. Next, we use expert guided interpretive analysis to label the roles and their descriptions.

3.4.3.1 Role Labeling with Expert Evaluation

Do our quantitatively identified clusters represent coherent roles in extremist movement participation? To evaluate, we invited a group of seven social psychology and social movement experts to first analyze and then label the clusters based on their representative characteristics. This group consisted of one senior professor, one assistant professor, one post-graduate researcher and four senior doctoral students. To reduce bias in role labeling, we selected the external evaluators who are not a part of the author group and were not involved in any of the work preceding or following this stage. We showed them mean feature values for every cluster. Additionally, we selected top 5 representative extremist accounts from each cluster—accounts with closest distance to the cluster centers. We compiled the list of top 10 most representative posts from each selected account ranked by post likes [120]. Based on this information, we asked the evaluators to come up with labels and descriptions for each cluster. In all, every evaluator looked at 250 Facebook posts (5 clusters × 5 accounts × 10 posts) and recorded the possible labels and descriptions. Finally, the first author and the

Role	Frequency	Example texts used by the accounts while sharing links from extremist websites
Solicitors	5.2%	*"...Sign here to demand her [Rep. Maxine Waters] immediate resignation"* *"...Join us in signing thank you card for President Trump"* *"...Stand with us to take back our streets"*
Educators	10.6%	*"...Escaping from motherhood: how it destroys society"* *"...We believe that we have the duty to instruct people in the truth of Tradition. Even if it destroys their party"* *"...The need for an ethnocentric society amidst "globalism" "*
Flamers	18.4%	*"...genuine Christians know that homosexuality is an abomination before GOD! "* *"...MURDERED in cold blood. Emergency: Gunfire, bodies and BLM murderers"* *"...house democrats vote to allow female genital mutilation..." (false information flag by Facebook pops up)*
Motivators	29.4%	*"...Senator Dan Halls stands with us in a passionate commitment to strengthening religious freedom"* *"...FREE SPEECH WINS!!! Supreme court rules pregnancy centers can't be forced to advertise abortion"* *"...we WILL have the COURAGE to defend ourselves! borders, language and culture MATTER"*
Sympathizers	36.4%	*"...White South Africans petition Trump to allow them to migrate to the US"* *"...A jihadi cult member running for Congress as Democrat from Alaska"* *"...They will only see Italy on postcard: Italy turns away another migrant ship"*

Table 3.4: Roles and the corresponding percent of extremist accounts in the dataset. The examples are of the texts written by the extremist accounts while sharing links from the extremist websites

evaluators together, worked and selected the best label for every cluster. We present the identified roles and comments on the expert evaluation process in the next section.

3.4.4 Results: Roles in Online Extremist Movements

Here I describe the roles played by extremist accounts and their typical behaviors. Table 3.4 displays the frequency of the roles in the dataset alongwith the example text written by the extremist accounts while sharing links from extremist websites. We identified five roles in the online extremist movements.

1. **Solicitors:** These are the accounts that solicit participation from their readers for signing petitions, attending rallies etc. On average, around 20% of their links come extremist domains and they post extremist content with fairly consistent trend throughout the six month period (trend feature values close to 0). These accounts use high group identity language such as "we", "our", "us" compared to other roles. Evaluators also described them as "recruiters." One evaluator mentioned:

   > *"These groups appear to be soliciting action for their hate. To some extent, they seem pretty keen on doing something about the groups they hate and are actively sharing/liking posts to promote action"*

2. **Educators:** Educators have distinctively high amount of extremist content in their link sharing. On an average, 50% of their links come from extremist domains. Additionally, the extremist links posts get more likes and comments compared to

other material on these pages/groups. They post the extremist content with consistently high rates (trend feature > 0) throughout the six months. In qualitative evaluation, the experts pointed out that these groups share intellectual material and appear serious and sincere in propagating the fundamentals of extremist ideologies. The evaluators also suggested alternate labels such as "preachers" and "intellectuals." According to one evaluator:

> *"...they seem to take effort to make logical arguments. They are not necessarily showing anger towards other groups but are instead more focused on highlighting their own group's worth logically/analytically"*

3. **Flamers:** These accounts spew toxic and inflammatory content. Around 5% of their links belong to extremist domains and the messages on the links and the link text itself often contains language suggesting anger and injustice. In other words, these pages/groups have the highest proportion of anger and injustice related words while disseminating extremist content. The extremist links posted on these accounts get higher number of shares compared to the rest of the content. The experts also described them them as "fear mongerers" for attempting to cause general outrage. Immediately after looking through the posts, one evaluator commented:

   > *"these are clearly very strong, divisive and toxic posts"*

4. **Motivators:** Around 7% of the links by motivators are sourced from extremist domains. Evaluators pointed out that motivators use exceptionally positive language. While posting extremist content, they stress on the achievements and rewards associated with extremist activities. Motivators also express opinions with highest proportions compared to the other roles. Experts noted that these accounts engage in policy activism focusing on policies protecting and defending cultural and moral values. Evaluators also mentioned:

   > *"it almost looks like they are celebrating the in-group [people and organization involved in the extremist movement] and the sensationalized news about the in-group"*

5. **Sympathizers:** These accounts post extremist content links with lowest rates (2% of their Facebook link posts) and sporadically throughout the six month

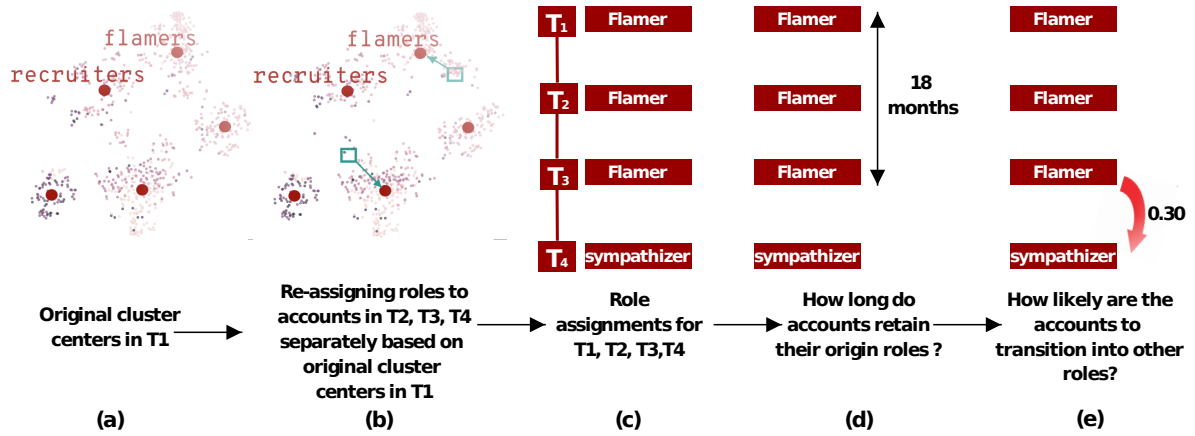


Figure 3.7: Figure showing method steps for analyzing role dynamics. (a) We identified cluster centers for each role in RQ1 using the account activity in T_1 (Jan 2018-Jun 2018). (b) We use those cluster centers identified in T_1 and re-assign extremist accounts to roles based on their activity in T_2, T_3 and T_4. (c) For every account in the dataset, we get cluster assignments for all time periods. (d) To measure role retention, we calculate the number of time periods for which the extremist accounts maintain their original roles from T_1. (e) In role transition, we calculate the probability with which the extremist accounts may transition to different roles.

period. They also show low engagement in terms of likes, shares and comments on the extremist link posts. According to the experts, these groups are on the fringe of extremist ideology and might be only slightly interested in extremist causes. Experts also referred to them as "observers." One evaluator described *sympathizers* as: *"They look more like general conservative interest groups"*

3.4.5 Measuring Role Dynamics

The roles we identify are data-driven. The underlying characteristics of the roles are derived mainly from the theoretical studies on physical protest events. Unlike physical social movements where members can commit to protests, meetings or other events, social media provides a dynamic, evolving space for members to engage or disengage from the social movement as they like without much accountability [56]. Hence, we also analyze the dynamics of roles based on how long the accounts retain their initial roles (role retention) and their transition probability to another role (role transition). Figure 3.7 displays the method steps taken to measure role retention and role transition.

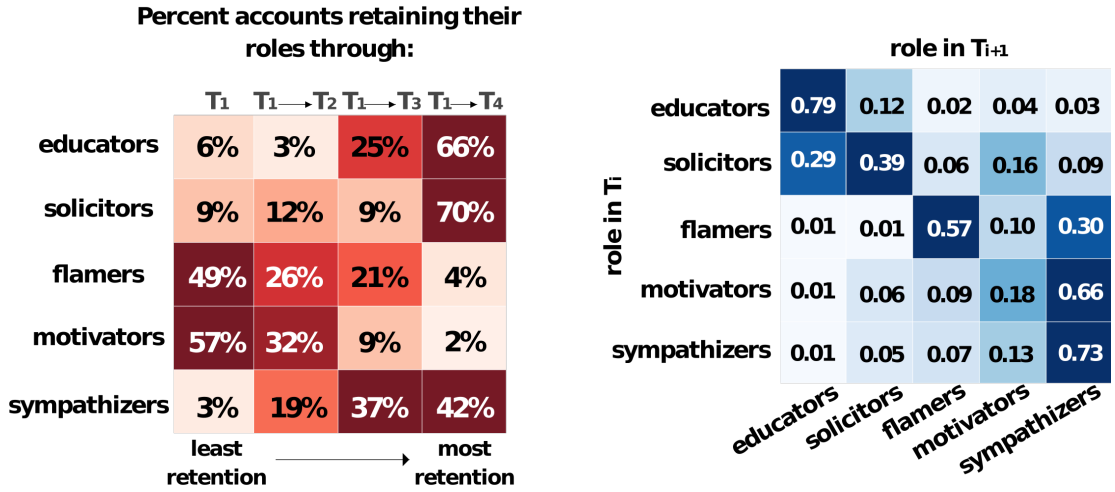
Measuring Role Retention: By retention, we measure how long the extremist accounts adhere to their originally identified roles. Recall that our data spanning two years

(Jan'18 to Dec'19) is sliced into four windows, each six months duration—T_1, T_2, T_3, T_4. We initially identified roles using the account activity in the first time period, T_1 (Jan 2018 - Jun 2018). To measure role retention, we use the cluster centers (mean values for each feature for each cluster) from T_1 and re-assign roles to the extremist accounts using their activity in T_2, T_3, T_4. For every extremist account, we now have the role assignment for T_1, T_2, T_3, and T_4. For example in Figure 3.7 (c), an account stays in the *flamer* role for three consecutive time periods. That is, the role identified in T_1 is retained for (3 X 6) 18 months consecutive months. Note that all time windows T_1, T_2, T_3, and T_4 represent distinct calendar months between 2018 and 2019. Using the initially defined cluster centers in T_1 allows us to compare the accounts' activities in T_2, T_3, T_4 with respect to their own past states. How long do extremist accounts adhere to their initially identified role in T_1 over the subsequent future time windows? To answer, we calculate the number of continuous time windows across which an account retains its initially identified role. Finally, for each role, we calculate the proportion of accounts maintain their roles across just one (T_1), two ($T_1 \rightarrow T_2$), three ($T_1 \rightarrow T_3$) or all four ($T_1 \rightarrow T_4$) time windows.

Measuring Role Transition: How likely are the extremist accounts to move from one role to another? For example, how likely are *sympathizers*—fringe supporters of the extremist movement—to transition to *educators*—accounts that actively distribute large proportion of extremist content? From the role retention analysis, we know which role every extremist account plays in each of the time windows—T_1, T_2, T_3, T_4. For every account, we consider the role played by that account in a particular time window T_i as the *state* S_i that account is in. Consequently, for every account we have sequence of four states corresponding to each of the time windows. The example account in Figure 3.7 (e) has states: *flamer* $\rightarrow$ *flamer* $\rightarrow$ *flamer* $\rightarrow$ *sympathizer*. Using such state sequences of all extremist accounts, we then calculate the state transition, or the role transition probability. Specifically, we calculate the pairwise transition probabilities for each pairs of roles. High probability of transition *solicitor* $\longrightarrow$ *educator* will indicate that an account currently playing the role of *solicitor* is likely to transition into *educator* in the next time window with high probability.

3.4.6 Role Dynamics in Extremist Information Sharing

Retention: Figure 3.8 (a) displays the initial roles (rows) and number of time periods (T_1, T_2, T_3, T_4) for which the role was retained by an account. We find that 66% of



(a) (b)

Figure 3.8: Figure presenting the results of role dynamics analysis. (a) indicates role retention—proportion of extremist accounts that retain their originally identified role through subsequent time windows in the dataset. For example, 70% of the *solicitors* retain their role throughout $T_1 \rightarrow T_4$. Whereas 49% of *flamers* retain their roles for only the initial T_1 period. (b) displays the role transition probability matrix. Rows indicate the role in T_i and columns indicate the role in T_{i+1}. The cells indicate the probability of transition from role in T_i to T_{i+1}. For example, *solicitors* may transition to *educators* in the next time period with 0.29 probability.

the *educators* and 70% of the *solicitors* retain their initial role for all four time periods, that is for the entire two years. Whereas, 49% of the *flamers* and 57% of *motivators* transition to another role just after T_1 (6 months). *Educators* and *solicitors* can be viewed as an elite group in extremist movements—members who distribute (information) resources and actively recruit others [161]. On the other hand, *flamers* and *motivators* are supporters who exhibit low engagement with the links from extremist websites. Our results suggest that roles more core to the extremist movement such as *educators* and *solicitors*) are more stable. In other words, *educators* and *solicitors*) are more likely to maintain their roles and consequently their behaviour surrounding the participation in extremist movements for longer periods compared to others.

Transition: Figure 3.8 (b) displays a transition matrix for roles. The values in the cell indicate the probability by which an account in one role (row) would transition to another (column) in the next six months. Based on the main diagonal, accounts in most role are more likely to retain the same role in the next time window. For example, *educators* will stay *educators* in the next six months with 0.79 probability. Similarly the probability of *flamer* $\longrightarrow$ *flamer* is 0.57. Notably, *solicitors* and *educators*—roles

with highest engagement with links from extremist websites—have highest transition probabilities with each other compared to any other roles (*solicitor* ⟶ *educator* = 0.29 and *educator* ⟶ *solicitor* = 0.12). Moreover, *flamers* and *motivators* can transition to *sympathizers* with 0.30 and 0.66 probabilities respectively, indicating that roles with less engagement with extremist content are also less stable.

3.5 RQ3: Coordinated Political Amplification in India

Increasingly online platforms are becoming popular arenas of political amplification in India. With known instances of pre-organized coordinated operations, researchers are questioning the legitimacy of political expression and its consequences on the democratic processes in India. In this paper, we study an evolved form of political amplification by first identifying and then characterizing political campaigns with lexical mutations. By lexical mutation, we mean content that is reframed, paraphrased, or altered while preserving the same underlying message. Using multilingual embeddings and network analysis, we detect over 3.8K political campaigns with text mutations spanning multiple languages and social media platforms in India. By further assessing the political leanings of accounts repeatedly involved in such amplification campaigns, we contribute a broader understanding of how political amplification is used across various political parties in India. Moreover, our temporal analysis of the largest amplification campaigns suggests that political campaigning can evolve as temporally ordered arguments and counter-arguments between groups with competing political interests. Overall, our work contributes insights into how lexical mutations can be leveraged to bypass the platform manipulation policies and how such competing campaigning can provide an exaggerated sense of political divide on Indian social media.

3.5.1 Collecting cross-platform data

3.5.1.1 Political events and keywords

In this paper, we study the political message amplification campaigns on Indian social media. To contextualize our analysis, we focus on two recent politically divisive events that affected the whole nation and attracted partisan debate on social media. We used keywords and hashtags related to these events to collect data from Facebook and Twitter.

	Keywords	Timeline	#Messages	#accounts
Farmers' protests	FarmersProtest,KisanAndolan,किसान, आंदोलन, किसान, AIKSCC, farmer, farmers, Kisaan,किसान_बिल,बिल,Farmers,Kisaan,kisan,KisanBill,KisanProtest, TractorMarch,FarmerPolitics, "भारत बंद"",bharatbandh,"bharat bandh", कृषि_कानून,कृषि,FarmersBill,BharatBand,भारत_बंद,खेती,farmlaws,Tractor2Twitter, कृषि_कानूनों,FarmLaws2020,FarmersDelhiProtest,फार्मला	Oct 2020 - Dec 2020	98,297	4,593
			701,345	171,345
CAA	CAAProtest,IndianCitizenshipActProtest,"citizenship amendment act",CAA_NRC, ShaheenBagh,शाहीनबाग़, CAA,NRC,नागरिकता,नागरिकता_संशोधन_बिल,शाहीन, नागरिकतासंशोधनकानून,Shaheen,CAB,NRCBill,नागरिकताकानून, CitizenshipAmmendmentAct,CABBill,CitizenshipAmendmentBill, नागरिकता_संशोधन_विधेयक,"नागरिकता संशोधन विधेयक",एनआरसी,सीएए	Nov 2019 - Jan 2020	94,489	4,218
			602, 967	125,827

Figure 3.9: Table describing dataset keywords, timeline, and the number of posts. Note that we collect only original tweets and posts, excluding retweets and reshares.

Famers' protests: Farm acts, passed in the parliament of India in September 2020, sparked nationwide protests by farmers' organizations. Farm unions were primarily protesting the entry of corporations in crop trading facilitated by the farm acts along with legacy issues such as high farmer suicide rates and low agricultural income in India. The protests were highly politicized across the political spectrum in India with the ruling political right (BJP) standing in support of the farmers' bills while the oppositional left (INC, AAP, etc.) aligned with farmers' unions in the protests.
Citizenship Amendment Act (CAA): This amendment to the Citizenship Act was proposed by the Government of India under the leadership of the right-wing BJP party. It offered a pathway to Indian citizenship for persecuted religious minorities from Afghanistan, Bangladesh, and Pakistan who are not Muslims. The amendment was opposed by non-BJP politicians and student organizations, causing polarizing tensions across the nation which led to violent protests and rallies from both sides.

We start building our list of keywords from previous works [67, 68] around the Farmers' protests and CAAs. We aim to make the keyword list more generic to increase the coverage of the dataset, while still ensuring that we capture the relevant data. For example, we strategically include general words like "Kisan" or "Esee" which are transliterations of the words "farmers" and "CAA" and are more likely to be used only in the Indian context. Table 3.9 records the keywords used in collecting the data.

3.5.1.2 Facebook dataset

Facebook groups have been known to play key roles in recruiting supporters for politicians and political parties during elections [276]. In fact, Facebook cyber security expert commented that during the 2019 Parliamentary elections in India, Facebook groups and pages were designed to look independent but were actually linked to political parties trying to conceal their identities [215]. Such Facebook groups were

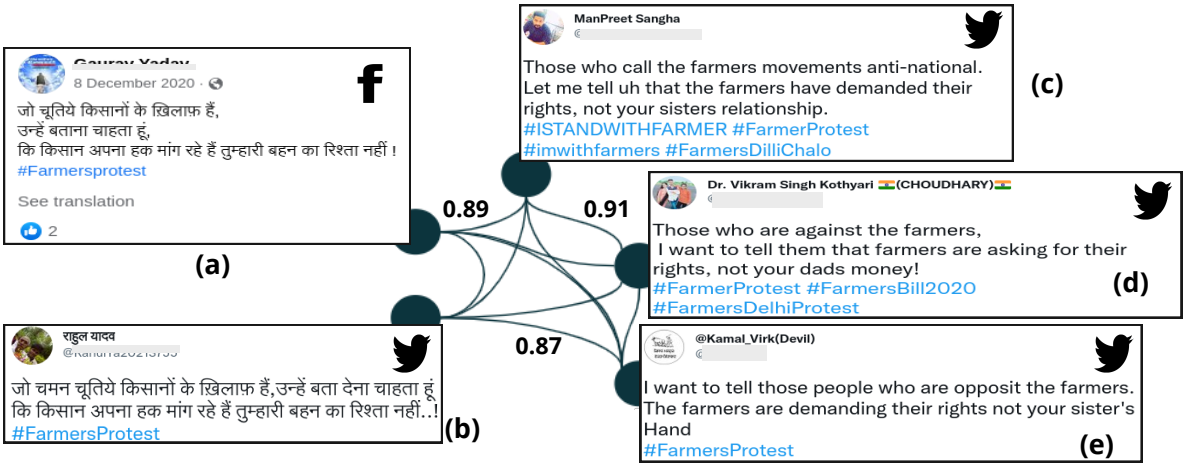


Figure 3.10: Examples of lexical mutants in an amplification campaign. Two messages (nodes) are connected together if they have high cosine similarity. A clique of nodes connected like this represents an amplification campaign with lexical mutants.

also found to be linked with fake or bot accounts that spread misinformation and influential content relevant to elections [96].

We use CrowdTangle's post search API[13] to extract posts made on Facebook pages and groups relevant to the political events described above. We use the list of keywords and hashtags mentioned in Figure 3.9 to collect public posts from Facebook groups and pages. Note that CrowdTangle or any official Facebook API does not offer any authorship information for the posts.

3.5.1.3 Twitter dataset

We also collect tweets relevant to the two political events using the Twitter Academic Research API. Similar to the Facebook dataset, to analyze hidden amplification campaigns, we only preserve the original tweets, excluding retweets and quote tweets. Overall details of the dataset are available in Figure 3.9.

3.5.2 Identifying Political Campaigns with Lexical Mutations

A key to identifying amplification campaigns beyond copypasta is to identify messages that are similar to each other in terms of core content but are not explicit re-posts or re-tweets [187]. Unlike retweets or reposts, messages with similar content are treated as different by the platform and the link between the original and the copy is hard to detect [187]. Consider, for example, the messages in Figure 3.10. All of the messages

[13] https://github.com/CrowdTangle/API/wiki/Search

have similar content with slight lexical variations, posted by different users on different social media platforms. Below, I outline the methodology for identifying groups of messages with lexical mutations across multiple languages on Indian social media.

3.5.2.1 *Extracting multilingual embeddings:*

A usual approach to identifying messages with slight lexical mutations may be to compare the edit distanced [153], message keywords [222] or to use sentence embeddings with cosine similarity [187]. However, posts on Indian social media contain multiple languages. For example, in Figure 3.10, all the messages have similar content but some are in different languages. In this case, simple token-based word embeddings trained in English, or other fuzzy matching methods will not work.

Instead, we use multilingual sentence embeddings to represent texts across multiple Indian languages and also capture the semantic similarity between paraphrased or reframed texts. There are several pre-trained multi-lingual models available, such as sentence-BERT [208], Language-Agnostic SEntence Representations (LASER) [11] and multilingual BERT [72]. LASER was found to outperform multilingual BERT for Hindi text classification [131]. Since a significant portion of our text data is in Hindi, we use the LASER model which works with more than 90 languages containing more than 28 different kinds of alphabets.

3.5.2.2 *Identifying similarity threshold for lexical mutation:*

Similarity between two texts can be calculated with the cosine similarity between the multilingual sentence embeddings described above. A cosine similarity of 1 indicates perfect similarity between the two texts. Examples in Figure 3.10 all contain a similar underlying message with variations in languages and phrases. To extract groups of texts such as this, we need to determine a threshold value of cosine similarity above which we can consider two texts to be lexical mutants of each other. We determine this threshold empirically by manually analyzing pairs of texts with different cosine similarity scores. Specifically, we randomly sampled 20 pairs of sentences for cosine similarity values each, starting from 0.5, with increments of 0.05. We labeled each pair as either 1—to indicate whether a lexical mutation still resulted in preserving the underlying message—or 0. The number of pairs scored as 1 in the sample of 20 naturally kept increasing with the cosine similarity score. All samples with a score of 0.85 were labeled as 1. Hence, we chose 0.85 as the cosine similarity threshold. In other

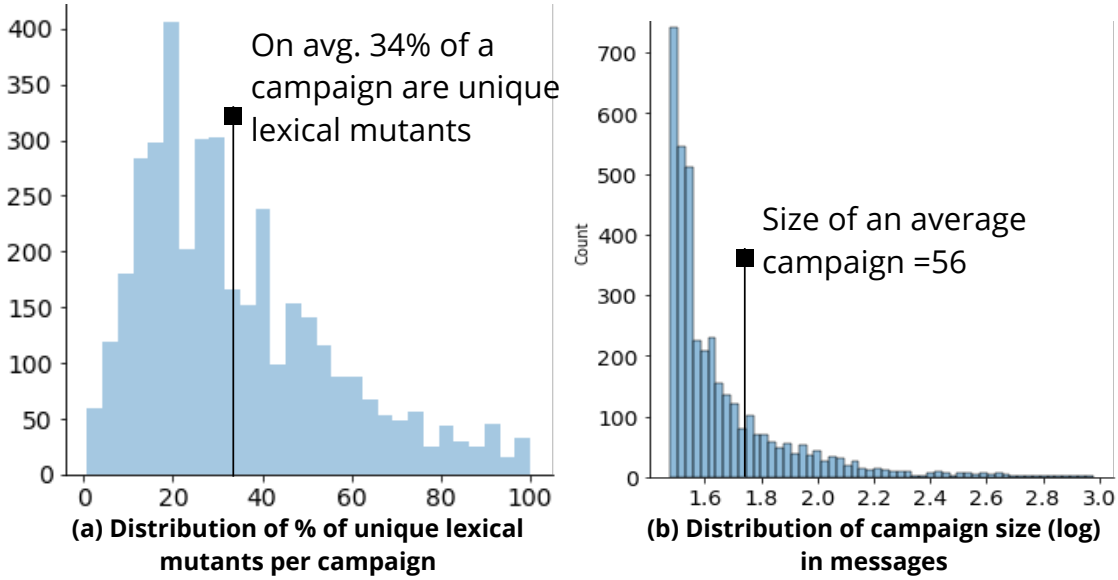


Figure 3.11: Distributions of the proportion of unique lexical mutants per campaign (a) and campaign sizes (b)

words, for the rest of the downstream analysis, "lexical mutant" posts will mean posts with cosine similarity between multilingual embeddings equal to or above 0.85.

3.5.2.3 Finding amplification campaigns with lexical mutations

After identifying lexical mutants through pairwise similarity of multilingual embeddings, to analyze amplification campaigns on a larger scale, it is important to determine large groups of similar messages. For example, there are several other messages similar to the examples in Figure 3.10 with high semantic similarity to each other. How can we find large groups of lexical mutants? We next take a network-based approach to identify groups of lexical mutants.

3.5.2.4 *Network of texts:*

Two messages are connected with each other through an edge if they have a cosine similarity equal to or more than 0.85 between their multilingual embeddings. In a network such as this, finding groups of lexical mutants will be equivalent to finding clusters of nodes that are all connected to each other through high cosine similarity. In other words, clusters of completely connected nodes will represent texts which are all lexical mutants of each other. This is commonly referred to as finding cliques [53]. A clique is a sub-graph in which all nodes are connected to each other through an edge

(example Figure 3.10). In a network of messages connected with high cosine similarity, all similar messages will form a clique.

3.5.2.5 *Finding lexical mutants through cliques:*

Computing cliques in large networks is a computationally expensive task. However, the method in which our network is constructed—drawing edges between nodes only when there is high cosine similarity—allows for a large number of connected components. For example, in the Farmers' protest dataset, we had a total of 799K messages (nodes). Out of which 301,342 nodes had at least one edge. The resulting network had around 2.1K connected components. Connected components make for completely disjoint subgraphs and provide a much more computationally affordable space to find cliques. In fact 1,857 of the total components were also perfect cliques, suggesting that connected components could be a good approximation for finding cliques in a network such as this. A similar approach has proven successful in finding coordination networks in White Helmets [187]. We consider each clique found with this method as a single campaign with lexical mutants.

3.5.2.6 Results: Lexical mutant amplification campaigns

In total, we find 2,558 amplification campaigns relevant to Farmers' protests spanning over 231,896 and 1,268 campaigns in the CAA dataset spanning over 146,465 messages. In the results, we only include campaigns with at least 10 messages with no two messages shared by the same Twitter handle, Facebook group, or page. While is it easier to identify and exclude multiple similar messages posted by a Twitter handle, it is challenging to establish this in the Facebook data, given that we dont have user-level information on Facebook groups or page posts. For every clique, we only consider one message per Facebook group or page. It is possible that we are still considering messages posted by the same user in different Facebook groups, inflating the sizes of the cliques. To find out the extent of this, we manually analyze the authorship distribution for Facebook data points in a random sample of 1000 cliques. We found that 78% of the Facebook messages in a clique come from different users.

An average amplification campaign has 56 messages posted by different users (Figure 3.11 (b)). We found 219 campaigns with at least 100 messages from different user accounts and the largest campaign in the dataset contains 1,232 messages. We find that 29% of all messages in the Farmers' protests dataset and 21% of all messages in the CAA dataset were part of an amplification campaign. Moreover, on average, 34% of

the messages in the amplification campaigns are unique lexical mutations (Figure 3.11 (a)). The rest of the messages are copypastas of different lexical mutants, indicating that including lexical mutants helps in identifying larger amplification campaigns than simple copypastas.

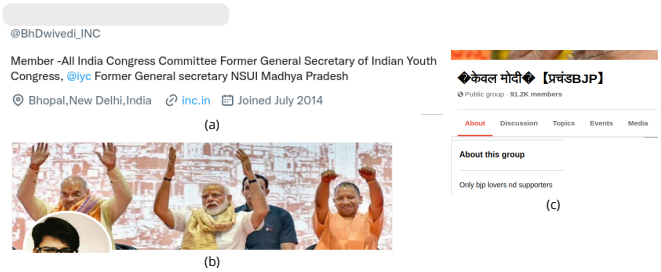


Figure 3.12: Examples of account bio, pictures, and descriptions in the dataset

3.5.2.7 *Evaluating identified amplification campaigns:*

How accurately does the methodology described above identify political amplification with lexical mutations? To evaluate, we analyze a random sample of 200 cliques in the dataset and label for the similarity of the messages in the clique. We use strict labeling criteria— 1 if all messages in the clique satisfy our definition of lexical mutations (see Introduction) and 0 if at least one message is semantically dissimilar to the rest of the clique. We find that 97% of the sampled cliques have all messages with common underlying content, indicating that our methods can identify amplification campaigns with lexical mutants with high confidence.

3.5.3 Characterizing hidden amplification campaigns

Previously, we identified 2,558 amplification campaigns relevant to Farmers' protests and 1,268 campaigns around Citizenship Amendment Act (CAA). While the previous literature has largely focused on studying political influence by the Indian right-wing, in this question we examine the extent of hidden amplification across the political spectrum in India. Moreover, we also analyze the expanse of political amplification across platforms and evaluate dominant narratives. Toward this goal, we first identify the political leaning of the accounts involved in the amplification campaign and measure the use of political amplification on Facebook and Twitter across different parties. We start by outlining our process for identifying the political leanings of social media accounts.

3.5.3.1 Labeling accounts with political leanings

The amplification campaigns detected in the earlier research questions, spread over nearly 40K social media accounts. Currently, there is only one large-scale dataset—NivaDuck—by Panda [188] that records the political leanings of 18,500 Twitter accounts. However, only 368 accounts from NivaDuck overlap with the accounts in our dataset. Hence, we manually analyze the accounts in our dataset and record their political leanings. To better manage annotation resources, we focus only on repeat offenders—accounts that repeatedly participate in different amplification campaigns. In total, we labeled 493 Facebook groups or pages and 1,631 Twitter accounts that are involved in 5 or more amplification campaigns using the steps described below.

Labeling Twitter handles: We first cross-referenced the Twitter handles with the NivaDuck dataset [188] and borrowed the labeles readily available. To label the remaining Twitter handles, we first look for cues in the Twitter bio, username, profile image, and background image for explicit political party affiliation. For example, the Twitter profile in Figure 3.12 (a) explicitly declares an association with the Indian National Congress (INC) political party. In some cases, the accounts also signal political affiliation through profile or Twitter background pictures, as shown in 3.12 (b). For the accounts that do not signal explicit political affiliation, we read through the 20 most recent tweets and also consider the messages in the amplification campaigns associated with the account. We describe the annotation scheme in the next sections.

Labeling Facebook groups and pages: Our dataset contains 493 Facebook groups and pages that host messages used in at least 5 or more amplification campaigns. We label political leanings based on the group's name, description, rules, and recent posts. We find that in most cases, Facebook groups and pages had a clear political affiliation signaled either through group name or description. For example, the Facebook group in Figure 3.12 (C) is restricted only to the BJP supporters. For the other groups, we follow the same methodology as before and assign political leanings based on the recent posts and associated amplification campaigns.

3.5.3.2 *Political leanings labeling scheme:*

The annotation process described above, resulted in the following labeling scheme. We focus on three political parties—BJP, INC, and AAP—that had the largest number of social media accounts in the NivaDuck dataset [188]. We also include % accounts belonging to each leaning in brackets:

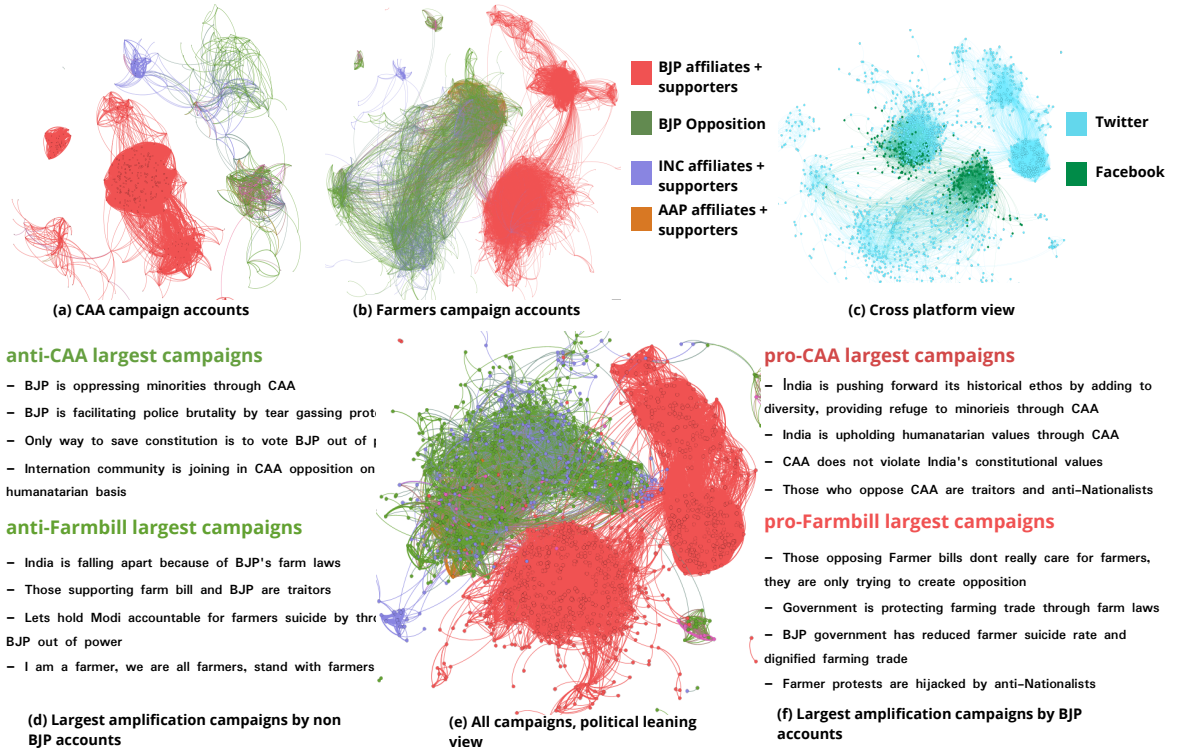


Figure 3.13: (a) and (b) represent the network of users involved in political amplification relevant to CAA and Farmers' protests respectively. (e) is the overall user network across both events. Nodes are the user accounts and colors represent political leaning. The precise percentage of accounts belonging to different political leanings is reported in Table 3.5. (c) represents the combined network of users across Facebook and Twitter. The node color distinguishes Facebook users (green) from Twitter users (blue). (d) and (f) list messages from the largest amplification campaigns from (d) non BJP accounts and (e) BJP accounts

- **BJP affiliates (17%):** Accounts that explicitly signal affiliation with the ruling right-wing political party. Many of these accounts are created in support of the leading BJP politicians and the 2024 general elections. Widely protested Farmers' bills and Citizenship Amendment Act (CAA) were advocated under BJP leadership and support.

- **BJP supporters (35%):** Accounts that did not explicitly affiliate with the BJP but frequently posted content in support of BJP's mission and politicians.

- **INC affiliates (2%):** Accounts that openly affiliate with the leading opposition and Indian left-wing party—Indian National Congress (INC). INC joined the general opposition to Farmers' bills and CAA.

- **INC supporters (10%):** Accounts that do not clearly affiliate with INC but post content supporting INC politicians and mission.
- **AAP affiliates (0.6%):** Accounts affiliated with Aam Admi Party (AAP). AAP is often referred to as the "third opposition" with appeals to the common-man identity in India. AAP also joined in the opposition to Farmers' bills and CAA.
- **BJP opposition (32%):** A large chunk of the accounts that while not affiliating with any specific political party, explicitly oppose BJP politicians and policies.

3.5.3.3 Results: Characterisitcs of amplification campaigns

In the previous analysis, we identified political amplification campaigns around Farmers' protests and CAA and labeled the political leanings of the accounts participating in the campaigns. Here we discuss partisanship, social media platform use, and narratives in the most widespread amplification campaigns. Figure 3.13 displays various views of the network of users involved in the amplification campaigns. Nodes are user accounts and edges connect two users that are involved in the same amplification campaign. Various network views are colored differently to represent user learning and social media platforms. For example, in (a), (b), and (e) node colors correspond to political leaning while in (c), node colors are based on the social media platform of the account.

Political amplification and partisanship: Figure 3.13 (e) represents the network of users involved in political message amplification across both, Farmers' protests and CAA. Overall, 38% amplification campaigns spread through BJP accounts, 40% spread through BJP opposition accounts, and 22% amplification campaigns are propagated through accounts of other parties such as INC and AAP. More granular results across events are present in Table 3.5. Overall, we find that accounts all across the Indian political spectrum participate in political amplification in an equitable way.

After analyzing the networks of users that repeatedly participate in the amplification campaigns, we observe that BJP accounts are strongly clustered together without having edges to any other political leanings. On the other hand, non-BJP accounts (BJP opposition accounts with no specific party affiliation, INC accounts, and AAP accounts) all participate in common amplification campaigns.

Political amplification across platforms: Figure 3.13 (c) displays the network of user accounts involved in amplification with node color representing the social media platform (blue: Twitter, green: Facebook). In sum, 65% of the detected campaigns

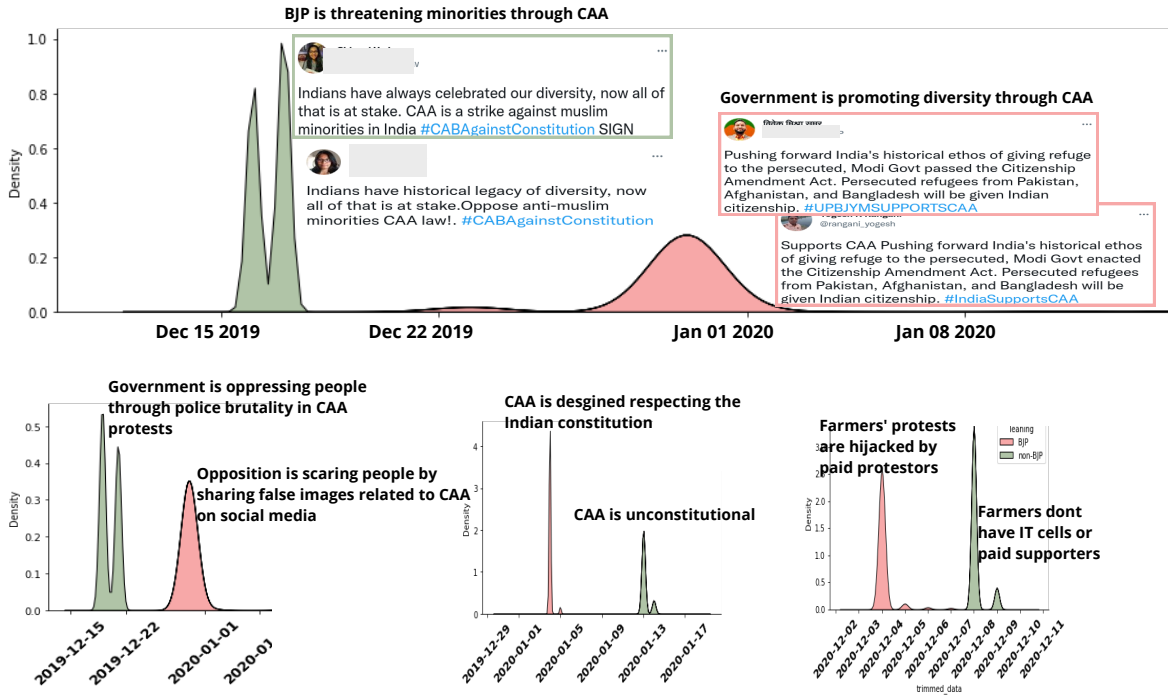


Figure 3.14: Timelines of largest amplification campaigns. In each plot, x-axis represents the actual calendar dates, and y-axis displays the proportion of messages in the amplification campaign posted on that date. We specifically analyze the timelines of the largest campaigns on the opposite sides of the political spectrum that make contrasting claims.

reside only inside Twitter, while 5% are confined to Facebook. 30% of the amplification campaigns spread across both platforms. Given the size of the datasets and

Prominent political claims in amplification campaigns: We further analyze the dominant claims in amplification campaigns across different political leanings in Figure 3.13 (d) and Figure 3.13 (e). Across both events, the largest amplification campaigns by BJP and non-BJP accounts could be considered direct antitheses of each other. For example, the largest anti-CAA campaign by non-BJP accounts claims that BJP is oppressing minorities through CAA. Whereas, the largest amplification campaign by BJP accounts claims that India is continuing their historical legacy of providing refuge to minorities through CAA. Similarly, non-BJP accounts attempt to depict Farm Bill supporters as traitors whereas BJP accounts criticize the opposition to Farm Bill for its political opportunism. Noting the pattern of contrasting political claims in the amplification campaigns, we further analyze the timelines of the contrasting campaigns in Figure 3.14. Specifically, we investigated 63 campaigns that had at least 500 messages and manually identified claims and counterclaims as described above. Not all campaigns

accounts from	CAA	Farmers' protests	Overall	TW	FB
BJP	47%	35%	38%	37%	39%
BJP Opposition	35%	44%	40%	41%	40%
Other parties	18%	21%	22%	22%	21%

Table 3.5: Participation in political amplification by accounts of different political leanings. Here we report on the proportion of accounts with different leanings in different political events and platforms.

had a counterpart in their political opposition, however, we recorded 20 campaigns that had counter-narratives in the form of other campaigns.

We find that the claims and counter-claims follow precise temporal orders. For example, the bulk of messages suggesting that BJP is threatening minorities through the CAA (Figure 3.14) were posted by non-BJP accounts around December 15th, 2019. Amplification campaigns by BJP contesting this claim were spread during the subsequent days claiming that CAA promotes diversity and the historical legacy of providing refuge to minorities. Similar trends followed in the 17 other contrasting campaigns. Overall, the temporal order political messaging suggests that amplification campaigns are possibly being used as devices to counter the narratives from both sides of the Indian political spectrum.

3.6 Discussion and Implications

In this chapter, I presented my research investigating the ways in which hate groups utilize various online platforms to frame content. Moreover, by analyzing how various accounts participate in hateful information sharing, I displayed how information in hate movements travels through online accounts assuming different roles such as educator, solicitors, etc. I also extended the research in communication practices around problematic information by studying political amplification in India. In this discussion, I will reflect more deeply on the findings reported in this chapter.

3.6.0.1 Do hate groups use Facebook and Twitter differently?

We find that on Facebook, *fear* is prominently used as a motivating agent. Previous literature suggests that *fear* appeals is a common mechanism for hate groups to strategically

recruit like-minded people that are predisposed to the group's ideology [166]. Our framing analysis, shows that hate groups indeed use more *fear* appeals on Facebook. This indicates that on Facebook, hate groups' might be directing their communication towards a more like-minded audience, one that already aligns with their ideological worldview. Further, hate groups claim to be oppressed and put out calls for *membership* at a higher rate on Facebook compared to Twitter. Messages that pose such negative threats to one's existence and sovereignty are particularly persuasive [151]. This might suggest that hate groups use Facebook to not only direct their hateful agenda to an ideologically aligned like-minded audience, but they are doing it effectively, through clever persuasion strategies. Whereas on Twitter, hate groups direct messages to describe out-group as *immoral* or *inferior*. Scholars suggest that when faced with diverse or initially reluctant audience, hate groups use different recruiting strategies than what they would use with a like-minded audience. Specifically, they try to present inaccurate or distorted perception of the out-group by stressing on the out-group's negative stereotypes [166]. This may suggest that hate groups might be using Twitter to specifically cater towards a diverse audience and using framing strategies to bring initially hesitant users into their core follower base.

3.6.0.2 What are the Possible goals behind social media usage by hate groups?

Previous studies on SMOs show how SMOs use different social media platforms for different purposes [14]. In positioning hate groups within the SMO perspective, we are interested in examining how hate groups might be using these platforms for their extremist agenda. Here, we identify two possible dimensions of hate group activity on the platforms.

Group radicalization and recruitment on Facebook Researchers argue that the in-group's psychological need to survive, be significant and important for their cause can be associated with the radicalization process [145]. Specifically, McCauly et al. [163] carve out factors relevant in group radicalization. They argue that radicalization can be associated with beliefs like: "we are a special or chosen group who have been unfairly treated and betrayed (*oppression*), no one else cares about us or the system has failed us (*failure*), and the situation is dire—our group and our cause are in danger of extinction (*fear* appeals)." In our analysis, we observe that the rhetoric of *oppression-failure-fear* is more frequent on Facebook compared to Twitter. Moreover, Facebook has more *membership* calls (see Figure 3.2), links to personal mailing lists and recruitment forums. These results suggest that the Facebook audience of hate groups might be

more susceptible to extremist radicalization and successful recruitment, compared to Twitter.

Mass education and image Control on Twitter Communication scholars studying hate groups have observed that they often try to "educate" their audience [166]. Specifically, they attempt to spread negative news that address the problems associated with the out-group and stress positive aspects of the in-group itself. The efforts to educate a wider audience and promote positive self-image are commonly discussed together by other researchers as well [77]. We find that hate groups use Twitter to predominantly share news from known news media. They also dehumanize the out-group by describing them as inferior or immoral while presenting a positive self-image through *status* enhancement and effectiveness (*efficacy*) of their proposed solution (e.g., "..doing the god's work in fighting the LGBT mafia.." or, "..fighting the good fight.."). Presenting positive self-image and ideology-aligning information can enable hate groups to appeal to the general public. What better way to do that than to use Twitter to gently coax the follower base into a more radical world of hate.

3.6.0.3 Effects of the Current Censorship Models

Recently, Facebook issued a wave of bans on known White Supremacists and neo-nazi account holders [84] which resulted in popularity of alternate platforms such as `gab`, `bitchute` (also demonstrated in our domain networks, see Figure 3.4). Particularly, a deeper look at messages annotated with *policy* and *membership* frames and the domains in the *promotion* category, demonstrate that White Supremacy hate groups are quickly adapting and finding ways to subvert censorship. Some accounts offered detailed guides describing how to get around censorship by installing Virtual Private Networks (VPN) and by avoiding using specific terms. We observe that promotion of alternate social media and ways to bypass censorship is only evident in White Supremacy accounts. Together these results suggest that social media companies are selectively censoring only one type of extreme ideology—White Nationalists—whereas other hate groups are still thriving online and gaining followers. Should social media companies selectively ban specific hate ideologies or, for that matter, any content that originates from known hate groups? Policy experts have posed bilateral arguments around the notions of freedom of speech and a platform's responsibility in restricting and moderating hateful communication. For example, Aswad's work outlines several challenges associated with deriving meaningful online content moderation policies while also aligning them with international human rights law [12]. McDonald et al.'s work also

laid out the challenges in online governance of extremist content, including lack of clear directive and the inability of moderation algorithms to distinguish different types of extremism [157]. Together, our findings and discussion indicate the need for further research exploring the design of online censorship and moderation models—one that carefully balances the arguments around policy, human rights law, and the need to make online spaces safer for a diverse population.

In this chapter, we looked at how hate movements thrive on social media, both, through cross platform communication by hate group leaders and information mobilization by numerous unaffiliated social media participants. Our results can offer insights into how participatory activism advances extremist movements, how various roles are located on the pathways to deeper engagement into extremism, and what could be the possible effects of interventions in countering online extremism undertaken by these roles.

3.6.0.4 Online Extremist Movements and Participatory Activism

Scholars have attributed the advancement of social movements to the successful distribution of resources through its participants [161]. We observe that through participatory activism, extremist accounts, in various roles, adequately use social media to spread various types of information resources. For example, *educators* and *solicitors* dedicate a large proportion of their Facebook activity to distributing extremist content for educating and soliciting the readers into extremist movements. In sum, by disseminating information through their Facebook accounts, mass educating the readers about their agenda, and soliciting funds and participation in the movements, *educators* and *solicitors* are creating human and material resources [161]. Moreover, by prominently sharing misinformative content and using toxic language, *flamers* may be raising emotional resources that create opportunities for public outrage and eventually, collective action [263] to advance the hateful agendas of their extremists movements. This distributed system of online information mobilization—distribution of various information resources through various roles online—can be compared to the democratization process in participatory activism [144]. The digital democratization process specifically consists of more equitable sharing of informational resources amongst the participants [42]. Take for example, the Facebook page of Alliance Defending Freedom (ADF)—an *educator* account in our study, representing a leading anti-LGBTQ organization [235] that the Southern Poverty Law Center has tracked for decades. This organization started with a small group of christian leaders advocating for discredited

practice of conversion therapy, criminalization of LGBTQ sexual acts and opposition to the transgender rights. ADF started with 84K Facebook page likes in 2012. Today, ADF's Facebook page has over 1.7 million page likes and over 1.6 million followers. Our dataset also revealed that over 1K other Facebook groups and pages have already shared content linking to ADF's website. With 1.6 million direct followers on the ADF's official Facebook page and an additional indirect exposure to users through shares on other Facebook pages, the material created by ADF is able to reach a vast audience. This may suggest that, through participatory activism, the picture of online extremism has now shifted from a few selected radical websites and accounts to a spectrum of allies with access to extremist information and the affordances to share it with the mass.

3.6.0.5 Trajectories of Extremist Movement Participation

Klandermans et. al. proposed a trajectory of social movement participation comprising four steps [139, 140]. First, people must sympathize with the ideals and goals of the movement, thus turning into potential targets for mobilization. Next, they must be targeted by core members' mobilization attempts. Next, they must develop motivation to participate in the movement and finally overcome possible barriers and engage in collective action [241]. Through our analysis, we identify a group of accounts played the role of *sympathizers*—pages/groups expressing sympathy towards the extremist causes without getting heavily involved. 36.4% of the accounts in our dataset are *sympathizers*. Based on Klandermans's models, *sympathizers* can also be viewed as the biggest potential group of supporters for the extremist movements. Interestingly, we also see that *educators*, *solicitors*, and *flamers* have high influence on *sympathizers* in spreading extremist content. This is the second step in the Klandermans's model, whereby *sympathizers* are targeted for mobilization. In other words, sympathizers may lie on the first two steps of the Klandermans's trajectories of participation. The third step consists of participants who have developed motivation for participating. The *flamer* and *motivator* roles are primarily driven by motivating factors such as anger, injustice, and the sense of achievement. Hence, *flamers* and *motivators* might be on the third step of Klandermans's trajectory. Finally, we believe that *solicitors* and *educators* are on the last step of the participation trajectory as they actively try to educate and proselytize others through collective action. In summary, based on Klandermans's comprising, *developing sympathy → getting targeted by mobilization → developing motivation → collective action*, is equivalent to the following role transitions: *sympathizers →*

(*flamers* or *motivators*) → (*educators* or *solicitors*), which together represents a trajectory of deeper engagement into the extremist movements online. Can targeted interventions for counter-extremism stop users from getting induced into the deep trenches of extremist movements? Below, we discuss which roles could potentially benefit from interventions designed for countering online extremism.

3.6.0.6 Theoretical Implications: Parallels Between Theoretical and Online Roles

Some of our roles correspond to the categories of participants identified in theoretical research based on physical protest events and social movements. For example, the *educators*—accounts that primarily focus on distributing links from extremist domains—may correspond to "constituents" as described by McCarthy and Zald [161]; "constituents" are primary distributors of resources. Similarly, our *solicitors*—who actively solicit participation via donations and gatherings—may correspond to "beneficiary constituents" [161] who stand to gain from the success, funds, and connections emerging from the movement. The *sympathizers* category may be similar to "bystanders"—a group of third-party participants, as defined by Turner et. al. [251], who might acknowledge grievances related to the issues of social movement and take a sympathetic stand. However, the *motivator* role does not resemble any of the theoretically described categories. We believe that the *motivator* role is specifically relevant in the online setting for relaying positive news and wins related to the extremist causes. Additionally, *flamers* also do not correspond to any theoretical roles. Through our data driven methods, we are able to surface these new roles that characterize online participation in extremist social movements. We believe that our framework for identifying roles based on the characteristics of participation can be extended to other social movements as well.

3.6.0.7 Practical Implications: Interventions for Online Extremism Engagement

While this study focuses on identifying roles, it can also inform the design of interventions for countering extremism. Our results suggest that while accounts core to the extremist movements—*educators* and *solicitors*—tend to retain their roles, others are more likely to transition to different roles. For example, *flamers* and *motivators* become *sympathizers* with high probability. *Flamers*, *motivators* and *sympathizers* also show more sporadic engagement with sharing extremist links compared to the *educators* and *solicitors*. A study by Siegel and Badaan revealed that targeted interventions against hate speech, such as sanctions on hateful messages, leads users to tweet less

hateful content, especially if the individuals are less engaged with the hate speech in the first place [226]. On the other hand, accounts that frequently see or produce hostile language are less likely to get deterred by sanctions and may even express backlash [226]. Other researchers also report that rather than conforming to the community norms upon receiving sanctions, the producers of hostile content are more likely to move to other platforms [177] or find creative ways of continuing their hate speech [44]. For example, recall that the Facebook page, *Pissed off White Americans*, described in the Introduction, shared videos that are now banned on YouTube. However, they still made the extremist videos available to the readers by hosting them on bitchute.com which has been described as the "hotbed for violence and hate" [2]. Considering this, our results suggest that *flamers*, *motivators* and *sympathizers*—accounts infrequently exposed to extremist content—might benefit most from targeted interventions designed to counter extremism. On the contrary, *educators* and *solicitors* may retaliate or relocate to alternate platforms in response to an intervention.

3.6.1 Political Amplification in India

In the last research question, I analyzed how political parties leverage amplification campaigns with lexical mutations to popularize contrasting political stances on Indian social media. Focusing on two key recent events —Farmers' protests and the introduction of the Citizenship Amendment Act (CAA)—we first identified over 3.8K amplification campaigns across Facebook and Twitter. Next, we characterize the use of amplification campaigns across multiple political parties in India and multiple and social media platforms. Our results provide an updated understanding of political amplification by looking beyond the popularly studied platform Twitter and by considering the multi-party political landscape of India.

3.6.1.1 *Hidden amplification and platform manipulation policies:*

In one of the previous studies [127], researchers found that local political leaders in India seek out participation through WhatsApp groups to spread messages from "Tweet banks". One such message quoted in their paper explicitly instructs users to alter the wording in the template: *"Note-Please don,Äôt just copy-paste the sample tweets, please alter it a bit."*[[127] Pg. 9]. Our results indicate that users might be getting these types of instructions on a larger scale for campaigns of different political parties. Moreover, we argue that this kind of hidden amplification might be designed to bypass the platform

manipulation and spam policies [14]. For example, on Twitter, copypasta campaigns are subject to review, and the platform is committed to reducing the visibility of such content. In fact, repeatedly violating copypasta rules is considered a severe violation of the platform manipulation policy. In our study, we identify thousands of user accounts that repeatedly participate in amplification campaigns (Figure 3.13). The fact that we were able to detect large-scale campaigns and repeat offender users, who are still active on the platforms, may suggest that hidden amplification may be proving effective in evading platform governance policies. In the future, a more standardized and scalable approach such as the one suggested in RQ1 could be required to counter large-scale platform manipulation. For example, our entire pipeline to detect amplification campaigns with lexical mutants could be used in real-time to detect political amplification. Detecting messages involved in amplification could help platforms limit their visibility soon after they are published, and also keep track of repeat offenders across different political events.

3.6.1.2 Indian political amplification campaigns and reactionary politics

Our study offers a unique look into the political influence exerted by accounts from various political leanings in India. By considering the multi-party political spectrum in India, we provide an essential context to previous research focusing on only one political party. More importantly, our results reveal the reactionary nature of political influence in India (Figure 3.14). Specifically, by investigating the temporal order of contrasting amplification campaigns, we observe that large-scale amplification campaigns arise in the form of arguments and counter-argument between different political viewpoints. Researchers studying the role of copypasta in reactionary politics around neo-Nazism in the United States propose that such reactionary politics often gives rise to radicalization [250]. Specifically, influencing public arguments and counter-counterarguments through amplification can harm democratic equality and can create an exaggerated sense of democratic divide in the public eye. Our methods can also be extended to study the temporal patterns in political amplification campaigns outside of India.

[14] https://help.twitter.com/en/rules-and-policies/copypasta-duplicate-content

3.7 Limitations

This chapter presented completed research on various communication practices employed in communities of problematic information. This work has some limitations which are discussed separately for each research study.

3.7.0.1 Limitations in cross-platform analysis

Our data is limited to 72 hate groups, constrained on a three-months time span, and focused on only one source for identifying hate groups—SPLC. Hence, we do not argue for generalizability across all possible hate communities in the U.S. or its representativeness outside of the three months period. However, we want to note that the three months window was randomly selected and was not marked by any major socio-political event that could have potentially affected our results.

3.7.0.2 Limitations in studying information mobilization in extremist movements

First, our dataset contains only US-based extremist websites and most hold far-right political ideology. However, this skew toward far-right might not correctly represent the political scenarios from other countries. For example, unlike the U.S., Germany has observed increased political violence from both, far-left and far-right ideological groups [**?**] and is known to have a history of violence from both ends of the political spectrum [132]. Hence, while applying our study results in the context of other countries, researchers need to be cautious about the distribution of extremist ideologies in our dataset. Next, we compiled extremist accounts based on their sharing behavior, specifically the number of unique links they shared from known SPLC-designated extremist websites. While this is a common methodological choice made when choosing users/accounts for studying social media activity, a stricter selection criteria can be beneficial. For example, in addition to the frequency of extremist link posts, extremist accounts can be selected based on the topics discussed in the posts. Moreover, while our data collection spans the activity of extremist accounts in 2018 and 2019, it was collected post-hoc, in May 2020. Once a Facebook page/group is banned, its data is no longer accessible through CrowdTangle or any official Facebook API. Thus, considering the recent Facebook bans on white nationalist accounts, our study doesn't contain the data for the extremist accounts that were active in 2018-2019 but got banned before May 2020. In the future, researchers can collect data in real-time and extend our study to understand what roles among the extremist accounts face the highest moderation.

Finally, our results reveal the information ecosystem of extremist movements on just one platform—Facebook. The problem of extremism is also evident on other platforms such as Twitter and Gab. Indeed, researchers have shown that the extremist movements leverage different social media platforms towards different goals such as radicalization and mass education [195]. We encourage future researchers to extend our methods to model roles in extremist movements on other platforms or even, across the platforms.

3.8 Conclusions

This chapter presented my research on information sharing and mobilization practices in communities of problematic information. The first two research questions contribute to the understanding of how hate groups utilize various platforms for offering different framings of hate ideologies, and how hateful information is mobilized through accounts with different social roles. We leveraged three months of Facebook and Twitter activities of 72 SPLC-designated hate groups spanning five ideologies. We found that hate groups complain about being oppressed, use *fear* appeals, promote calls to join their movement, and share informational resources and opinion pieces more on Facebook. On the other hand, they call out their targets by depicting them as immoral or inferior, demand governmental or legislative *policy* changes, and motivate their audience through the positive self-image on Twitter. Our findings indicate that hate groups might be using Facebook to radicalize already like-minded audiences while using Twitter to educate a more ideologically diverse set of followers. Moreover, we identify five social roles in online hate movements: *educators solicitors*, *flamers*, *motivators*, and *sympathizers*. Our findings offer a perspective on how participatory activism might be advancing hate movements and how various roles may be targeted for mobilization.

In the last research question, I analyze Indian political amplification campaigns with lexical mutations. By modeling messages in national political events, I uncover over 3.8K political amplification campaigns across languages, social media platforms, and political parties in India. This type of amplification also raises concerns about the legitimacy of political expression on social media in rising democracies.

CHAPTER 4

DISENGAGEMENT FROM COMMUNITIES OF PROBLEMATIC INFORMATION

Published Works

Shruti Phadke, Mattia Samory, Tanushree Mitra. 2022. "Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. Proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2022.

Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. Characterizing Social Imaginaries and Self-Disclosures of Dissonance in Online Conspiracy Discussion Communities. Proc. ACM Hum.-Comput. Interact. 5, CSCW2, Article 468 (October 2021), 35 pages. https://doi.org/10.1145/3479855

With the rise in divisive and antisocial communities of problematic information, such as racial extremism or climate change debate or anti-vaccination attitudes, it has become important to explore mechanisms through which social media users could be dissuaded from participating in such online interactions. Building on the previous work studying communication practices in communities of problematic information, this chapter studies disengagement from such communities. A primary step in studying disengagement from conspiracy theory discussions could be to observe the organic process of how users find pathways in and out of conspiracy discussion communities. For example, what could be the different trajectories of engagement after users join online conspiracy theory discussions? What are the differences in the quality of engagement between users who increasingly participate more as compared to the users who eventually disengage? And finally, what are the key social, psychological or

epistemological mechanisms that make people leave behind their conspiracy beliefs?

This chapter grapples with these questions by first contrasting users that travel through various engagement pathways inside conspiracy theory discussion and then investigates disengagement by using cognitive dissonance as an underlying mechanism in the case study of QAnon community. Below, I outline the background works inspiring the research in this chapter.

Specifically, by studying the online QAnon discussion subreddits, this chapter proposes cognitive dissonance as one of the disengagement mechanisms from conspiracy theory discussions.

4.1 Background

Research studying the natural processes of disengagement from conspiracy theory discussions or other problematic content is limited. However, there are some key studies that provide valuable insights into the role of technology, information, and social psychology in disengagement from problematic content. Below, I will summarize the existing research and theories that motivate the research questions in this chapter.

4.1.1 Research exploring disengagement from problematic content

Most existing research observes the process of disengagement from conspiracy theorizing isolated from the influence of internet or social media. However, given the role online conspiracy theorizing can play into mobilizing participants offline [28, 102], I argue that it is important to understand the process of radicalization and recovery from conspiracy belief in online discussions. Specifically, understanding the natural mechanisms by which user adopt engagement or disengagement trajectories inside online conspiracy theory discussions can provide insights for designing interventions that can artificially pave the way out of conspiracy theorizing for online users. In the first research question, I first characterize various engagement pathways inside conspiracy theory discussions and then contract various pathways using theory guided measures.

4.1.2 QAnon conspiracy theory

In the second research question, I study cognitive dissonance as one of the mechanisms through which users may disengage from conspiracy theory belief inside the QAnon

community. Here I will first provide some background into what is QAnon conspiracy theory.

In October 2017, a user on the image board, 4chan, signed off as "Q" and posted a comment prophesying the arrests of Hillary Clinton and her staff. In successive posts, "Q" purported the arrests of several other politicians associated with the "Deep state"—a conspiracy theory claiming that a coalition of politicians in the U.S. run a shadow government involved in corruption and cronyism [17]. In the next several months, the discussions around Q's posts took 4chan by the storm. Q presented themselves as an anonymous, high ranking U.S military official, with insider information about the U.S. government. Together, "Q", a mysterious, prophetic leader and "anons", Q's followers comprise the QAnon community. Specifically, Q predicts various political events in their Q-drops—messages posted to image boards such as 4chan, 8chan. Subsequently, followers of QAnon start piecing together the clues left in Q's posts and predictions. In fact, Q-drops are carefully crafted to contain cryptic messages such as "*the wormhole goes deep*" or "*future proves past*" which are meant to be clues for the followers to decipher. By encouraging followers to "open their eyes" and "search for truth" Q has institutionalized knowledge production practices that are unparalleled by other conspiracy movements [190]. Specifically, Q-followers are called "bakers" who assemble the "crumbs" (clues) left by Q into coherent pieces [6, 190]. This social construction of knowledge then produces unambiguous certainty through alternate reality [190], a characteristic commonly associated with new religious movements [16]. In this shared alternate reality, Q and their audience identify themselves as actors in a larger movement by using specific designations such as "anons", "patriots", "bakers". A large part of QAnon's alternate reality and worldview is represented by the use of symbolic language whose shared meaning is understood only within the community.

QAnon community has started attracting research attention due to its clear mobilizing potential. Specifically, previous studies explored the topics and dissemination of Q's messages on various online platforms and found that QAnon borrows theories from other conspiracies such as Pizzagate [189], shares moral values with Christian theology [167] and QAnon followers are likely to use violent rhetoric on Twitter [202]. While the existing studies provide valuable characterization of the QAnon movement across platforms [5], deeper psychological and sociological exploration is required to deter increased engagement with QAnon [103]. Our work fills this gap by establishing QAnon social imaginaries symbolizing collective interpretation of reality by QAnon. Next, I discuss the role of social imaginaries in the conspiracy communities.

4.1.3 Conspiracists and Social Imaginaries

Conspiracy theorizing generally consists of a belief that a covert operation is being carried out by a group of conspirators or secret organizations to influence events [134, 200]. Conspiracy theorizing is able to produce a certain aesthetic pleasure that enables people to form social imaginaries—coherent, collective imagination of social existence by a set of people [152]. Social imaginaries refer to the ways people imagine their social existence, their relationship between different social groups, deeper normative notions and the expectations of reality born out of such norms [246]. For example, a social imaginary of a conspiracy theorist can consist of irrational interpretation of reality, such as, "there is a global lobby trying to enslave common people" or "conspiracists being ridiculed by ignorant mass are further proof of the subversion by elites" [152] or QAnon's purported worldview that "America is run by a cabal of pedophiles and Satan-worshippers who run a global child sex-trafficking operation and QAnon are force of good stopping them" [178]. Conspiracy theorists may interpret everyday, trivial events as plots of manipulation by conspiring agents [86]. Conspiracists' social imaginaries can also determine the differences between the insiders, those who know what's really going on, and the outsiders, those who live in the ignorance of reality perceived by the insiders [39, 152]. In sum, social imaginaries lie at the heart of the belief systems and are a way for groups to rationalize their sense of reality and even find purpose in the collective action [246]. Thus we argue that it is important to understand the conspiracists' social imaginaries. After we understand what are the social imaginaries or worldviews of conspiracy theory believers, we can next investigate how such worldviews can get fractured. Next, I discuss cognitive dissonance theory which offers one possible mechanism that can induce conspiracy theory disbelief.

4.1.4 Conspiracy Theorizing and Cognitive Dissonance

While not all conspiracies are false or impossible [18], many suffer inconsistencies, contradictions, and general epistemological challenges [243, 275]. Realizing such inconsistencies may induce a state of dissonance among conspiracy followers. How do conspiracy believers react when their beliefs are contradicted? Researchers found that mistrust in authorities or governing bodies is sufficient to overwhelm the contradictions between individual conspiracy theories [275]. For example, Wood [275] found that participants who distrust the official story of Princess Diana's death, often simultaneously support contradictory accounts of the same events such as Princess Diana

faked her own death or she was murdered [275]. A follow up study revealed that participants rationalized belief in contradictory claims by viewing them as probable alternate explanations of the same event [156].

Festinger coined the phenomenon of believing in contradictory, inconsistent ideas as "cognitive dissonance" [88, 90]. In a famous immersive ethnographic study, Festinger and colleagues infiltrated a UFO religion in Chicago where the cult leader had prophesied that the world will end on December 1954. Festinger and colleagues revealed that after the prophesied date and obvious signs of world not ending, members of the group experienced cognitive dissonance. According to the theory of cognitive dissonance proposed by Festinger [88], dissonance can result from various individual or social factors. For example, dissonance can result from involuntary or voluntary exposure to information that directly contradicts previously held beliefs. Further, the simple act of having to choose between two contradictory ideas can also intensify the experience of dissonance. Moreover, dissonance can result from the conflict between individually held and socially accepted beliefs. When confronted with such dilemmas, individuals may change their perception of contradictory ideas, find overlap between the two ideas or completely reverse their previously held beliefs [89]. Indeed, Festinger's study showed that after experiencing dissonance, different believers reacted in different ways. While some strengthened their convictions and even recruited newer members, others left the cult. In other words, being in the state of dissonance, where beliefs are challenged or contradicted, may lead people to change their behaviors or attitudes [88, 203, 240].

4.1.5 Cognitive dissonance and QAnon

Similar to the cult leader in Festinger's study, Q—the leader of QAnon—has made several predictions that never came true [248]. For example, amongst hundreds of other predictions, Q's very first prediction about Hillary Clinton's arrest in 2017 has provably failed [248]. Can Q's failed predictions, similar to the UFO religion studied by Festinger, induce dissonance in Q's followers? What other fracture points in QAnon belief can induce cognitive dissonance? In this light QAnon makes for an ideal case to study dissonance in conspiracies, where we can analyze both social imaginaries created by Q and the self-disclosures of dissonance by Q followers on the fracture points of the social imaginaries.

What happens after people express dissonance? Researchers studying addictive behaviors found that higher levels of cognitive dissonance can help people deconstruct

their previously held norms and beliefs and thus, pave the pathway for recovery[29, 99, 256]. In this work, we analyze changes in user engagement with QAnon discussion communities after they express dissonance with QAnon.

4.2 Research Questions

Guided by the background and theories explained above, I now outline research questions undertaken in this chapter. With each research question, I will explain the gap in the literature that motivates the question, and describe methods and findings in short.

- **RQ1: What are the similarities and differences in various engagement pathways in online conspiracy theory discussions?** After joining online conspiracy theory discussions, users beliefs may strengthen or dissolve over the time. Despite the clear implications of online CT engagement, we know little about an individual's journey through CT discussion communities and how they become increasingly engaged (or not) with conspiratorial worldviews. This study provides just such an understanding. In this research question, I first characterize users' longitudinal engagement in conspiracy theory discussions online. This research exploration will follow individuals from the first interaction with conspiracy theory content online, through the engagement with CT-related groups, to the eventual disassociation from conspiracy theories. Next, using various novel theory-guided measures of radicalization, this study contrasts users with increasing engagement in conspiracy theories with those who eventually disengage to identify online behaviors specific to the pathways out of online conspiracy theory discussions.

- **RQ2: How do users express dissonance inside QAnon community and how does it relate with their disengagement?** Believing in contradictory ideas induces the state of cognitive dissonance that can prompt individuals to take belief rejecting actions [88]. In particular, experiencing dissonance with conspiracy belief can motivate individuals to depart from conspiratorial views. Hence, studying how conspiracists express dissonance with their beliefs is crucial towards understanding the pathways of recovery from conspiracy theories. How can we identify expressions of dissonance with conspiracies? Based on the QAnon social imaginaries and other theoretically-informed constructs of belief and dissonance,

such as the language of doubt or tentativeness [83], credibility cues [170], integrative complexity [62], this study creates a computational framework to identify self-disclosures of dissonance in Reddit's QAnon discussion communities. With an interrupted time series analysis this study finds that user contributions in QAnon communities decrease significantly, immediately after dissonance disclosure, but not after expressing belief, while their overall engagement on Reddit stays the same. Not only disclosures of dissonance are followed by a decrease in contributions, but also by the departure of the users from the community. In particular, users who disclose dissonance disproportionately more than belief, are those most likely to leave the community.

4.3 RQ1: Contrasting pathways of engagement inside conspiracy theory discussions

RQ1 first aims to uncover various longitudinal patterns of engagement, or pathways, that users follow *after* online joining conspiracy theory discussions before analyzing user behaviours across various engagement trajectories. Next I will discuss data collection and analysis methods for characterizing user engagement pathways in online conspiracy theory discussion communities.

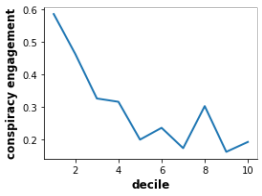
4.3.1 Data and Subjects

Reddit is a great venue for studying the evolution of user engagement in online conspiracy discussions as it offers explicit community based structure in terms of subreddits, where an individual's activity on Reddit can be tracked across thematically diverse communities. To mark the users' entry into Reddit's CT discussion world, we look at users' first comment into `r/conspiracy`—Reddit's biggest and most popular conspiracy discussion community. As of this writing, `r/conspiracy` amassed around 1.7M subscribers and has been growing at an average of 100K subscribers every year over the past five years[1]. The subreddit claims to be a widely inclusive community that invites all domains of conspiracy theory discussions. Given the huge number of contributions made everyday in diverse set of conspiratorial topics, `r/conspiracy` makes for a great place for users' entry into Reddit's conspiracy theory world.

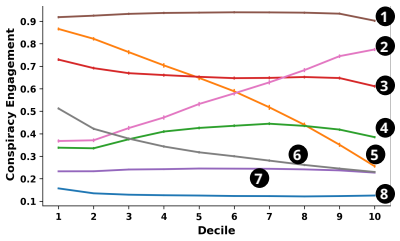
[1] `https://subredditstats.com/r/conspiracy`

1st decile	3rd decile	5rd decile
technology 47.3%	technology 39.8%	C_S_T 80.7%
conspiracy 35%	conspiracy 41.2%	conspiracy 16.1%
trashy 11.6%	C_S_T 15.6%	politics 1.8%
politics 4.2%	cringe 1.9%	news 1.4%
news 1.9%	news 1.5%	

(a) Example of a user's activity across subreddits in different deciles



(b) Example of a user's conspiracy engagement trajectory



(c) Initial representative trajectories of conspiracy theory engagement after clustering

Figure 4.1: Figure illustrating various stages of characterizing conspiracy theory engagement of Reddit users. (a) Example of a user's Reddit activity in different subreddits grouped by deciles. The percentages symbolize the proportion of comments or submissions made to that subreddit in a decile. (b) Example of a user's conspiracy theory engagement trajectory calculated with user activity data as displayed in (a) in combination with *conspiracy scale* using the formula 4.1. (c) Initial results of clustering individual user's conspiracy theory engagement trajectory. The initial clustering resulted in 8 different trajectories reflecting the general patterns of steady high, increasing, decreasing and steady low engagement in conspiracy theory discussions on Reddit.

Reddit users differ in their frequency and levels of contribution. To compare evolution of all users on equal footing, Reddit activity can be split in ten equal deciles of contribution volume. For example, if a user x makes their first comment in `r/conspiracy` on 1st January 2018 and makes n comments on Reddit since then, their entire Reddit activity *after* 1/1/2018 is split into equal, time ordered batches of $n/10$ comments (example, Figure 4.1 (a)). A similar approach for mitigating temporal shifts in user activity has been validated in studying far-right radicalization on Twitter [265]. Using these temporally ordered bins of users' Reddit activity, we can next characterize users' conspiracy theory engagement pathways.

4.3.2 Characterizing engagement pathways in Online conspiracy theory discussions

To understand a user's journey through Reddit after participating in conspiracy theory discussions, we first need to model each user's longitudinal development of CT engagement. Simply put, the engagement pathways can capture the amount, over time, at which the users engage in conspiracy theory discussions on Reddit. This can be achieved by first characterizing how conspiratorial are the subreddits that users participate in.

4.3.2.1 Characterizing the conspiratorial nature of the subreddits

My previous work [197] described in Chapter 2 contains a well validated *conspiracy scale* (Figure 2.1 b) that maps subreddit across the scale of similarity to `r/conspiracy` vs. `r/science`. Using this scale, subreddits can be mapped on a scale from -1 to 1 where 1 represents the highest similarity to `r/conspiracy`. Using the *conspiracy scale* we assess the proportion of users' Reddit activity in subreddits that discuss conspiracy theories.

4.3.2.2 Measuring user engagement in conspiracy theory discussion subreddits

The conspiracy theory engagement can measure captures what proportion of the user's Reddit activity is dedicated to CT discussion subreddits in each decile. The weighted average of contributions in each subreddit weighted by that subreddit's score on the *conspiracy scale* can reflect the amount of Reddit activity users spend in discussing conspiracy theories. More formally, for a user u, with N contributions in a decile i, their conspiracy engagement score C_u^i will be calculated as:

$$C_u^i = \sum_{j=1}^{J} \frac{n_j s_j}{N} \tag{4.1}$$

where n_j is number of contributions in the j_{th} subreddit and s_j is the subreddit's score on the *conspiracy scale*. C_u^i is bounded between 0 to 1 with higher scores indicating higher proportion of engagement in the CT discussion subreddits.

4.3.2.3 Modeling trajectories of users' conspiracy theory engagement

Every user's conspiracy theory engagement trajectory can be measured as a time series of C_u^i over ten deciles example, Figure 4.1 (b). What are the common patterns of conspiracy theory engagement? One way to find common patterns in users' conspiracy theory engagement is to perform unsupervised clustering of the user trajectories [105]. Longitudinal clustering of user trajectories could result in grouping together online users who have similar patterns of engagement in online conspiracy theory discussions. Figure 4.1 (c) displays initial clustering results where I identified 8 prominent patterns of longitudinal engagement in conspiracy theory discussions on Reddit. The Y axis represents the proportion of user activity invested in conspiracy theory discussion subreddits and the X axis represent the temporally ordered deciles in the users' Reddit lifetime. The clusters of trajectories presented in Figure 4.1 (c) indicate that there are four prominent patterns of engagement in online conspiracy theory discussions:

- **Steady high engagement (1 & 2 in Figure 4.1 c):** Users who consistently maintain high proportion of Reddit activity in conspiracy theory discussion subreddits throughout their lifetime.
- **Steady low engagement (4, 7 & 8 in Figure 4.1 c):** Users who consistently maintain low proportion of Reddit activity in conspiracy theory discussion subreddits throughout their lifetime.
- **Increasing engagement (2 in Figure 4.1 c):** Users who increase their activity in conspiracy theory discussion subreddits over time.
- **Decreasing engagement (5 & 6in Figure 4.1 c):** Users who decrease their activity in conspiracy theory discussion subreddits over time.

These results indicate that there are distinct patterns of longitudinal engagement in conspiracy theory discussion subreddits on Reddit. But how robust is this characterization of conspiracy theory engagement pathways? Next, I explain the measures for validating the conspiracy theory engagement pathways.

4.3.2.4 Validating conspiracy engagement pathways:

We invited 6 evaluators proficient in statistics and data analysis to manually assess the quality of conspiracy engagement pathway assignments in two annotation tasks.

1. **1. User trajectory labeling:** We asked the evaluators to label a user conspiracy trajectory plot—for example, the trajectory displayed in Fig. 4.1 (b)—as either *steady high*, *increasing*, *decreasing* or *steady low*. As instructions, we additionally provided C_u thresholds for each pathway and demonstrated sample trajectories from each pathway as a guideline for annotations.
2. **2. User-decile activity labeling:** We showed user contributions over subreddits in every other (1st, 3rd, 5th, 7th, 9th) decile (Fig. 4.1 (a)) and asked the evaluators to label the user activity by one of the four conspiracy pathways.

We randomly selected 12 users from each conspiracy pathways separately for each of the tasks. Each trajectory was labeled by two evaluators. We consider a true positive assessment for a trajectory only when *both* evaluators agree on the conspiracy pathway label. Evaluators labeled trajectories in task 1 with an accuracy of 78%, while 83% validation samples received perfect agreement. Task 2 resulted in accuracy of 84%

accuracy with a perfect agreement of 96%. The validation performance across both tasks suggests that the computational conspiracy pathway assignments are cohesive and can be inferred through both, user trajectory plot (task 1) and the user's raw activity data (task 2).

How do users on increasing engagement pathways differ from users who are on decreasing engagement pathway? Below, I briefly describe some of the theoretical models that are be used to contrast users with different types of engagement in conspiracy theory discussions online.

4.3.3 Models and measures for contrasting users on various conspiracy theory engagement trajectories

Users' behaviour after joining conspiracy theory discussions could be modeled across several dimensions such as in-group language, conspiracy epistemology, group connections. The RECRO model proposed by Neo [175], is a pathway-based theoretical model that views engagement into problematic content as an internet-mediated process involving all, individual, epistemic and social factors. Researchers have used RECRO in qualitative analyses of online anti-vaccination discussions finding that social media provides a strong platform for the first three phases [260]—Reflection, Exploration, Connection. Hence, we operationalize the first three phases of RECRO to contract user behaviors on different pathways of engagement into conspiracy theory discussions.

4.3.3.1 Phases of RECRO model

Reflection: The Reflection phase details the vulnerabilities and psychological predispositions that increase one's receptivity towards radicalization [175]. This is a phase where personality and psychological factors motivate the individuals to open-up, also described as "cognitive opening", to alternate belief systems. Other researchers also agree on the importance of psychological footprints such as anger and heightened emotions in online radicalization [64]. After the cognitive opening, users begin to form radical worldviews in the Exploration phase.

Exploration: Here, individuals begin to make sense of new information and narratives by forming alternate worldviews in a way that fosters eventual radicalization [175]. Individuals are primed to form a new, alternate worldview that resonates with

their interests and epistemological needs [175]. Specifically in relation to conspiracy theorizing, researchers propose a "monological belief system," describing it as a stable cognitive style that dictates the perceived functioning of the social world [109]. Monological conspiracy worldviews offer a general set of assumptions, such as cover-ups by powerful people, that are portable across multiple CTs and socio-political phenomena, independently from their specific topic or context [109]. This affords applying CT to any socio-political phenomena, independently from the specific topic or context of an event [94]. Hence, the monologicality hypothesis paints the picture of a closed-minded CT believer with a strong mobilization potential, and of a CT ecosystem of broadly applicable, interconnected, mutually supporting ideas. This hypothesis though is contested. Competing research presents a possibility of better educated, open, and socially active CT believers who might restrict their interests to specific conspiracy topics [94]. This work, for the first time, analyzes the generality or specificity of CT belief by modeling how individuals *explore* the world of online conspiracies after their initial exposure.

Connection: Here, individuals interact to form group bonds with like-minded people [175]. As opposed to the Reflection phase capturing individual predisposition, the Connection phase describes how bonds with a group of peers advance the radicalization process. Specifically, cohesion or conformity to one's social group [61], small-group dynamics [207] and feelings of group affiliation [64] are strongly associated with radicalization.

4.3.3.2 Characterizing conspiracy engagement through the phases of RECRO model

Characterizing Reflection Phase: The Reflection phase captures the psychological predispositions of users towards adopting radicalization narratives online [175]. Predisposition towards radicalization can be visible through the psycho-linguistic footprints left by the users online [64]. Specifically, researchers found that language reflecting anger, anxiety and heightened emotions was used by online radicalized groups [64].

Hence, to measure the language related to anger and anxiety, we use anger and anxiety lexicons, respectively, from Linguistic Inquiry and Word Count (LIWC) dictionary [245]. LIWC encodes words capturing affective, emotional and cognitive processing expressions and is often used for psycho-linguistic analysis of online texts. To measure emotionality, we calculate the average compound VADER sentiment scores [122]. In total, we calculate 3 linguistic features to characterize the Reflection phase.

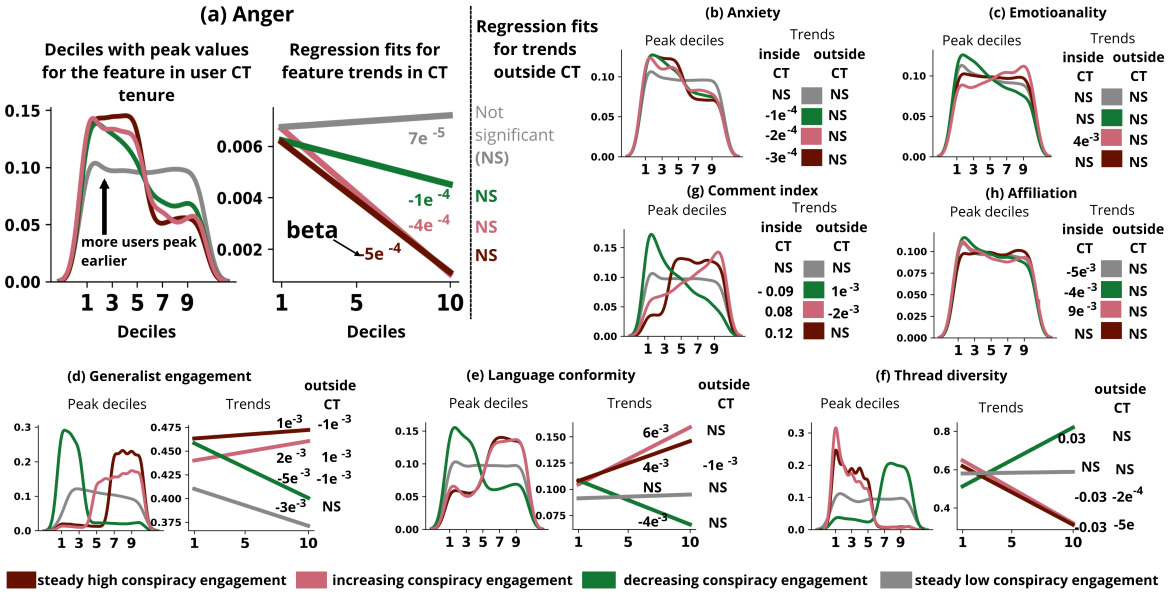


Figure 4.2: Figure presenting the peak deciles distributions and linear regression trends grouped by the conspiracy engagement pathways. In (a) we present an enlarged view of a typical result showing deciles with peak values and trends inside and outside of CT subreddits. For all features, peak decile distributions represent the density plots for deciles at which the users attain the highest feature value. For example, users on all pathways show highest values for anger in earlier deciles (subfigure a). Trend in each subplot represents the linear regression fit for various pathways over deciles. For every line we denote the β coefficient if the trend is significant. Non significant trends ($p > 0.05$) are denoted with NS. We show trend coefficients for feature calculated both, inside and outside CT subreddits. Due to space limitations we show the actual trend lines only for 4 features.

Characterizing Exploration Phase: The Exploration phase describes a period in which users develop alternate worldviews that advance the radicalization process. Specifically in conspiracy theorizing, scholars have debated whether conspiracy theory belief evolves into a monological worldview—a tendency to analyze all events through the lens of conspiracy theorizing [109]. Previous researchers have concluded that online discussions could be useful in understanding the users' conspiracy worldview [274]. To understand how online users explore the world of Reddit CTs, here we characterize conspiracy worldviews by calculating conspiracy generalist or specialist engagement. Specifically, we create a *generality scale*, that scores a subreddit based on the generality of topic discussions.

Intuitively, more general subreddits will host content that is less *exclusive* across Reddit. For example, `r/conspiracy` hosts political conspiracies on topics also dis-

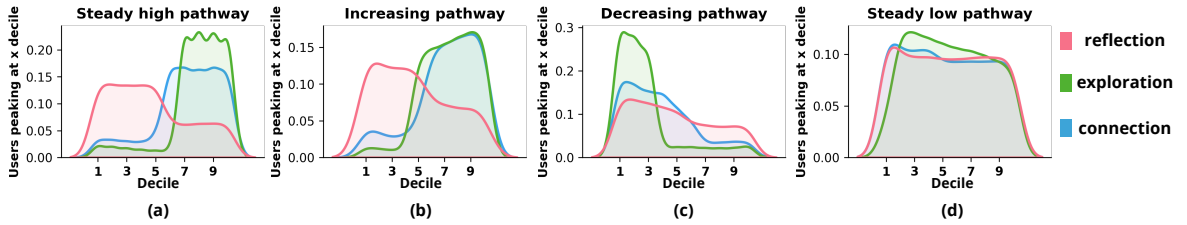


Figure 4.3: Figure outlining peak deciles for features in various RECRO phases for (a) *steady high* (b) *increasing* (c) *decreasing* and (d) *steady low* pathways. Peak in earlier deciles means that users attain highest values for that phase, early on.

cussed in `r/politics` and medical conspiracies on topics in `r/science`. However, subreddits such as `r/moonhoax` host specific conspiracy topics, that are less likely to be popular across rest of the Reddit. We leverage this intuition and build "subreddit-entity network".

Subreddit-entity network is a graph in which subreddits are the nodes connected based on common content. To find the content representative of a subreddit [120], we analyze the top 200 submissions made in every subreddit and extract named entities (names of people, places, organizations etc.) from the submission text. We create an edge between two subreddits if top posts in both subreddits mention the same entity. To further improve the quality of edges, we consider the inverse term frequency of the shared entity across the entire corpus. The total edge weight between two subreddits is the sum of inverse term frequency of the entities shared between the subreddits.

In the subreddit-entity network, intuitively, more generalist subreddits will share more entities with other subreddits. In other words, generalist subreddits might be in a densely connected neighborhood connecting various sub-graphs. Accordingly, generalist subreddits may be *influential* in the subreddit-entity network. Hence, to assess the degree of generality of a subreddit, we calculate the eigenvector centrality for all subreddits which is a measure of influence, where a node is considered to be influential if it is connected to other influential nodes. We consider the eigenvector centrality as the subreddit generality score where subreddits with higher eigenvector centrality hosting more general discussions.

With the *generality scale* we measure whether users increasingly engage in more generalist or more specialist CT discussion subreddits. Using the same computation as the conspiracy engagement score, we calculate the generalist engagement score in every decile as the weighted average of a user's contributions in subreddits weighted by the subreddits' generality score. Higher generalist score would indicate high engagement in subreddits with general discussions.

Characterizing Connection Phase: In the Connection phase, individuals form social connections to support and reinforce their alternate worldviews, facilitating the relational bond between an individual and a wider radical movement [175]. To characterize the this phase, we capture language conformity and group connections established by users with conspiracy communities.

Cohesion or conformity with one's social group is a fundamental requirement in the process of radicalization [61]. Especially in CT discussions, groups conform by establishing shared interpretations of reality around them [39]. This process of interpreting reality, or meaning-making, often manifests into the insider language used by conspiracy groups [152]. Hence, we measure language conformity by assessing how much of the subreddit's characteristic language does the user use.

To measure a user's language conformity, we consider each subreddit the user contributes in as her "social group" and calculate the overlap between the language used by the user and that subreddit's characteristic language. To understand the characteristic language for each subreddit, we utilize Sparse Additive Generative models (SAGE) [81] that uses a regularized log-odds ratio to compare word distributions across various text corpora. We compare the word distributions in the text corpus of each subreddit with that of all other subreddits using SAGE. As a result, for each subreddit we obtain a lexicon of 1000 words that distinctively represent the language used in that subreddit. Finally, we calculate a user's language conformity in a subreddit s as the intersection of words used by the user in subreddit s with the SAGE lexicon of the subreddit, normalized by total word count used by the user in s.

Furthermore, interactions within small groups of like minded people can help in creating unambiguous shared narrative of events that is instrumental to the process of radicalization [207]. Discussions within small groups can also limit the number of diverse opinions and information users might get exposed to, thus contributing to what researchers call as "crippled epistemeology" [264]. Hence, we characterize users's small group interactions by analyzing the diversity of audience and repeated contributions in comment threads populated by users.

4.3.3.3 Contrasting various pathways of engagement through RECRO phase operationalizations

1. **Linear regression fits for trends:** For each user on each conspiracy engagement pathway, we have ten values of all features corresponding to every decile. To understand how features evolve over time, we fit a linear regression line with

deciles as the independent variable and the feature value as the dependent variable. The magnitude of the fit coefficient (β) tells us the degree of increase and decrease over time and the p-value indicates whether the fit is significant. We display all trends lines and coefficients inside the 171 CT subreddits in Fig. 4.2 along with the coefficients for trends *outside* of conspiracy subreddits as well. Trend coefficients and significance outside CT show whether the trends we observe inside are specific to conspiracy discussion.

2. **Users' conspiracy tenure with peak feature values:** We characterize phases of radicalization that, in theory, take effect one after another. Hence we next analyze how soon after CT joining, users attain peak values for features. For every user, we pick the decile with highest feature value and plot the density distribution of peak decile for all users grouped by conspiracy pathway. Density peak in early deciles would mean that users attain highest value for that feature immediately after initial participation in conspiracy discussions.

3. **Users' conspiracy tenure with peak phase features:** While the previous analysis displays how individual features peak across decile, here we group the feature belonging to same radicalization phases and plot similar peak density plots for each pathway (Fig. 4.3). For example, Fig. 4.3 (a) indicates that for *steady high* pathway, Reflection features peak immediately after initial conspiracy participation whereas Exploration and Connection peak later in the CT journey. Visualizing this phase progression can inform whether users develop RECRO phases successively in time.

4.3.4 Results: Contrasting conspiracy engagement pathways

4.3.4.1 How do users on display markers of reflection phase?

We characterized the Reflection phase using anger (Fig. 4.2 (a)), anxiety ((Fig. 4.2 (b)) and emotionality (Fig. 4.2 (c)), expressed inside and outside of CT subreddits. Overall we find that use of language related to anger and anxiety decreases over time for users on all pathways. There are no significant trends for emotionality inside or outside of CT, except for users on the *increasing* pathway who show increasing emotionality inside CT over time ($\beta = 4e^{-3}$). Earlier peaks in Reflection features may indicate what Neo [175] describes as "cognitive opening" where individuals turn to internet to express their grievances and vulnerabilities. Do all users advance to subsequent phases of

radicalization after the cognitive opening? To find out, we next analyze the results of the Exploration phase.

4.3.4.2 How do users explore Reddit's conspiracy world?

To characterize the Exploration phase, we investigate how users develop conspiracy worldview through generalist engagement (Fig. 4.2 (d)). Higher generalist engagement would indicate that users engage in general CT discussions, thus developing monological worldview. We find that users on *steady high* ($\beta = 1e^{-3}$), and *increasing* pathway ($\beta = 2e^{-3}$) increasingly participate in the generalist CT subreddits. Interestingly, *steady high* users show *reduced* generalist engagement outside of CT subreddits ($\beta = -1e^{-3}$). Conversely, users on *decreasing* pathway increasingly contribute in specialist CT discussions ($\beta = -5e^{-3}$) and have highest generalist engagement only in the earlier deciles. We ran an additional robustness check to ensure that this result is not an artifact of correlation between *conspiracy similarity scale* and *generality scale*. Overall, we find that users with *steady high* and *increasing* CT engagement may also adhere to monological conspiracy worldview by increasingly participating in general conspiracy discussion subreddits.

4.3.4.3 How do users make connections inside conspiracy communities?

We measure group bonding through language conformity (Fig. 4.2 (e)), audience diversity in threads (Fig. 4.2 (f)), repeated comments (Fig. 4.2 (g)) in threads, and affiliation related language (Fig. 4.2 (h)). Overall, we find that users on *steady high* ($\beta = 4e^{-3}$) and *increasing* ($\beta = 6e^{-3}$) pathways develop high language conformity with CT subreddits. However, users on *steady high* pathways show reduced language conformity outside of CT subreddits ($\beta = -1e^{-3}$). Interestingly, users on *decreasing* pathway ($\beta = -4e^{-3}$), despite exhibiting early high engagement, never develop as high language conformity with CT subreddits in comparison to the other cohorts. Hence, early lexical conformity could be one of the important precursor of sustained CT engagement. Users with *steady high* engagement also participate repeatedly (β=0.12) in smaller discussion groups with less audience diversity (β = -0.03). Users on *increasing* pathways also show similar trends. These results suggest that users on *steady high* and *increasing* pathways repeatedly show engagement in discussions with less diverse user base inside CT subreddits.

4.3.4.4 How does conspiracy radicalization evolve?

Fig. 4.3 shows the deciles in which users attain peak feature values in different phases. We observe that in *steady high* (Fig. 4.3 (a)) and *increasing* (Fig. 4.3(b)) pathways, users show higher feature values in the Reflection phase right after starting CT participation and develop high Exploration and Connection feature values in later deciles. This may suggest that the internet-mediated conspiracy radicalization does evolve through different phases over time. Interestingly, users on *decreasing* pathway show high feature values for all phases only early on, while for users on *steady low* pathway, there is no discernible peak for these phases. We discuss the implications of these results in the discussion section.

4.4 RQ2: QAnon disengagement and cognitive dissonance

The previous research question investigated how users on decreasing conspiracy theory engagement pathway differ from those on the increasing pathway. This section dives deeper into understanding the possible reasons for disengagement based on cognitive dissonance theory explained in the Background section. To characterize users' self-disclosures of dissonance, we first understand users' social imaginaries using 4chan and Reddit data. Below, I start with explaining the data collection process.

4.4.1 Datasets

4.4.1.1 Q-drops Dataset: Q-drops from 4chan and 8chan

In the beginning of the QAnon conspiracy, Q posted messages (Q-drops) on image boards, and the followers discussed Q-drops across various Reddit communities. In this RQ, we first analyze Q-drops to characterize social imaginaries established by the leader Q for the QAnon community. Specifically, Q drops were made on 4chan and 8chan boards that are anonymous, ephemeral forums revolving around posting images along with text. 4chan and 8chan are infamous for hosting controversial content resulting in multiple temporary bans. We use `qalerts.app`, a website that aggregates Q-drops from image boards. While the exact agency of `qalerts.app` is unknown, it is a common resource used by QAnon communities[2] and also by other researchers

[2]See QAnon and the Great Awakening group on Gab.com

studying QAnon [5]. There are several other Q-drop aggregation sites, however, most of them contain nearly similar record of Q posts (see Table 1 in [5]). We downloaded the first 2166 Q-drops that were made between October 2017 and September 2018 to allow for consistent time period between the analysis of social imaginaries and cognitive dissonance.

4.4.1.2 QAnon Subreddits Dataset: Reddit Dataset of 12 Banned QAnon Communities

After the emergence of Q on 4chan in September 2017, the QAnon movement popularized on Reddit [33]. To reach a more mainstream audience, prominent QAnon followers created a subreddit called r/CBTS_Stream, short for *Calm before the Storm*–a popular saying in the QAnon community indicating impending arrests of the deep state politicians. However, r/CBTS_Stream was banned in March 2018 for violating Reddit's terms of content policies. This ban resulted in the creation of new subreddits such as r/greatawakening2, r/BiblicalQ and others that combined accrued more subscribers than the original r/CBTS_Stream [34]. Finally, 17 of these new communities were also banned by Reddit in September 2018 for repeated violation of content policies [34]. We identified the 17 banned QAnon related subreddits from various press mentions [34, 212]. To obtain the data for banned subreddits, we used Reddit Pushshift Dataset[3] [20]. Specifically, we queried the Pushshift data from Google Bigquery and obtained the submissions and comments from 12 of the 17 banned subreddits. The data for the rest of the 5 subreddits is not present on Pushshift nor through the official Reddit APIs. In total, we have 96,068 submissions and 1,104,096 comments made by 33,561 users across 12 subreddits.

4.4.2 Characterizing QAnon social imaginaries: The QAnon Canon

4.4.2.1 Iterative Inductive Coding

After collecting words and phrases from Q-drops dataset, we began to organize them into themes based on the semantic relationships. Two authors of this work were involved in this inductive analysis. Our focus was on grouping words that represent similar meanings in the QAnon ideology. For example, the words "boom", "moab" (mother of all booms), "april showers", "red october" all generally allude to a future event, significant to the take down of the deep state. Similarly, the words

[3] `https://files.pushshift.io/reddit/comments/`

"anons", "bakers", "autists", "patriots" are all used to refer to insiders of the QAnon community. Note that to understand the insider language of QAnon, we refer to the words cross-listed across various Q-drops and also other online resources such as list of abbreviations on `qalerts.app` and various news articles describing the Q-drops and QAnon language [60, 248]. We iteratively developed the categories through a discursive process. We found saturation at five categories that capture the social imaginaries of the QAnon community. Below, I explain the five dimensions of QAnon social imaginaries—*movement*, *expectations*, *practices*, *heroes*, and *foes*—resulting from the qualitative analysis along with examples.

4.4.3 Five Dimensions of QAnon Social Imaginaries

1. **Movement:** The *movement* category signifies the collective identity of the QAnon community. Movement includes Q-team (a group of anonymous people believed to be working with Q), Q research team (a group of Q followers that organize and research the Q-posts) and several other designations (anons, bakers, patriots) that collectively represent the Q-followers. Slogans such as WWG1WGA (Where we go one we go all), and WRWY (We Are With You) are used to reinforce faith and motivate collective action in the QAnon movement.

2. **Expectations:** Through the promises of arrests of deep state agents, Q sets expectations for their followers. Expectations are about both good and bad events in the context of the QAnon community. For example, several Q-drops predict arrests of specific deep state politicians while others warn the readers about false flag events, forecasting covert operations of various governments and cabals.

3. **Practices:** An important part of QAnon community is hunting for clues provided by Q. Q instructs their followers to follow certain knowledge construction practices. For example, Q asks their followers to organize and archive Q-posts, connect the dots and dig for the truth.

4. **Foes:** Q routinely releases the names of celebrities, politicians and law enforcement agents who are purportedly associated with the deep state. Deep state agents are believed to be involved in a satanic cult with an international child sex trafficking ring. Simply put, foes of QAnon are portrayed as the enemies of the QAnon movement.

5. **Heroes:** According to Q, while the law agencies, media and a large part of the government is considered to be controlled by the deep state, there are some "good guys" who fight for the American people. Q, Donald Trump and former U.S Attorney General Jeff Sessions are at the top of this list. Heroes often know more than they choose to reveal to the QAnon community for reasons of national security and are believed to be experts at undercover work.

As a byproduct of the content analysis process, we recorded relevant words and phrases in each of the five dimensions of the QAnon social imaginaries. We refer to this as a **seed lexicon** for QAnon social imaginaries. In the next section, we understand how QAnon followers communicate the social imaginaries established by Q, by quantitatively expanding the seed lexicon.

4.4.3.1 Building QAnon Canon

We quantitatively expand the seed lexicon into QAnon Canon—a dictionary capturing words and phrases in each of the dimension. We use the QAnon subreddits dataset of 1.2M comments and posts to understand various ways in which phrases from the seed lexicon are expressed. We utilize rigorous quantitative methods that use sentence parsing and semantic similarity of words to expand a seed set of 75 phrases to over 403 phrases. Finding various expressions for phrases from public discourse is a challenging task. Specifically, pronouns are often used to refer to the noun phrases (ex. using 'she' instead of 'Hillary'). This is called *coreference*. In order to find similar phrases, we first need to resolve the coreferences. After resolving the coreferences, we characterize meanings of various phrases as vectors and use interactive mixed-methods approach to find various expressions of phrases in the seed lexicon.

By iteratively finding phrases similar to the seed lexicon, we obtained QAnon Canon—a lexicon of 403 phrases encoding how QAnon followers communicate QAnon social imaginaries. We are able to recover various expressions for named entities. For example, QAnon Canon contains multiple expressions for Hillary Clinton (hillary, HRC, HC, killary, billary, alice in wonderland, clintons) and Barack Obama (Obama, Hussein, Obamas, ObamaHillaryCIA, Barack, Obummer, 0bama). Similarly we were also able to find lexically and semantically similar expressions of various phrases symbolizing the movement (q-team, q-analyst, q-research, q-clearance), practices (q-drops, qdrops, q drops, qposts, q-posts) and expectations (layoffs, mass exodus). We

have made QAnon Canon publicly available for other researchers to use [4]. In the next steps, we use words from QAnon Canon as linguistic features for identifying expressions of belief and dissonance in the QAnon community.

4.4.4 Compiling Factors in Self-Disclosure of Belief and Dissonance

Referring to the QAnon Canon and prior literature on expressions of belief and dissonance, we compile lexical, stylistic, and document level features for classifying belief and dissonance in QAnon. We prefer hand-picked, theoretically motivated features over pre-trained sentence embedding models to preserve the interpretability of various features in identifying belief and dissonance. We discuss the importance of different features in Section 4.4.6.2.

1. **QAnon Canon (403 features):** The lexicon, created in the previous chapter, captures the social imaginaries of the QAnon community, which can be elicited in affirming belief [9]. Similarly, disagreement with the social imaginaries can signal dissonance with the QAnon worldview. Hence, we calculate the frequency of each of the 403 phrases in the QAnon Canon in user comments.

2. **Linguistic Inquiry and Word Count (LIWC) (19 features):** LIWC encodes words that capture affective, emotional and cognitive processing expressions and is often used for analysis of online texts [245]. For example, LIWC categories of tentativeness (includes words such as *maybe*, *perhaps*) and certainty (*always*, *never*) were specifically found to be relevant in the expressions of doubts in online reviews [83]. Following a similar rationale, we include 17 other relevant LIWC categories[5]. For example, we include conjunctions (*and*, *but*, *whereas*) that may be present in arguments that combine contradictory claims ("I am trying to trust Q *but* my patience is running out") [71].

3. **Integrative Complexity (IC) Score (1 feature):** IC is a psychometric that captures people's ability to recognize multiple perspectives and connect them together [242]. IC is closely related to expressing belief and attitudes [62]. IC scores range from 1 to 7, where 1 indicates no evidence of IC and 7 indicates the presence of

[4] `https://social-comp.github.io/ConspiracyTraces/`

[5] feel, discrepancy, second person pronouns, differentiation, religion, third person singular pronouns, causation, first person pronouns, anger, hear, third person plural pronouns, insight, sadness, see, negations, conjunctions, anxiety

overarching perspectives with detailed connections. We calculate the IC score using the model published by Robertson [210] [210].

4. **Credibility Cues (8 features):** Perceived credibility and the evaluation of the common knowledge are strongly associated with belief and disbelief [170, 201]. We include credibility cues such as booster words (*actually*, *evidently*), hedge words (*in my view*, *in general*), modal words (*hypothetical*, *improbable*) and evidentials (*know*, *guess*) that are associated with perceptions of credibility [170]. We also calculate sentiment scores [30] and number of quotations [69] in a comment that might indicate uncertainty. Additionally, we include number of questions that might signal information needs [172].

5. **Community Feedback (2 features):** On Reddit, comments that are well-received by the conspiracy community are awarded upvotes while ill-received comments get downvotes [185]. In particular, we expect comments expressing dissonant views to receive negative community feedback. Hence, we calculate the comment score (an aggregation of upvotes and downvotes). To contextualize the feedback received, we also compute the synchronicity of the comment score with the score of its parent comment. To calculate synchronicity, we subtract the parent comment score from the child comment score.

6. **Generic Document Level features (50 features):** Finally, we calculate *smooth inverse frequency* (SIF) document embeddings that capture the overall semantics of a comment by combining the embeddings of its words [10]. We calculate 50 dimensional SIF embeddings and use them as the baseline for evaluating the features discussed above.

In total, we calculate 483 features for every Reddit comment or post in the 12 QAnon subreddits. Next, I describe various sampling methods used to create a labeled dataset for classification.

4.4.5 Creating a dissonance classifier

4.4.5.1 Creating a Labeled Dataset

To understand belief and dissonance at scale, we need a large labeled dataset—ideally, the whole 1.2M comments in the study. We rely on a smaller, high quality labeled

dataset that is manually vetted and use a high-precision classifier to extend the labeling to the rest of the data. Yet, even annotating the smaller dataset is challenging: labeling expressions of belief and dissonance in a community like QAnon is a complex and nuanced task which requires sizable theoretical background and expertise with the QAnon social imaginaries. Thus, we cannot rely on crowdworkers for the task. Moreover, self-disclosures of belief and dissonance are rare occurrences, with the great majority of the QAnon subreddits revolving on discussing details of the theories and phatic talk, posing technical challenges to sampling informative instances of belief and dissonance to label. We address these challenges using a expert-in-the-loop mixed-methods approach [76].

A common strategy to make the most out of constrained labeling resources is active learning [223]. It lets an *interim* classifier choose which next comment would be the most informative if labeled, given the ones already labeled. The intuition behind it is that many comments are similar to each other (e.g., phatic comments making up the majority of the discussions), and labeling multiple instances would not offer a downstream classifier any new information. Instead, labeling a diverse set of comments would better serve the classifier to explore the variety in the whole subreddit. Especially, the comments about which classifier is the most uncertain at any point, are the ones that, if labeled, would most likely help it discern between classes in the future. Hence, given a pool of comments, active learning selects the most helpful one for an expert to annotate, and re-trains itself adding the newly annotated comment. This procedure repeats until the classifier performance converges, or until the annotation resources (the experts) are exhausted.

Through a pilot annotation, we found that the classes of interest—belief, dissonance, and neutral comments—are extremely imbalanced. Comments expressing disbelief, especially, amounted to only 6% of the comments in a random sample, while neutral comments amounted to 80%. Hence, we differentiate the pools of candidate comments to feed into the active learning loop, to trade off between exploring the large variety of comments in the whole dataset, and labeling a meaningful number of belief and dissonance comments. We select two distinct pools of comments: the first sampled at *random*, to represent data variety; the second sampled with a *biased* strategy to surface a higher number of instances of belief and dissonance. We perform active learning on each pool separately, and then use the comments labeled in both active learning runs to train the final classifier and extend the labeling to the whole dataset.

For the final classifier, we experimented with different aggregation techniques,

from combining the labeled data and training a single classifier, to training classifiers separately on each labeled datasets and combining their predictions into a single score. The latter approach performs best on a held-out validation set. Next, we discuss the details of sampling the random and biased data pools, performing active learning, and training the final classifier.

4.4.5.2 Creating Pools of Unlabeled Data

Because of the inherent disproportion of neutral comments in comparison to those expressing belief and dissonance in the QAnon subreddits, classifiers trained on such data may be less accurate on the latter classes [13] even after training on large labeled data [40]. In order to accurately model belief and dissonance in QAnon, we need a higher proportion of labeled instances of such classes. At the same time, to build a classifier that generalizes well to real data distributions, we also need labeled samples that represent the overall dataset. Hence, we first create two unlabeled sample pools—*random samples* and *biased samples*.

1. **Random Unlabeled Sample:** Random sample contains 100K comments and posts selected uniformly at random. Random sampling can be representative of the dataset [149] where various types of expressions of belief and dissonance can occur at their natural frequencies inside the QAnon subreddits.

2. **Biased Unlabeled Sample:** We use a cluster-based sampling technique to include belief, dissonance and neutral expressions in similar quantities [279]. Specifically, we perform K-Means clustering on the entire dataset and select samples that are closest and farthest from each cluster centroid. We use the elbow method—plotting explained variance in clustering as a function of number of clusters—to determine optimal number of clusters as 3. From every cluster, we collect the 20K closest and farthest samples to each centroid, for a total of 120K posts and comments.

We use these two pools of unlabeled data to select the samples to label.

4.4.5.3 Labeling Dataset with Active Learning:

Combined, the random and biased data contains 220K comments, which are still too large for complete manual labeling. To trade-off between the manual labor of labeling many comments and the downstream classification performance, which depends on a

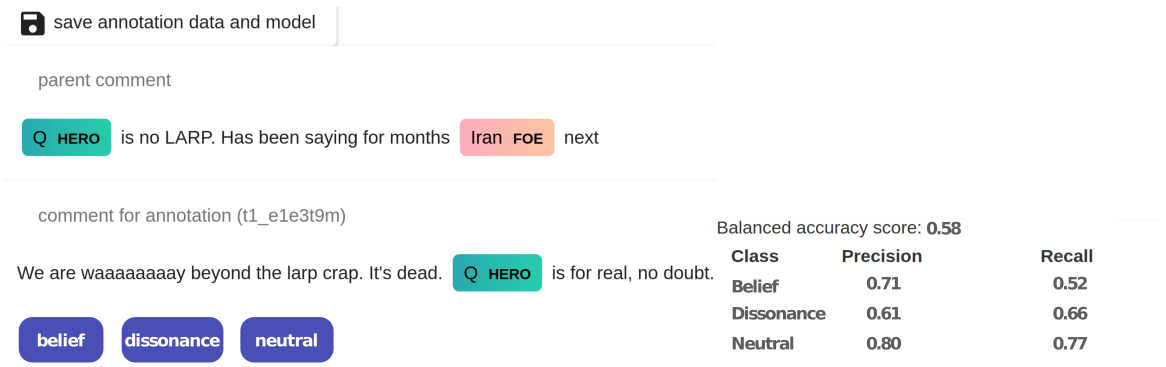


Figure 4.4: Figure showing the dissonance annotation interface. Every comment and its parent comment is annotated with categories of QAnon social imaginaries. Annotators label every comment as belief, dissonance or neutral. The interface also displays naive classification scores updated every 10 samples.

large enough labeled dataset, we devise an interactive labeling process based on active learning. An active learning classifier selects the next unlabeled comment to label according to its sampling strategy; after annotators label the comment, the classifier adds the comment to the labeled dataset, retrains itself, and selects the next comment to label [184]. This process repeats until the satisfactory classification performance is achieved. The crux of the learning strategy lies in sampling the new data to be labeled. We use a popular sampling strategy uncertainty sampling [119].

4.4.5.4 Building Classifiers from Labeled Data

We use this labeled data to develop a classifier that reliably identifies expressions of belief and dissonance in the whole QAnon subreddits. In this section, I explain our selection of the classifier model, and details about its training and prediction procedures.

Selecting the Classifier Model: Although we labeled a total of 2,371 comments and posts from the random and biased samples, the labeled dissonance instances were still fewer compared to the belief and irrelevant. Since class imbalance may bias classifier performance, we balance the data. We undersample the data while maintaining above 0.6 precision for all classes. We experimented with various undersampling strategies and found the best performance with one-sided undersampling [146]. One-sided sampling removes the noisy, under-performing examples from the majority classes while preserving all the examples of the minority class. After undersampling, we end up with 336 and 900 samples from the random and biased labeled datasets respectively.

To choose best classifier we use the `auto-sklearn`[6] toolkit that searches over a wide variety of models optimizing for performance [91]. We find the optimal model for each of the two labeled datasets—namely, a Random Forest model [117] for the random labeled data and an Extreme Gradient Boosting (XGBoost) model [48] for the biased labeled data.

Training and Predicting with Classifier Models: We perform hyperparameter optimization to tune model parameters. Different models require tuning of internal parameters such as learning rate, estimators, etc., to generalize to unseen data [52]. We determine the best model parameters through an extensive grid search in a cross-validation scheme. We test over 15,000 combinations of hyperparameters, optimizing for chance-adjusted, balanced classification accuracy.

After fine-tuning both classifiers, we combine their predictions on the whole dataset. Specifically, we use a strict consensus pooling method to determine class assignments. Consensus pooling assigns a particular class (belief, dissonance, or neutral) to a sample if *both* classifiers agree on the predicted class. Disagreements are removed from the predictions. We also experimented with different pooled prediction strategies that do not require removing the ambiguous samples with slightly lower scores. A single classifier trained on the combination of the two labeled datasets performs significantly worse.

4.4.5.5 Characterizing the types of Dissonance Self-Disclosures

Using the classifiers designed in the previous steps, we label the dataset of 1.2M comments and posts. We remove around 500K samples with prediction disagreements. In the remaining automatically labeled dataset, we find that over 43K comments and posts (6%) are labeled as dissonant. What are the different ways in which users express dissonance? We qualitatively analyze a random sample of 500 comments and posts expressing dissonance, focusing on how dissonance relates to the dimensions of QAnon social imaginaries, such as how they refer to collective practices and expectations. We report the results of the qualitative analysis in the next sections

	Random Forest (RF) Classifier				XGBoost Classifier (XGB)				RF+XGB Consensus	
	training		validation		training		validation		validation	
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall
belief	0.61	0.46	0.55	0.41	0.60	0.54	0.66	0.58	0.71	0.54
dissonance	0.62	0.76	0.52	0.69	0.60	0.59	0.60	0.64	0.70	0.76
neutral	0.71	0.71	0.67	0.69	0.64	0.61	0.70	0.73	0.79	0.79
balanced accuracy	0.66		0.60		0.63		0.69		0.79	

Table 4.1: Training and validation performances for individual classifiers and final consensus classifier. The final consensus classifier clearly outperforms the individual classifiers across all precision scores.

4.4.6 Results: Dissonance classifier

4.4.6.1 Classification Performance

Table 4.1 shows the training and validation performance of the individual classifiers as well as the ensemble classifier consensus-pooling their predictions. Individual classifiers have comparable performances. RF shows 0.66 accuracy in the training set (cross-validated) and 0.60 accuracy on a held-out validation test, while XGB 0.63 and 0.69 respectively. The consensus classifier clearly outperforms both RF and XGB individually. In particular, precision exceeds 0.7 for all three classes. We further include distribution plots for validation precision and recall over 100 training iterations. The distributions show the mean precision/recall values along with the standard deviations. Overall our results show relatively stable precision and recall values, lessening potential concerns of model overfitting.

4.4.6.2 Important Indicators of Belief and Dissonance

Which features are most impactful in classifying belief or self-disclosures of dissonance? Usually, important features in the model can be interpreted using classification coefficients. Such model explanations are easiest in binary classification where one coefficient can be interpreted in terms of either of the two classes. However, our classification consists of three classes: belief, dissonance and neutral. We need to understand how different features are impactful in classifying *each* of the classes. To understand feature importance, we use Multitask Elastic Net model, which solves three regression problems with three classes while sharing the same feature space. Hence, we can get feature importance separately for every class. For example, positive feature coefficients

[6] `https://github.com/automl/auto-sklearn`

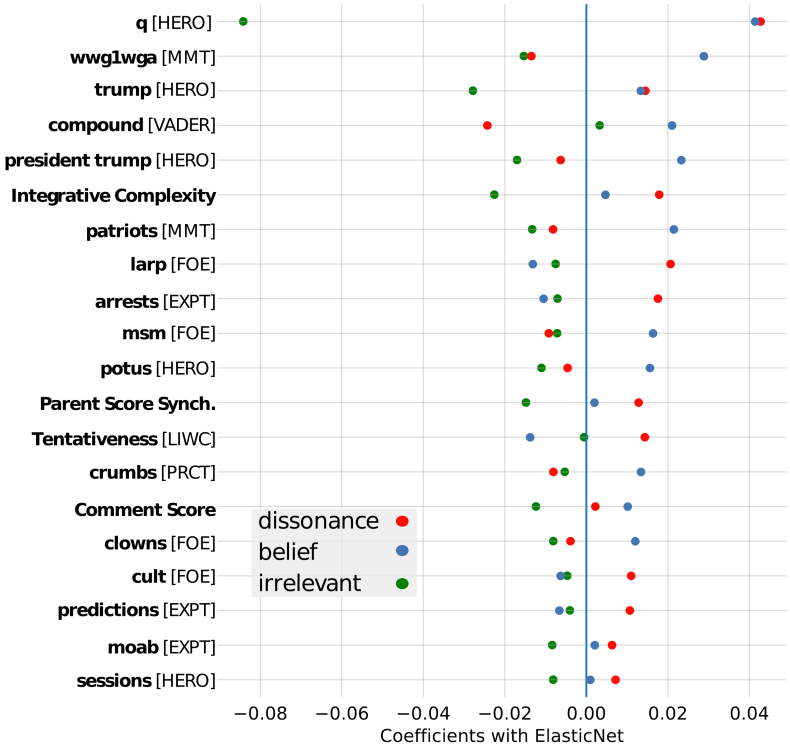


Figure 4.5: Plot showing most predictive features for each class. Negative coefficients indicate that the low value of feature is associated with the presence of the class. For example "wwg1wga", a QAnon movement related word, is present more in the comments expressing belief (0.03 coefficient for belief) whereas dissonance and neutral comments have less occurrences of the same word (-0.015 for both). Note that the importance of features was separately calculated for each class using a Multitask Elastic Net model. Hence, the coefficients for features for each class are independent of the others. We annotate each feature with the feature category whenever applicable. Abbreviations for QAnon Canon features are; HERO:heroes, FOE: foes, MMT: movement, EXPT: expectations, PRCT: practices.

for belief class does not imply negative coefficient for the dissonance class. Figure 4.5 displays the most important features per class. Negative coefficients indicate that the low value of feature is associated with the presence of the class. For example, mention of "Q" is positively associated with both, belief (0.04) and dissonance (0.04) whereas, neutral comments have low mentions of "Q" (0.08). Interestingly, while both belief and dissonance have higher mentions of "trump" (0.01 and 0.01), only comments with belief mention "president trump" (0.02) indicating respect for QAnon *heroes*. Comments expressing belief also have higher proportion of *movement* related words ("wwg1wga" (0.03), "patriots" (0.02),) and higher positivity (compound VADER (0.02)). Dissonance disclosures however, mention more *expectations* related words

Legitimacy of the movement	*...training your FOLLOWERS to put your voice before Q's - as we see from some - that can definitely be a problem ....this makes me question if this whole thing is even worth it..*
	i knew this movement was going to shit when the "Cult of Q" occupied this sub. YOU NEED TO CHILL THE F' DOWN
Unfulfilled expectations	*Fool me once shame on Q, fool me twice.. Been following Q from beginning. How many failed predictions does it take before you realize its bs?*
	Ok, I'm getting off the Q-Anon train right here. He has consistently said, "Big news week", "Past Unlocks the Future", "Have Faith, Patriots are in Control", but nothing ever happens ! We have all been strung along like lemmings.
Ineffective practices	*this is just ridiculous. Q sends us down the rabbit hole, asks us to "dig deeper" but nothing makes sense!*
	In all fairness, Q has only dropped vague crumbs and we are supposed to find the truth off of that? Is it too much to ask for more proof?
Distrust in heroes	*I'm seriously starting to doubt Q is a White House insider. Most of what he posts is easily found in the news and on conspiracy sites.*
	Trump staff found holding devils signs in their hand. I think trump is controlled too by the jews/nwo..
Trust in foes	*Without more evidence than a few vague posts from Q, I'm not going to believe that Barack Obama was sexually abusing that girl.*
	I disagree with Q on the AJ [Alex Jones] matter. I have listened to infowars for years. What's the point in alienating people like Alex who are clearly on our side

Table 4.2: Types of dissonance related to social imaginaries in QAnon. We conducted qualitative analysis of predicted dissonant comments. This table presents examples related to main observations.

("arrests" (0.02), "predictions" (0.01), "moab" (0.01)). Dissonance self-disclosures also have higher integrative complexity (IC) score indicating the presence of multiple argumentative perspectives. In general, we find that semiotic language captured by QAnon Canon is strongly indicative of belief ("wwg1wga", "president trump", "patriots", "msm", "potus", "crumbs", "clowns", "obummer" etc.) and dissonance ("larp", "arrests", "cult", "sessions", "moab" etc.).

4.4.6.3 Points of Dissonance in QAnon

In the qualitative analysis of the dissonant comments, we examined how dissonance is expressed along the social imaginaries of QAnon. Table 4.2 lists various points of fracture in the QAnon social imaginaries along with example comments. Several users express dissatisfaction with various components of the QAnon movement. For example, some comments expressed how YouTubers profiting off of QAnon movement could harm the unity. Moreover, some users expressed concerns with overzealous nature of and the credibility of the others in the QAnon movement. Next, disappointment over Q's failed prophecies and unfulfilled promises is one of the primary point of dissonance for some users. Several users refer to specific phrases used by Q to express dissatisfaction over unmet expectations. While some users doubted the effectiveness of knowledge construction practices instructed by Q, distrusting the legitimacy and power of heroes is perhaps the most common point of dissonance in the QAnon followers. Some users express concerns over the mysterious identity of

Q. We did not find many comments explicitly defending the enemies of the QAnon movement. However, some users expressed concerns over lack of evidence for vilifying deep state politicians. Some also expressed disappointment when Q de-legitimized popular right wing celebrities such as Alex Jones.

4.4.7 User Engagement after Dissonance Self-Disclosure

Dissonance can induce various behavioral changes, such as strengthening commitment [89], recruiting others [88] or even reversing one's belief [89, 164]. Thus, we study how users change their engagement patterns within the QAnon subreddits after expressing dissonance.

4.4.7.1 Observing Changes in User Engagement After Dissonance

We use Interrupted Time Series (ITS) analysis to characterize user contributions before and after dissonance self-disclosure. ITS is a quasi-experimental statistical method that is used to analyze change in longitudinal data after an intervention or policy change. For example, researchers have used ITS to observe how dramatic events change user engagement in Reddit conspiracy communities [216]. Here, we consider dissonance self-disclosure as the *intervention* that determines changes in user contributions. With ITS, we can characterize whether, and by what degree, the trends in contributions after the intervention differ significantly from before. ITS analysis involves solving the following linear regression:

$$contributions \sim b_0 + b_1 T + b_2 D + b_3 P \tag{4.2}$$

where T represents the time step of the observation, D is a binary variable representing whether the time step is before or after the intervention and P encodes the time steps after the intervention. For example, in a model with time step of one week and an observation window starting 2 weeks prior to the intervention, the 2nd week after the intervention will be encoded as $T = 5$ (5 weeks since start of the observation window), $D = 1$ (after intervention), $P = 3$ (3 weeks after intervention, including the week of intervention). b_0 is the model intercept. The coefficient of T, b_1 indicates the slope of trend in the outcome variable (contributions in this case) *before* the intervention. b_2 indicates the change in level *starting at* the intervention whereas b_3 indicates the change in slope *after* the intervention. Therefore, the actual slope of trend after

the intervention is derivable adding b_1 to b_3. The ITS regression directly indicates whether the pre-intervention trend b_1 and change in level at the intervention b_2 are statistically significant. Since ITS does not model directly the actual slope of trend after the intervention, but only the change with respect to the trend before, we corroborate its statistical significance via a separate piece-wise linear regression.

4.4.7.2 The ITS Setup:

Given that the QAnon communities were banned within 11 months of their creation, a week is an appropriately short time-step to measure the immediate effects of the intervention. We define an observation window of total 13 weeks, centered at the intervention[7]. We compute the number of contributions (comments and posts) that users made each week in the QAnon communities, normalized by the number of contributions throughout their lifespan. In other words, each observation shows what fraction of contributions users made in QAnon communities within that specific week. Normalizing this way allows us to compare all users' contributions from the scale of 0 to 1. We employ several other robustness measures to ensure that users with inherently short contribution spans do not influence the analysis. For example, we consider only users who have at least one contribution before and after the 13-week observation window and also at least one contributions before and after the intervention within the observation window. We compare how user engagement changes after expressing dissonance within and outside the QAnon community by repeating the ITS analysis but including contributions made outside of the QAnon subreddits. As a further point of comparison, we also consider belief, instead of dissonance, as an intervention which may determine changes in engagement within the QAnon subreddits.

4.4.8 Results: Changes in User Contributions after Dissonance

To perform ITS analysis, We first identify users who express dissonance in the manually labeled dataset. We rely on the manually labeled dataset, rather than the larger but automatically labeled one, to be fully confident in our identification of dissonant comments and by association, dissonant users. Figure 4.6 displays the ITS and regression results. We find that user contributions inside QAnon decrease significantly (b_2 = -0.02) immediately after expressing dissonance (Figure 4.6 (a)). The same effect is *not*

[7]we also experimented with observation windows of 5, 7, 9, 11, 15, 17 and 21 weeks observing similar results.

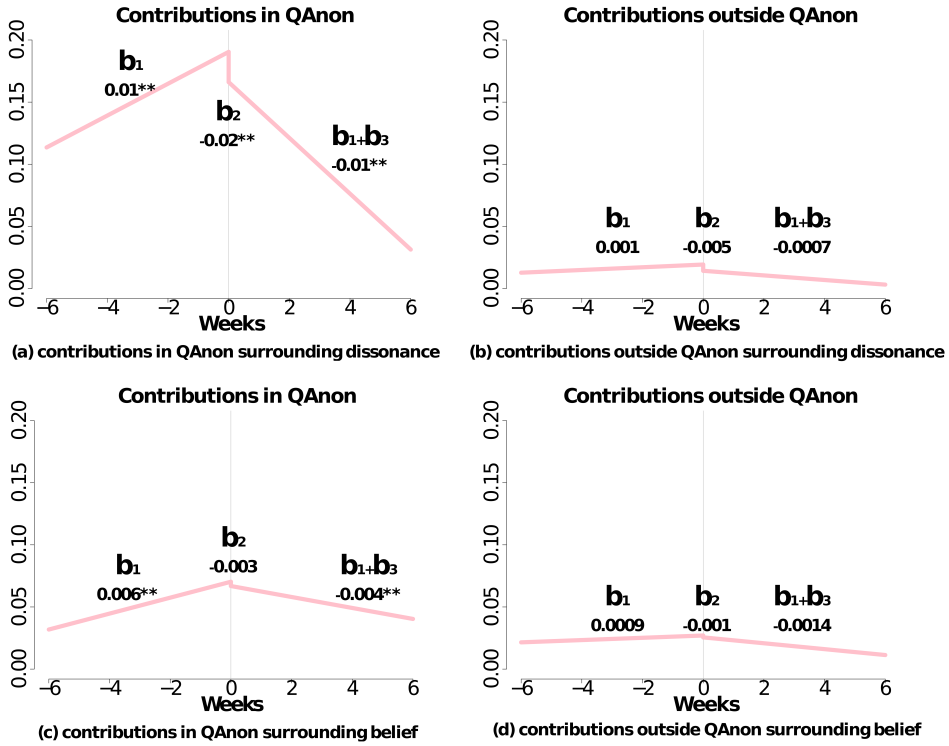


Figure 4.6: ITS plots for user contributions with different interventions. Two asterisks (**) indicate that the coefficient is statistically significant ($p < 0.05$). As shown in (a), immediately after the dissonance (week=0) there is significant decrease in the user contributions inside QAnon subreddits (b_2 = -0.02). However, as indicated in (c) there is no significant change in contributions after expressing belief. In the long term, contributions inside QAnon decrease with higher rate after expressing dissonance ($b_1 + b_3$ = -0.01) compared to belief ($b_1 + b_3$ = -0.004). Moreover, (b) and (c) indicate that there are no significant changes in user contributions outside of QAnon after expressing dissonance or belief.

significant after the users express belief (Figure 4.6 (c)). Further, in the long term as well, contributions inside QAnon decrease at a higher rate after expressing dissonance (b_3 = -0.02) (Figure 4.6 (a)) compared to expressing belief (b_3 = -0.004) (Figure 4.6 (c)). Is this effect a byproduct of users reducing their overall Reddit activity? This does not appears be the case. We find no significant changes in the user activity outside the QAnon subreddits (Figure 4.6 (b) and (d)). While the analysis in Figure 4.6 is based on the 2,371 manually labeled comments spanning over 1,498 users, we also repeat the entire ITS analysis on the complete dataset of 700K comments with labels predicted by the classifier. We find similar results indicating that user contributions decrease significantly soon after dissonance.

4.5 Discussion and Implications

In this chapter, I presented first large scale modeling of long term engagement and radicalization in the online conspiracy communities using a longitudinal digital trace data of 36K Reddit users. Looking deeper into QAnon conspiracy theory groups, I investigated cognitive dissonance as a mechanism through which users may experience disbelief and disengage from conspiracy theory discussion spaces. Below, I discuss some of the key insights and implications generated from this work.

4.5.0.1 Monologicality as a varying process:

While not all conspiracy believers adopt a general conspiratorial worldview as the primary sensemaking device [94], through characterizing conspiracy trajectories we find that two groups of users (those on *steady high* and *increasing* engagement trajectories) do contribute prominently in general conspiracy subreddits that host all types and topics of conspiracies. Are users on *steady high* and *increasing* pathways predisposed to monological thinking? The original theoretical proposition by Goertzel [109] describes monologicality as a stable cognitive style, trait or disposition. However, looking specifically at users on *increasing* pathway, the discussion spaces they engage in become more generalist over time. This suggests that monological conspiracy worldviews can develop over time. In fact, our quantitative results align with qualitative observation of **(author?)** [94] depicting monologicality as a variable *endpoint* of various social processes rather than a cognitive predisposition. In particular, counter to the popular rabbit-hole metaphor, individuals who show signs of radicalization do not seem to narrow their interests down to fringe theories. Instead, such individuals adopt venues of generalist CT discussions together with their idiosyncratic lingo. This observation, on the one hand, purports a parallelism between monological worldviews and radicalization. On the other, it begs the question of what types of online discussion environments harbor the potential for mobilization of radicalized individuals: topically and socially fringe spaces that may host extreme ideas, or comparatively mainstream spaces that afford perception biases of false consensus?

4.5.0.2 Cognitive Dissonance Reduction Strategies

Our results indicate that QAnon followers expressed dissonance about the legitimacy of the QAnon movement, unfulfilled expectations, ineffective practices and distrust in the heroes of the movement. What are the consequences of experiencing disso-

nance? Researchers posit that experiencing cognitive dissonance induces the state of psychological discomfort [88]. To deal with this discomfort, people employ several dissonance reduction strategies [88, 164]. For example, they trivialize the cause of dissonance and self-affirm their belief system [228]. Once an individual trivializes the point of contradiction, other discrepancies can no longer arouse dissonance. Consider this Reddit comment from our dataset for example:

> *I dont really care if Q is real. He has just reinforced my commitment to dig deeper. That's what a real Q dude would do no matter what. DIG DEEPER*

People also find strategies to rationalize the cause of dissonance [164]. Similarly, QAnon followers often rationalized Q's failed predictions by giving alternate explanations of the failures or creating more consonant interpretations of reality.

> *...[you] are not thinking critically in context of everything that comprises Q's message and content. He hasn't "failed". he's made statements that people have misinterpreted and then blamed on Q for the misinterpretations not being correct.*

More importantly, however, researchers state that cognitive dissonance can lead to attitude and behavior changes rejecting previously held beliefs [89]. Meaning, experiencing dissonance with conspiracies may lead people to abandon conspiracy beliefs and pave the way for recovery from conspiratorial worldview. Indeed, our analysis indicates that user contributions lowered after expressing dissonance, and that dissonance increased just before user's departure from the QAnon community. This is mirrored in the findings of our qualitative analysis. We found instances of users indicating that they were leaving the QAnon subreddit as a result of dissonance.

> *Q Anon is A psyop!!!! I am out, this board has been infiltrated. Something good, to something terrible, real quick. Don't fall for this Q stuff. Think for yourselves.*

> *This is the last one for me I think. I got hyped for "the memo" I got hyped for "raw footage". Arrests of the cabal within the week or I am out.*

While acknowledging that dissonance may not always lead to positive behavior change, our results suggest that exploring dissonance as a possible intervention for online conspiracy engagement is a promising future direction. Several other studies have explored dissonance based interventions [95] for reducing implicit racial prejudice [116] and promoting positive social behavior [165]. Below, we discuss how dissonance can be used as an intervention to motivate positive behavior change in conspiracy communities.

4.5.0.3 Intervention for Recovery from Online Conspiracy Discussion Engagement

We showed how users spontaneously disclosed experiences of dissonance and how this correlates with changes in behavior, especially focusing on the effects of disengaging from the conspiracy community. Dissonance can also be introduced externally as an intervention to *induce* such behavior. In fact, similar interventions based on "hypocrisy paradigm" [97], that encourage participants to explore the differences between their internally held beliefs and their public expression, have been tested in settings ranging from mental health through addiction recovery to prejudice reduction [95, 116]. For example, participants with high implicit racial prejudice were asked to write an essay on racial justice and fairness. Publicly expressing views contradictory to the implicit beliefs led the participants to reduce the prejudicial behavior [116]. Our results show that such "hypocrisy" exists in the QAnon conspiracy community. For example, the scenario where users want to be part of the QAnon community while at the same time dislike some aspects of the QAnon movement or doubt the QAnon heroes.

This chapter also offers ways to systematically compile social imaginaries from the point of view of the conspiracist themselves. This may help design community centered hypocrisy interventions based strategies that could nudge the conspiracists to explore the differences between the social imaginaries of the community and their internally held beliefs. For example, building on the qualitative analysis of fracture points in conspiracy belief, one could build interventions that question the infallibility of heroes and the promises made by the movement leaders. Moreover, our results may also help contextualize past successful interventions within the social imaginaries of a specific community and to recast them as hypocrisy interventions, such as questioning the efficacy of the movement to rigorously derive truth [55] or to be effective against foes [239].

Furthermore, the computational framework for the dissonance classifier can be used to identify the central causes of dissonance in the community. These fracture points can be insisted or expanded via strategic interventions. Our results can also inform which interventions might not be successful. For example, in QAnon, "MSM" (mainstream media) is heavily distrusted and is considered as a foe. Interventions citing news articles from mainstream sources maybe met with instant criticism, despite their credibility. Finally, we also offer ways to measure the outcomes of interventions, and therefore to select the ones that are most effective. In sum, conspiracy social imaginaries and computational dissonance detection offer powerful tools to design contextually-informed interventions.

4.5.0.4 Ethical Considerations

Researchers also need to consider social, psychological and ethical consequences of designing such intervention systems. For example, Sunstein [243] argue that sowing the seeds of doubt in conspiracy theory communities is most (or perhaps only) effective when done from within the community [243]. Skepticism coming from outsiders may be deemed illegitimate, or even be construed as part of a larger conspiracy attempting to undermine the conspiracists' truth [155]. Dissonance causes psychological discomfort, and therefore interventions inducing dissonance should weigh harms against benefits. It is also important to consider *whether* certain conspiracies need intervention by accounting for the researchers' socio-political biases. These are but few of the ethical issues that intervention designers should consider before interacting directly with social media participants.

4.6 Limitations

This work has some limitations that should be acknowledged. First, we characterize the conspiracy engagement trajectories by using the contribution volumes of users, a measure commonly used as a strong latent proxy for user engagement [112]. While this affords studying the complete evolution of contributions in the CT communities, analyzing contribution volume alone can limit the interpretability of quality of contributions. For example, it is possible that some users may be contributing troll posts while keeping the same contribution volume as others. A more nuanced measure of CT engagement could involve analyzing text and context of the user contributions. Second, this work offers empirical insights on how users escalate through the formative phases of radicalization. Yet, it would be crucial to unpack when and how this potential is turned into action in the Resolution and Operational phases. Our work provides a framework for experimental designs in this direction.

Further, while our cognitive dissonance classifier gives precision above 0.7 in a complex three-class classification problem, it relies on relatively simple features. Its performance would likely improve further after incorporating stylistic and meta-linguistic features and a larger labeled dataset. Finally, while our analysis provides initial evidence about changes in user engagement following dissonance disclosures, we do not make causal claims. Studies aiming at deriving causal connections between disclosures of dissonance and behavior change should adopt controlled experimental designs and should account for potential confounders.

4.7 Conclusions

In this work, I investigated the association between online conspiracy theory discussions and various pathways of engagement. Through an ensemble of computationally derived features backed by theoretical models, I investigated how users on the increasing engagement pathway show distinctive behavior compared to those on the decreasing engagement pathway. Moreover, looking into the QAnon conspiracy theory subreddits, I uncovered five dimensions along which believers expressed dissonance with the QAnon worldview. I further provided evidence that user contributions inside QAnon communities decrease immediately after self-disclosures of dissonance and that high levels of experienced dissonance correlate with users ultimately leaving the communities. These results show that users *do* express dissonance inside their communities and dissonance can be explored further as a possible intervention for online conspiracy engagement.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Through my doctoral research, hope to generate insights into various ways technology could affect users' belief in problematic information. Some of the work presented in Chapter 4 already gives empirical evidence that not only disclosures of dissonance are followed by a decrease in contributions, but also by the departure of the users from the community [198]. How can this knowledge inform the design of interventions for countering belief in problematic information? In this chapter, I briefly outline possible next steps based on my dissertation.

Specifically, I envision two future directions: First, 1) building a rigorous, multi-disciplinary research foundation for countering problematic information online, and then 2) applying this knowledge to design ethical, impactful, and fair interventions for online communities of problematic information.

5.1 Future Directions

5.1.1 Multidisciplinary research foundation for countering problematic information online

My doctoral research and the works of other scholars indicate that people may get disillusioned by problematic narratives through a combination of epistemic and social factors. Therefore, as a first step towards reducing problematic content online, I want to formalize the effect of various factors such as information, social context, narratives,

language, and offline identities through pilot experiments in open discussions and user studies.

For instance, in the case of online conspiracy theory discussions, it would be valuable to understand the interplay of cross-cutting information and social interactions that result in more factual discourse. As a former anti-Vaxxer in our study points out—*"I just had to type in the word 'debunk' in all those anti-vaccination videos I watched on YouTube. I was always one word away from getting out of the rabbit hole "*. The accounts by interview participants motivate me to study problematic information from the information retrieval perspective. Specifically, I am interested in modeling information retrieval habits surrounding problematic information across platforms and digital identities to pinpoint potential counterfactual scenarios that can lead users away from problematic content. I am also interested in exploring how insider language and identity-based framing increase trust in scientific information inside communities of problematic information.

Another question of interest is to what extent different socio-technical factors impact disengagement from problematic content. To what extent users' exits from problematic discussions are internet-mediated? Building on my work investigating patterns of disengagement in ex-conspiracy theory believers, I want to further characterize the role of intersectional factors and offline identities in reducing belief in problematic information. By developing a research foundation for various epistemic and social factors, I aim to contribute a knowledge base for researchers to design online nudges and interventions to reduce problematic content online.

5.1.2 Designing online prompts to reduce engagement in problematic information

Next, I want to research specific nudges designed to introduce doubt and critical reflection while consuming problematic content. Designing online nudges needs a multi-pronged approach that grapples with the ethics, impact, fairness, and scalability of online interventions.

First, I aim to contribute to a discussion on which types of problematic information warrant intervention, who should have the authority or the responsibility to intervene, and how online nudges can respect the freedom of speech and healthy debate. Moreover, I plan to undertake studies that estimate the positive and negative consequences of interventions against problematic information in open discussions. I believe that

creating a foundation for ethical and fair practices is a crucial step toward increasing trust and acceptance of online nudges against problematic information.

Second, given the expanse of online discussions, it is also essential to study the trade-off between the impact and scalability of online nudges. I have already started work in this direction with Google Jigsaw, where we are designing interventions based on rhetorical questioning to contest misconceptions in climate change denial discussions on Reddit. To make interventions more scalable, we are leveraging natural language generation models trained on the insider language of climate change denialists. Further, to test the impact, we are exploring various combinations of research designs ranging from single prompt interventions to dialogic interventions in Reddit discussion threads. I believe this line of research needs more in-depth future exploration, experimenting with the language and method of delivery, to create scalable and impactful interventions reducing problematic information online.

5.2 Conclusion

In this dissertation, I have documented my research exploring engagement, mobilization, and disengagement from communities of problematic information.

In Chapter 2, using a theory-driven framework of social factors across six dimensions, I perform a retrospective case-controlled study of *future conspiracists* and compare them with *non-conspiracists*. The research not only finds that the social factors are important but that conspiracy joining can be explained at least partially by at least one feature in each social factors group. Given these findings, I offer a unique, empirically backed perspective on the life cycle of conspiracists, echo chambers in conspiracy communities, and the effect of social exclusion in conspiracy engagement.

In Chapter 3, I presented three studies. The first study investigated framing strategies used by hate groups across online platforms finding that hate groups use affordances of Facebook and Twitter differently to propel narratives of radicalization and mass education. The second study characterizing roles in hateful information mobilization found that online accounts assume different roles such as educators and solicitors while mobilizing content related to the hate movement. Combined together, these works contribute cross-platform, empirical understanding of communication and mobilization practices used in hate movements in the West. In the third study, I analyzed political amplification through lexical mutants in India and demonstrated how multiple political parties leverage political amplification in reactionary online

campaigns.

In Chapter 4, I explore various mechanisms through which users may disengage from conspiracy theory discussions online. Specifically, I find that users who decrease their participation in conspiracy theory discussions show early signs of low language conformity and high topical specialization in conspiracy theorizing. Moreover, in another study, I find that cognitive dissonance is one of the mechanisms by which users may lose their conspiracy theory belief and exit online conspiracy theory discussion communities.

Through my completed research, I showcased how large-scale, theory-guided analysis of problematic information can provide insights into self-selection, group conformity, worldview, and social imaginaries of users engaged in communities of problematic information. I believe this work addresses an important need to apply social science theories at scale on large, organic datasets of social media discourse to gain deeper insights into the social processes in communities of problematic information.

BIBLIOGRAPHY

[1] M. ABALAKINA-PAAP, W. G. STEPHAN, T. CRAIG, AND W. L. GREGORY, *Beliefs in conspiracies*, Political Psychology, 20 (1999), pp. 637–647.

[2] ADL, *Bitchute: A hotbed of hate | anti-defamation league*. `https://www.adl.org/blog/bitchute-a-hotbed-of-hate`, August 2020.
(Accessed on 10/14/2020).

[3] S. AHMED, K. JAIDKA, AND J. CHO, *The 2014 indian elections on twitter: A comparison of campaign strategies of political parties*, Telematics and Informatics, 33 (2016), pp. 1071–1087.

[4] S. Z. AKBAR, A. PANDA, D. KUKRETI, A. MEENA, AND J. PAL, *Misinformation as a window into prejudice: Covid-19 and the information environment in india*, Proceedings of the ACM on Human-Computer Interaction, 4 (2021), pp. 1–28.

[5] M. ALIAPOULIOS, A. PAPASAVVA, C. BALLARD, E. DE CRISTOFARO, G. STRINGHINI, S. ZANNETTOU, AND J. BLACKBURN, *The gospel according to q: Understanding the qanon conspiracy from the perspective of canonical information*, arXiv preprint arXiv:2101.08750, (2021).

[6] A. AMARASINGAM AND M.-A. ARGENTINO, *The qanon conspiracy theory: A security threat in the making*, CTC Sentinel, 13 (2020), pp. 37–44.

[7] O. ARAZY, H. LIIFSHITZ-ASSAF, O. NOV, J. DAXENBERGER, M. BALESTRA, AND C. CHESHIRE, *On the" how" and" why" of emergent role behaviors in wikipedia*, in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017, pp. 2039–2051.

[8] B. ARDA, *The construction of a new sociality through social media: The case of the gezi uprising in turkey*, Conjunctions. Transdisciplinary Journal of Cultural Participation, 2 (2015), pp. 72–99.

[9] B. ARECHIGA, *Mythic pizza: Semiotic and archetypal significance in the conspiracy narrative known as' pizzagate'*, (2019).

[10] S. ARORA, Y. LIANG, AND T. MA, *A simple but tough-to-beat baseline for sentence embeddings*, in ICLR 2017, 2017, pp. 1–16.

[11] M. ARTETXE AND H. SCHWENK, *Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond*, Transactions of the Association for Computational Linguistics, 7 (2019), pp. 597–610.

[12] E. M. ASWAD, *The future of freedom of expression online*, Duke L. & Tech. Rev., 17 (2018), p. 26.

[13] J. ATTENBERG AND F. PROVOST, *Why label when you can search? alternatives to active learning for applying human resources to build classification models under extreme class imbalance*, in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, pp. 423–432.

[14] G. A. AUGER, *Fostering democracy through social media: Evaluating diametrically opposed nonprofit advocacy organizations' use of facebook, twitter, and youtube*, Public Relations Review, 39 (2013), pp. 369–376.

[15] J. A. BANAS AND G. MILLER, *Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies*, Human Communication Research, 39 (2013).

[16] E. BARKER, *New religious movements: their incidence and significance*, Routledge, 1999.

[17] J. BARNES, *Blaming the deep state: Officials accused of wrongdoing adopt trump,Äôs response - the new york times*.
`https://www.nytimes.com/2018/12/18/us/politics/deep-state-trump-classified-information.html`, December 2018.
(Accessed on 02/03/2021).

[18] L. BASHAM, *Malevolent global conspiracy*, Conspiracy Theories: The Philosophical Debate, (2006), pp. 93–106.

[19] F. R. BAUMGARTNER, S. L. DE BOEF, AND A. E. BOYDSTUN, *The decline of the death penalty and the discovery of innocence*, Cambridge University Press, 2008.

[20] J. BAUMGARTNER, S. ZANNETTOU, B. KEEGAN, M. SQUIRE, AND J. BLACKBURN, *The pushshift reddit dataset*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, 2020, pp. 830–839.

[21] H. S. BECKER, *Outsiders*, Simon and Schuster, 2008.

[22] H. BEIRICH AND S. BUCHANAN, *2017: The Year in Hate and Extremism*, tech. rep., Southern Poverty Law Center, 2018.

[23] R. D. BENFORD AND D. A. SNOW, *Framing processes and social movements: An overview and assessment*, Annual review of sociology, 26 (2000), pp. 611–639.

[24] A. J. BERINSKY AND D. R. KINDER, *Making sense of issues through media frames: Understanding the kosovo crisis*, The Journal of Politics, 68 (2006), pp. 640–656.

[25] A. BESSI, M. COLETTO, G. A. DAVIDESCU, A. SCALA, G. CALDARELLI, AND W. QUATTROCIOCCHI, *Science vs conspiracy: Collective narratives in the age of misinformation*, PloS one, 10 (2015), p. e0118093.

[26] A. BESSI, F. ZOLLO, M. DEL VICARIO, A. SCALA, G. CALDARELLI, AND W. QUATTROCIOCCHI, *Trend of narratives in the age of misinformation*, PLoS ONE, 10 (2015), pp. 1–16.

[27] P. BIYANI, C. CARAGEA, P. MITRA, AND J. YEN, *Identifying emotional and informational support in online health communities*, in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 827–836.

[28] A. BLAKE, *Vaccine conspiracy theorists become even more desperate after full fda authorization - the washington post*.
`https://www.washingtonpost.com/politics/2021/08/26/vaccine-conspiracy-theorists-become-even-more-desperate-after-full-fda-authorization/`, August 2021.
(Accessed on 08/28/2021).

[29] A.-M. BLIUC, D. BEST, M. IQBAL, AND K. UPTON, *Building addiction recovery capital through online participation in a recovery community*, Social Science & Medicine, 193 (2017), pp. 110–117.

[30] P. BORDIA AND N. DIFONZO, *Problem solving in social interactions on the internet: Rumor as social cognition*, Social Psychology Quarterly, 67 (2004), pp. 33–49.

[31] G. BOUMA, *Normalized (pointwise) mutual information in collocation extraction*, Proceedings of GSCL, (2009), pp. 31–40.

[32] A. E. BOYDSTUN, *Tracking the Development of Media Frames within and across Policy Issues*, (2014).

[33] B. C. BRANDY ZADROZNY, *How three conspiracy theorists took 'q' and sparked qanon*.
https://www.nbcnews.com/tech/tech-news/how-three-conspiracy-theorists-took-q-sparked-qanon-n900531, August 2018.
(Accessed on 02/17/2021).

[34] ——, *Reddit bans qanon subreddits after months of violent threats*.
https://www.nbcnews.com/tech/tech-news/reddit-bans-qanon-subreddits-after-months-violent-threats-n909061, September 2018.
(Accessed on 02/17/2021).

[35] D. G. BROMLEY AND A. D. SHUPE JR, *Financing the new religions: A resource mobilization approach*, Journal for the Scientific Study of Religion, (1980), pp. 227–239.

[36] O. W. BUREAU, *'manipulated media' controversy: Government vs twitter*.
https://www.outlookindia.com/website/story/india-news-manipulated-media-controversy-government-vs-twitter/383352, May 2021.
(Accessed on 01/14/2023).

[37] D. M. BUSS, *Selection, evocation, and manipulation.*, Journal of personality and social psychology, 53 (1987), p. 1214.

[38] L. D. BUTLER, C. KOOPMAN, AND P. G. ZIMBARDO, *The psychological impact of viewing the film" jfk": Emotions, beliefs, and political behavioral intentions*, Political psychology, (1995).

[39] M. BUTTER AND P. KNIGHT, *Routledge handbook of conspiracy theories*, Routledge, 2020.

[40] M. C LIN, *Active learning with unbalanced classes & example-generated queries*, in AAAI Conference on Human Computation, 2018.

[41] D. CARD, A. E. BOYDSTUN, J. H. GROSS, P. RESNIK, AND N. A. SMITH, *The Media Frames Corpus: Annotations of Frames Across Issues*, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), (2015), pp. 438–444.

[42] W. K. CARROLL AND R. A. HACKETT, *Democratic media activism through the lens of social movement theory*, Media, culture & society, 28 (2006), pp. 83–104.

[43] S. P. L. CENTER, *Hate map | southern poverty law center*. `https://www.splcenter.org/hate-map`, September 2019. (Accessed on 09/13/2019).

[44] S. CHANCELLOR, J. A. PATER, T. A. CLEAR, E. GILBERT, AND M. DE CHOUDHURY, *#thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities*, in Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, New York, New York, USA, 2016, ACM Press, pp. 1199–1211.

[45] E. CHANDRASEKHARAN, M. SAMORY, A. SRINIVASAN, AND E. GILBERT, *The bag of communities: identifying abusive behavior online with preexisting internet data*, in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, ACM, 2017, pp. 3175–3187.

[46] J. P. CHANG, J. CHENG, AND C. DANESCU-NICULESCU-MIZIL, *Don‚Äôt let me be misunderstood: Comparing intentions and perceptions in online discussions*, in Proceedings of The Web Conference 2020, 2020, pp. 2066–2077.

[47] M. CHAU AND J. XU, *Mining communities and their relationships in blogs: A study of online hate groups*, International Journal of Human-Computer Studies, 65 (2007), pp. 57–70.

[48] T. CHEN AND C. GUESTRIN, *Xgboost: A scalable tree boosting system*, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[49] X. CHEN, S. ZHAO, AND W. LI, *Opinion dynamics model based on cognitive styles: field-dependence and field-independence*, Complexity, 2019 (2019).

[50] D. CHONG AND J. N. DRUCKMAN, *Framing Theory*, Annual Review of Political Science, 10 (2007), pp. 103–126.

[51] M. CINELLI, S. CRESCI, W. QUATTROCIOCCHI, M. TESCONI, AND P. ZOLA, *Coordinated inauthentic behavior and information spreading on twitter*, Decision Support Systems, (2022), p. 113819.

[52] M. CLAESEN AND B. DE MOOR, *Hyperparameter search in machine learning*, arXiv preprint arXiv:1502.02127, (2015).

[53] J. CLARK AND D. A. HOLTON, *A first look at graph theory*, World Scientific, 1991.

[54] S. CLARKE, *Conspiracy Theories and Conspiracy Theorizing*, Philosophy of the Social Sciences, 32 (2002), pp. 131–150.

[55] J. COOK, P. ELLERTON, AND D. KINKEAD, *Deconstructing climate misinformation to identify reasoning errors*, Environmental Research Letters, 13 (2018), p. 024018.

[56] C. CORRIGALL-BROWN, *Patterns of protest: Trajectories of participation in social movements*, Stanford University Press, 2011.

[57] M. COSTELLO AND J. HAWDON, *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 (2018), pp. 55–60.

[58] M. COSTELLO, J. HAWDON, T. RATLIFF, AND T. GRANTHAM, *Who views online extremism? individual attributes leading to exposure*, Computers in Human Behavior, 63 (2016).

[59] M. COULTER, *Paypal shuts russian crowdfunder's account after alt-right influx*. `https://www.ft.com/content/7c4285b2-fe2f-11e8-ac00-57a2a826423e`, December 2018.
(Accessed on 09/13/2019).

[60] S. CRIMANDO, *Q-speak: The language of qanon*.

https://www.asisonline.org/security-management-magazine/latest-news/online-exclusives/2021/q-speak-the-language-of-qanon/, January 2021.
(Accessed on 04/12/2021).

[61] C. CROSSETT AND J. SPITALETTA, *Radicalization: Relevant psychological and sociological concepts*, The John Hopkins University, (2010).

[62] M. R. CZAJA, A. D. BRIGHT, AND S. P. COTTRELL, *Integrative complexity, beliefs, and attitudes: Application to prescribed fire*, Forest Policy and Economics, 62 (2016), pp. 54–61.

[63] K. DAELLENBACH AND J. PARKINSON, *A useful shift in our perspective: integrating social movement framing into social marketing*, Journal of Social Marketing, 7 (2017), pp. 188–204.

[64] A. DALGAARD-NIELSEN, *Violent radicalization in europe: What we know and what we do not know*, Studies in conflict & terrorism, 33 (2010), pp. 797–814.

[65] H. DARWIN, N. NEAVE, AND J. HOLMES, *Belief in conspiracy theories. the role of paranormal belief, paranoid ideation and schizotypy*, Personality and Individual Differences, 50 (2011), pp. 1289–1293.

[66] A. DAS AND R. SCHROEDER, *Online disinformation in the run-up to the indian 2019 election*, Information, Communication & Society, 24 (2021), pp. 1762–1778.

[67] S. DASH, A. ARYA, S. KAUR, AND J. PAL, *Narrative building in propaganda networks on indian twitter*, in 14th ACM Web Science Conference 2022, 2022, pp. 239–244.

[68] S. DASH, D. MISHRA, G. SHEKHAWAT, AND J. PAL, *Divided we rule: Influencer polarization on twitter during political crises in india*, arXiv preprint arXiv:2105.08361, (2021).

[69] M.-C. DE MARNEFFE, C. D. MANNING, AND C. POTTS, *Did it happen? the pragmatic complexity of veridicality assessment*, Computational linguistics, 38 (2012), pp. 301–333.

[70] C. H. DE VREESE AND H. BOOMGAARDEN, *News, political knowledge and participation: The differential effects of news media exposure on political knowledge and participation*, Acta Politica, 41 (2006), pp. 317–341.

[71] A. DECTER-FRAIN AND J. A. FRIMER, *Impressive words: linguistic predictors of public approval of the us congress*, Frontiers in psychology, 7 (2016), p. 240.

[72] J. DEVLIN, M.-W. CHANG, K. LEE, AND K. TOUTANOVA, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[73] G. DI DOMENICO AND M. VISENTIN, *Fake news or true lies? reflections about problematic contents in marketing*, International Journal of Market Research, 62 (2020), pp. 409–417.

[74] M. DIANI, *The concept of social movement*, The sociological review, 40 (1992), pp. 1–25.

[75] J. P. DIMOND, M. DYE, D. LAROSE, AND A. S. BRUCKMAN, *Hollaback!: the role of storytelling online in a social movement organization*, in Proceedings of the 2013 conference on Computer supported cooperative work, ACM, 2013, pp. 477–490.

[76] E. DINAN, S. HUMEAU, B. CHINTAGUNTA, AND J. WESTON, *Build it break it fix it for dialogue safety: Robustness from adversarial human attack*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4529–4538.

[77] K. M. DOUGLAS, *Psychology, discrimination and hate groups online*, The Oxford handbook of internet psychology, (2007), pp. 155–163.

[78] M. E. DUFFY, *Web of hate: A fantasy theme analysis of the rhetorical vision of hate groups online*, Journal of Communication Inquiry, 27 (2003), pp. 291–312.

[79] D. B. EDINGO, *Social media, public sphere and counterpublics: An exploratory analysis of the networked use of twitter during the protests against the citizenship amendment act in india*, The Journal of Social Media in Society, 10 (2021), pp. 76–101.

[80] B. EDWARDS AND J. D. MCCARTHY, *Resources and social movement mobilization*, The Blackwell companion to social movements, (2004), pp. 116–152.

[81] J. EISENSTEIN, A. AHMED, AND E. P. XING, *Sparse additive generative models of text*, Proceedings of the 28th International Conference on Machine Learning (ICML-11), (2011), pp. 1041–1048.

[82] R. M. ENTMAN, *Framing: Toward Clarification of a Fractured Paradigm*, Journal of Communication, 43 (1993), pp. 51–58.

[83] A. M. EVANS, O. STAVROVA, AND H. ROSENBUSCH, *Expressions of doubt and trust in online user reviews*, Computers in Human Behavior, 114 (2021), p. 106556.

[84] FACEBOOK, *Standing against hate | facebook newsroom*. https://newsroom.fb.com/news/2019/03/standing-against-hate/, March 2019.
(Accessed on 09/13/2019).

[85] R. J. FEISE, *Do multiple outcome measures require p-value adjustment?*, BMC medical research methodology, 2 (2002), p. 8.

[86] M. FENSTER, *Conspiracy theories: Secrecy and power in American culture*, U of Minnesota Press, 1999.

[87] M. M. FERREE AND F. D. MILLER, *Mobilization and meaning: Toward an integration of social psychological and resource perspectives on social movements*, Sociological Inquiry, 55 (1985), pp. 38–61.

[88] L. FESTINGER, *A theory of cognitive dissonance*, vol. 2, Stanford university press, 1962.

[89] L. FESTINGER AND J. M. CARLSMITH, *Cognitive consequences of forced compliance.*, The journal of abnormal and social psychology, 58 (1959), p. 203.

[90] L. FESTINGER, H. RIECKEN, AND S. SCHACHTER, *When prophecy fails: A social and psychological study of a modern group that predicted the destruction of the world*, Lulu Press, Inc, 2017.

[91] M. FEURER, K. EGGENSPERGER, S. FALKNER, M. LINDAUER, AND F. HUTTER, *Auto-sklearn 2.0: The next generation*, arXiv preprint arXiv:2007.04074, (2020).

[92] I. FOR ECONOMICS AND PEACE, *Gti-2020-web-1.pdf*. https://www.visionofhumanity.org/wp-content/uploads/2020/11/GTI-2020-web-1.pdf, November 2020.
(Accessed on 01/11/2021).

[93] B. FRANKS, A. BANGERTER, AND M. BAUER, *Conspiracy theories as quasi-religious mentality: an integrated account from cognitive science, social representations theory, and frame theory*, Frontiers in psychology, 4 (2013), p. 424.

[94] B. FRANKS, A. BANGERTER, M. W. BAUER, M. HALL, AND M. C. NOORT, *Beyond "monologicality"? Exploring conspiracist worldviews*, Frontiers in Psychology, 8 (2017).

[95] T. FREIJY AND E. J. KOTHE, *Dissonance-based interventions for health behaviour change: A systematic review*, British journal of health psychology, 18 (2013), pp. 310–337.

[96] S. FRENKEL AND D. ALBA, *In india, facebook struggles to combat misinformation and hate speech - the new york times*.
https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html, October 2021.
(Accessed on 04/21/2022).

[97] C. B. FRIED AND E. ARONSON, *Hypocrisy, misattribution, and dissonance reduction*, Personality and Social Psychology Bulletin, 21 (1995), pp. 925–933.

[98] J. FRIMER, R. BOGHRATI, J. HAIDT, J. GRAHAM, AND M. DEHGANI, *Moral foundations dictionary for linguistic analyses 2.0*, Unpublished manuscript, (2019).

[99] M. GALANTER, *Alcoholics anonymous and twelve-step recovery: A model based on social and cognitive neuroscience*, The American journal on addictions, 23 (2014), pp. 300–307.

[100] R. J. GALLAGHER, L. DOROSHENKO, S. SHUGARS, D. LAZER, AND B. FOUCAULT WELLES, *Sustained online amplification of covid-19 elites in the united states*, Social Media+ Society, 7 (2021), p. 20563051211024957.

[101] W. A. GAMSON, W. A. G. GAMSON, W. A. GAMSON, AND W. A. GAMSON, *Talking politics*, Cambridge university press, 1992.

[102] M. GARRETT, *Capitol riot exposes reach of qanon disinformation: "it was a drug" - cbs news*.
https://www.cbsnews.com/news/qanon-capitol-riot-reach/, January 2021.
(Accessed on 08/26/2021).

[103] A. GARRY, S. WALTHER, R. RUKAYA, AND A. MOHAMMED, *Qanon conspiracy theory: Examining its evolution and mechanisms of radicalization*, Journal for Deradicalization, (2021), pp. 152–216.

[104] M. GEMIGNANI AND Y. HERNANDEZ-ALBUJAR, *Ethnic and Racial Studies Hate groups targeting unauthorized immigrants in the US: discourses, narratives and subjectivation practices on their websites*, Ethnic and Racial Studies, 0 (2015).

[105] C. GENOLINI, R. ECOCHARD, M. BENGHEZAL, T. DRISS, S. ANDRIEU, AND F. SUBTIL, *kmlshape: an efficient method to cluster longitudinal data (time-series) according to their shapes*, Plos one, 11 (2016), p. e0150738.

[106] P. B. GERSTENFELD, D. R. GRANT, AND C.-P. CHIANG, *Hate Online: A Content Analysis of Extremist Internet Sites*, Analyses of Social Issues and Public Policy, 3 (2003), pp. 29–44.

[107] C. H. E. GILBERT, *Vader: A parsimonious rule-based model for sentiment analysis of social media text*, in Proc. ICWSM, 2014.

[108] D. GILBERT, *The jfk qanon cult in dallas is somehow getting weirder*. `https://www.vice.com/en/article/g5qwb3/jfk-qanon-dallas-cult-growing`, February 2022.
(Accessed on 05/05/2022).

[109] T. GOERTZEL, *Belief in conspiracy theories*, Political Psychology, (1994).

[110] E. GOFFMAN, *Frame analysis: An essay on the organization of experience.*, Harvard University Press, 1974.

[111] M. S. GRANOVETTER, *The strength of weak ties*, in Social networks, Elsevier, 1977.

[112] W. HAMILTON, J. ZHANG, C. DANESCU-NICULESCU-MIZIL, D. JURAFSKY, AND J. LESKOVEC, *Loyalty in online communities*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, 2017.

[113] J. HAWDON, *Applying differential association theory to online hate groups: a theoretical statement*, (2012).

[114] C. HEATH, C. BELL, AND E. STERNBERG, *Emotional selection in memes: the case of urban legends.*, Journal of personality and social psychology, 81 (2001).

[115] T. C. HELMUS, E. YORK, AND P. CHALK, *Promoting Online Voices for Countering Violent Extremism*, tech. rep., RAND Corporation, Santa Monica, CA, 2013.

[116] L. S. S. HING, W. LI, AND M. P. ZANNA, *Inducing hypocrisy to reduce prejudicial responses among aversive racists*, Journal of Experimental Social Psychology, 38 (2002), pp. 71–78.

[117] T. K. HO, *Random decision forests*, in Proceedings of 3rd international conference on document analysis and recognition, vol. 1, IEEE, 1995, pp. 278–282.

[118] R. HOFSTADTER, *The paranoid style in American politics*, Vintage, 2012.

[119] A. HOLUB, P. PERONA, AND M. C. BURL, *Entropy-based active learning for object recognition*, in 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, 2008, pp. 1–8.

[120] B. D. HORNE, S. ADALI, AND S. SIKDAR, *Identifying the social signals that drive online discussions: A case study of reddit communities*, in 2017 26th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2017, pp. 1–9.

[121] K. HRISTAKIEVA, S. CRESCI, G. DA SAN MARTINO, M. CONTI, AND P. NAKOV, *The spread of propaganda by coordinated communities on social media*, in 14th ACM Web Science Conference 2022, 2022, pp. 191–201.

[122] C. J. HUTTO AND E. GILBERT, *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*, in Eighth International Conference on Weblogs and Social Media - ICWSM '14, 2014, pp. 216–225.

[123] IACP, *Facebook and violent extremism awareness brief*. `https://www.theiacp.org/resources/document/facebook-and-violent-extremism-awareness-brief`, January 2020. (Accessed on 10/15/2020).

[124] I. ÍÑIGO-MORA, *On the use of the personal pronoun we in communities*, Journal of Language and Politics, 3 (2004).

[125] C. JACK, *Lexicon of lies: Terms for problematic information*, Data & Society, 3 (2017), pp. 1094–1096.

[126] C. JAFFRELOT, *The modi-centric bjp 2014 election campaign: New techniques and old tactics*, Contemporary South Asia, 23 (2015), pp. 151–166.

[127] M. JAKESCH, K. GARIMELLA, D. ECKLES, AND M. NAAMAN, # *trend alert: How a cross-platform organization manipulated twitter trends in the indian general election*, arXiv preprint arXiv:2104.13259, (2021).

[128] V. JENNESS, *Social movement growth, domain expansion, and framing processes: The gay/lesbian movement and violence against gays and lesbians as a social problem*, Social Problems, 42 (1995), pp. 145–170.

[129] K. JOB-SLUDER AND S. A. BARAB, *Shared" we" and shared" they" indicators of group identity in online teacher professional development*, Designing for virtual communities in the service of learning, (2004).

[130] S. G. JONES, C. DOXSEE, AND N. HARRINGTON, *The escalating terrorism problem in the united states*, (2020).

[131] R. JOSHI, P. GOEL, AND R. JOSHI, *Deep learning for hindi text classification: A comparison*, in International Conference on Intelligent Human Computer Interaction, Springer, 2020, pp. 94–101.

[132] S. JUNGKUNZ, *Towards a measurement of extreme left-wing attitudes*, German Politics, 28 (2019), pp. 101–122.

[133] M. KAMINSKAS AND D. BRIDGE, *Measuring surprise in recommender systems*, in Proceedings of the workshop on recommender systems evaluation: dimensions and design (Workshop programme of the 8th ACM conference on recommender systems), Citeseer, 2014.

[134] B. L. KEELEY, *Of conspiracy theories*, The journal of Philosophy, 96 (1999), pp. 109–126.

[135] J. M. KELLER, *Virtual feminisms: Girls,Äô blogging communities, feminist activism, and participatory politics*, Information, Communication & Society, 15 (2012), pp. 429–447.

[136] M. KELLY, *Facebook still hosts boogaloo extremist groups, report finds - the verge*. `https://www.theverge.com/2020/8/12/21365278/facebook-boogaloo-tech-transparency-right-wing-extremist-platform`, 2020.

(Accessed on 09/13/2020).

[137] E. KICIMAN, S. COUNTS, AND M. GASSER, *Using longitudinal social media analysis to understand the effects of early college alcohol use*, in Twelfth International AAAI Conference on Web and Social Media, 2018.

[138] T. KIILAKOSKI AND A. OKSANEN, *Soundtrack of the school shootings: Cultural script, music and male rage*, Young, 19 (2011), pp. 247–269.

[139] B. KLANDERMANS, *Mobilization and participation: Social-psychological expansisons of resource mobilization theory*, American sociological review, (1984), pp. 583–600.

[140] B. KLANDERMANS AND D. OEGEMA, *Potentials, networks, motivations, and barriers: Steps towards participation in social movements*, American sociological review, (1987), pp. 519–531.

[141] C. KLEIN, P. CLUTTON, AND A. G. DUNN, *Pathways to conspiracy: the social and linguistic precursors of involvement in Reddit,Äôs conspiracy theory forum*.

[142] Y. KOU, X. GUI, Y. CHEN, AND K. PINE, *Conspiracy Talk on Social Media*, Proceedings of the ACM on Human-Computer Interaction, 1 (2017), pp. 1–21.

[143] P. KREKÓ, *CONSPIRACY THEORY AS COLLECTIVE MOTIVATED COGNITION*, in The Psychology of Conspiracy, Routledge, 2015, ch. 10.

[144] M. KRONA, *5 isis,Äôs media ecology and participatory activism tactics*, The Media World of ISIS, (2019), p. 101.

[145] A. W. KRUGLANSKI, M. J. GELFAND, J. J. BÉLANGER, A. SHEVELAND, M. HETIARACHCHI, AND R. GUNARATNA, *The psychology of radicalization and deradicalization: How significance quest impacts violent extremism*, Political Psychology, 35 (2014), pp. 69–93.

[146] M. KUBAT, S. MATWIN, ET AL., *Addressing the curse of imbalanced training sets: one-sided selection*, in Icml, vol. 97, Citeseer, 1997, pp. 179–186.

[147] S. KUMAR, W. L. HAMILTON, J. LESKOVEC, AND D. JURAFSKY, *Community interaction and conflict on the web*, in Proceedings of the 2018 world wide web conference, 2018, pp. 933–943.

[148] C. LAMPE AND P. RESNICK, *Slash (dot) and burn: distributed moderation in a large online conversation space*, in Proceedings of the SIGCHI conference on Human factors in computing systems, 2004, pp. 543–550.

[149] P. LANCE AND A. HATTORI, *Sampling and evaluation, a guide to sampling for program impact evaluation*, Chapel Hill, North Carolina: MEASURE Evaluation, University of North Carolina, (2016).

[150] A. LEAVITT AND J. J. ROBINSON, *The Role of Information Visibility in Network Gatekeeping*, Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17, (2017), pp. 1246–1261.

[151] E. LEE AND L. LEETS, *Persuasive Storytelling*, American Behavioral Scientist, 45 (2002), pp. 927–957.

[152] M. LEONE, *Fundamentalism, anomie, conspiracy: Umberto eco,Äôs semiotics against interpretive irrationality*, in Umberto Eco in his Own Words, De Gruyter Mouton, 2017, pp. 221–229.

[153] J. LESKOVEC, L. BACKSTROM, AND J. KLEINBERG, *Meme-tracking and the dynamics of the news cycle*, in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 497–506.

[154] B. LEVIN, *Cyberhate: A Legal and Historical Analysis of Extremists' Use of Computer Networks in America*, American Behavioral Scientist, 45 (2002), pp. 958–988.

[155] S. LEWANDOWSKY, G. E. GIGNAC, AND K. OBERAUER, *The Role of Conspiracist Ideation and Worldviews in Predicting Rejection of Science*, PLoS ONE, 8 (2013).

[156] P. LUKIĆ, I. ŽEŽELJ, AND B. STANKOVIĆ, *How (ir) rational is it to believe in contradictory conspiracy theories?*, Europe's journal of psychology, 15 (2019), p. 94.

[157] S. MACDONALD, S. G. CORREIA, AND A.-L. WATKIN, *Regulating terrorist content on social media: automation and the rule of law*, International Journal of Law in Context, 15 (2019), pp. 183–197.

[158] T. MARTIN, *community2vec: Vector representations of online communities encode semantic relationships*, in Proc. of the Second Workshop on NLP and Computational Social Science, 2017.

[159] G. T. MARX AND J. L. WOOD, *Strands of theory and research in collective behavior*, Annual review of sociology, 1 (1975), pp. 363–428.

[160] J. MCCARTHY AND M. N. ZALD, *Social movement organizations*, The social movements reader: Cases and concepts, (2003), pp. 169–186.

[161] J. D. MCCARTHY, M. N. ZALD, G. LONG, A. OBERSCHALL, A. ORUM, K. PEARCE, J. SEIDMAN, AND B. WALTER, *Resource Mobilization and Social Movements: A Partial Theory'*, tech. rep.

[162] C. MCCAULEY AND S. JACQUES, *The popularity of conspiracy theories of presidential assassination: A bayesian analysis.*, Journal of Personality and Social Psychology, 37 (1979).

[163] C. MCCAULEY AND S. MOSKALENKO, *Mechanisms of political radicalization: Pathways toward terrorism*, Terrorism and Political Violence, 20 (2008), pp. 415–433.

[164] A. MCGRATH, *Dealing with dissonance: A review of cognitive dissonance reduction*, Social and Personality Psychology Compass, 11 (2017), p. e12362.

[165] B. M. MCKIMMIE, D. J. TERRY, M. A. HOGG, A. S. MANSTEAD, R. SPEARS, AND B. DOOSJE, *I'm a hypocrite, but so is everyone else: Group support and the reduction of cognitive dissonance.*, Group Dynamics: Theory, research, and practice, 7 (2003), p. 214.

[166] L. G. MCNAMEE, B. L. PETERSON, AND J. PEÑA, *A call to educate, participate, invoke and indict: Understanding the communication of online hate groups*, Communication Monographs, 77 (2010), pp. 257–280.

[167] D. T. MILLER, *Characterizing qanon: Analysis of youtube comments presents new conclusions about a popular conservative conspiracy*, First Monday, (2021).

[168] S. K. MITRA, *Adversarial politics and policy continuity: the upa, nda and the resilience of democracy in india*, Contemporary South Asia, 19 (2011), pp. 173–187.

[169] T. MITRA, S. COUNTS, AND J. W. PENNEBAKER, *Understanding anti-vaccination attitudes in social media*, in ICWSM, 2016.

[170] T. MITRA, G. WRIGHT, AND E. GILBERT, *Credibility and the Dynamics of Collective Attention*, Proceedings of the ACM on Human-Computer Interaction, 1 (2017), pp. 1–17.

[171] A. MORRIS, *Leadership in social movements aldon morris and suzanne staggenborg*, (2002).

[172] C. W. MORRIS, *Foundations of the theory of signs*, in International encyclopedia of unified science, Chicago University Press, 1938, pp. 1–59.

[173] M. B. MOUSSA, *Online mobilization in times of conflict: A framing-analysis perspective*, Arab and Media Society, (2013).

[174] T. E. NELSON, R. A. CLAWSON, AND Z. M. OXLEY, *Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance*, American Political Science Review, 91 (1997), pp. 567–583.

[175] L. S. NEO, *An Internet-Mediated Pathway for Online Radicalisation*, (2016), pp. 197–224.

[176] P. R. NEUMANN, *Options and Strategies for Countering Online Radicalization in the United States*, Studies in Conflict & Terrorism, 36 (2013), pp. 431–459.

[177] E. NEWELL, D. JURGENS, H. M. SALEEM, H. VALA, J. SASSINE, C. ARMSTRONG, AND D. RUTHS, *User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest*, Aaai, (2016), pp. 279–288.

[178] C. NEWS, *What is qanon? what does wwg1wga mean? the conspiracy theory that explains everything and nothing - cbs news*.
`https://www.cbsnews.com/news/what-is-the-qanon-conspiracy-theory/`, November 2020.
(Accessed on 02/04/2021).

[179] P. C. NOW, *Introduction to activism*, Recuperado de< http://www. permanentculturenow. com/what-is-activism, (2018).

[180] S. NÚÑEZ PUENTE, D. FERNÁNDEZ ROMERO, AND S. VÁZQUEZ CUPEIRO, *Online feminist practice, participatory activism and public policies against gender-based violence in spain*, Feminist Theory, 18 (2017), pp. 299–321.

[181] J. A. OBAR, P. ZUBE, AND C. LAMPE, *Advocacy 2.0: An analysis of how advocacy groups in the united states perceive and use social media as tools for facilitating civic engagement and collective action*, Journal of information policy, 2 (2012), pp. 1–25.

[182] A. OBERSCHALL, *Social conflict and social movements*, Prentice hall, 1973.

[183] D. O'CALLAGHAN, D. GREENE, M. CONWAY, AND J. CARTHY, *Uncovering the Wider Structure of Extreme Right Communities Spanning Popular Online Networks*, tech. rep.

[184] F. OLSSON, *A literature survey of active machine learning in the context of natural language processing*, (2009).

[185] L. OSWALD AND J. BRIGHT, *How do climate change skeptics engage with opposing views? understanding mechanisms of social identity and cognitive dissonance in an online forum*, arXiv preprint arXiv:2102.06516, (2021).

[186] N. OWEN, *The conscience constituent reconsidered*, in Other People's Struggles, Oxford University Press.

[187] D. PACHECO, A. FLAMMINI, AND F. MENCZER, *Unveiling coordinated groups behind white helmets disinformation*, in Companion Proceedings of the Web Conference 2020, 2020, pp. 611–616.

[188] A. PANDA, A. GONAWELA, S. ACHARYYA, D. MISHRA, M. MOHAPATRA, R. CHANDRASEKARAN, AND J. PAL, *Nivaduck-a scalable pipeline to build a database of political twitter handles for india and the united states*, in International Conference on Social Media and Society, 2020, pp. 200–209.

[189] A. PAPASAVVA, J. BLACKBURN, G. STRINGHINI, S. ZANNETTOU, AND E. DE CRISTOFARO, *" is it a qoincidence?": A first step towards understanding and characterizing the qanon movement on voat. co*, arXiv preprint arXiv:2009.04885, (2020).

[190] W. C. PARTIN AND A. E. MARWICK, *The construction of alternative facts: Dark participation and knowledge production in the qanon conspiracy*, AoIR Selected Papers of Internet Research, (2020).

[191] U. PAVALANATHAN AND M. DE CHOUDHURY, *Identity management and mental health discourse in social media*, in WWW, ACM, 2015.

[192] J. W. PENNEBAKER, *Linguistic inquiry and word count: Liwc 2001*, (2001).

[193] G. PENNYCOOK, Z. EPSTEIN, M. MOSLEH, A. A. ARECHAR, D. ECKLES, AND D. G. RAND, *Shifting attention to accuracy can reduce misinformation online*, Nature, 592 (2021), pp. 590–595.

[194] T. PETROVA AND S. TARROW, *Transactional and participatory activism in the emerging european polity: The puzzle of east-central europe*, Comparative political studies, 40 (2007), pp. 74–94.

[195] S. PHADKE AND T. MITRA, *Many faced hate: A cross platform study of content framing and information sharing by online hate groups*, (2020), pp. 1–13.

[196] ——, *Educators, solicitors, flamers, motivators, sympathizers: Characterizing roles in online extremist movements*, Proceedings of the ACM on Human-Computer Interaction, (2021).

[197] S. PHADKE, M. SAMORY, AND T. MITRA, *What makes people join conspiracy communities? role of social factors in conspiracy engagement. Best Paper Honorable Mention*, Proceedings of the ACM on Human-Computer Interaction, 4 (2020), pp. 1–30.

[198] ——, *Characterizing social imaginaries and self-disclosures of dissonance in online conspiracy discussion communities.Best Paper Honorable Mention*, Proceedings of the ACM on Human-Computer Interaction, (2021).

[199] ——, *Pathways through conspiracy: The evolution of conspiracy radicalization through engagement in online conspiracy discussions*, (2022).

[200] C. PIGDEN, *Popper Revisited, or What Is Wrong With Conspiracy Theories?*, Philosophy of the Social Sciences, 25 (1995), pp. 3–34.

[201] T. D. PILDITCH, J. K. MADSEN, AND R. CUSTERS, *False prophets and cassandra's curse: The role of credibility in belief updating*, Acta psychologica, 202 (2020), p. 102956.

[202] S. PLANCK, *Where we go one, we go all: Qanon and violent rhetoric on twitter*, Locus: The Seton Hall Journal of Undergraduate Research, 3 (2020), p. 11.

[203] D. PRIOLO, A. PELT, R. S. BAUZEL, L. RUBENS, D. VOISIN, AND V. FOINTIAT, *Three decades of research on induced hypocrisy: A meta-analysis*, Personality and Social Psychology Bulletin, 45 (2019), pp. 1681–1701.

[204] T. T. T. PROJECT, *Facebook,Äôs boogaloo problem: A record of failure | tech transparency project*.
https://www.techtransparencyproject.org/articles/facebooks-boogaloo-problem-record-failure, August 2020.
(Accessed on 10/15/2020).

[205] B. RAHIMI, *Vahid online: Post-2009 iran and the politics of citizen media convergence*, Social Sciences, 5 (2016), p. 77.

[206] S. RAJGARHIA, *Media Manipulation in the Indian Context: An Analysis of Kashmir-related Discourse on Twitter*, PhD thesis, Harvard University, 2020.

[207] J. REEDY, J. GASTIL, AND M. GABBAY, *Terrorism and small groups: An analytical framework for group disruption*, Small group research, 44 (2013), pp. 599–626.

[208] N. REIMERS AND I. GUREVYCH, *Making monolingual sentence embeddings multilingual using knowledge distillation*, 01 2020, pp. 4512–4525.

[209] H. RHEINGOLD, *Smart mobs: The next social revolution*, Basic books, 2007.

[210] A. ROBERTSON, L. M. AIELLO, AND D. QUERCIA, *The language of dialogue is complex*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 428–439.

[211] D. RODAN AND J. MUMMERY, *Activism and digital culture in Australia*, Rowman & Littlefield, 2017.

[212] M. ROTHSCHILD, *Qanon followers have limited options after reddit ban*.
https://www.dailydot.com/debug/qanon-movement-reddit-ban-voat-facebook/, September 2018.
(Accessed on 07/10/2021).

[213] K. SAHA AND A. SHARMA, *Causal factors of effective psychosocial outcomes in online mental health communities*, (2020).

[214] K. SAHA, B. SUGAR, J. TOROUS, B. ABRAHAO, E. KICIMAN, AND M. DE CHOUDHURY, *A social media study on the effects of psychiatric medication use*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 440–451.

[215] K. SAMBHAV AND N. RANGANATHAN, *How a reliance-funded firm boosts bjp,Äôs campaigns on facebook | business and economy news | al jazeera*. https://www.aljazeera.com/economy/2022/3/14/how-a-reliance-funded-company-boosts-bjps-campaigns-on-facebook, March 2022. (Accessed on 04/23/2022).

[216] M. SAMORY AND T. MITRA, *Conspiracies online: User discussions in a conspiracy community following dramatic events*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 12, 2018.

[217] ——, *'the government spies using our webcams' the language of conspiracy theories in online discussions*, Proceedings of the ACM on Human-Computer Interaction, 2 (2018), pp. 1–24.

[218] R. SAURI, *A factuality profiler for eventualities in text*, PhD thesis, 2008.

[219] J. A. SCHAFER, *Spinning the web of hate: Web-based hate propagation by extremist organizations*, Journal of Criminal Justice and Popular Culture, 9 (2002), pp. 69–88.

[220] B. F. SCHAFFNER AND P. J. SELLERS, *Winning with words: the origins and impact of political framing*, Routledge, 2009.

[221] E. SCHLUETER AND E. DAVIDOV, *Contextual sources of perceived group threat: Negative immigration-related news reports, immigrant group size and their interaction, spain 1996–2007*, European Sociological Review, 29 (2011), pp. 179–191.

[222] D. SCHOCH, F. B. KELLER, S. STIER, AND J. YANG, *Coordination patterns reveal online political astroturfing across the world*, Scientific reports, 12 (2022), pp. 1–10.

[223] B. SETTLES, *Active Learning Literature Survey*, tech. rep., 2009.

[224] S. SHEATHER, *A modern approach to regression with R*, Springer Science & Business Media, 2009.

[225] K. SHU, S. WANG, J. TANG, R. ZAFARANI, AND H. LIU, *User Identity Linkage across Online Social Networks: A Review*, tech. rep.

[226] A. A. SIEGEL AND V. BADAAN, # *no2sectarianism: Experimental approaches to reducing sectarian hate speech online*, American Political Science Review, 114 (2020), pp. 837–855.

[227] B. SIMON AND P. KLANDERMANS, *Toward a social psychological analysis of politicized collective identity: Conceptualization, antecedents and consequences*, American Psychologist, 56 (2001), pp. 319–331.

[228] L. SIMON, J. GREENBERG, AND J. BREHM, *Trivialization: the forgotten mode of dissonance reduction.*, Journal of personality and social psychology, 68 (1995), p. 247.

[229] M. SINGH, *India objects to 'manipulated' label on politicians tweets; asks removal of reference to 'indian variant' of coronavirus | techcrunch*. https://techcrunch.com/2021/05/21/india-twitter-politicians-tweets-manipulated/, May 2021. (Accessed on 01/14/2023).

[230] B. SMITH, *More than half of democrats believed bush knew - politico*. https://www.politico.com/blogs/ben-smith/2011/04/more-than-half-of-democrats-believed-bush-knew-035224, 04 2011. (Accessed on 05/24/2020).

[231] D. SNOW, A. TAN, AND P. OWENS, *Social movements, framing processes, and cultural revitalization and fabrication*, Mobilization: An International Quarterly, 18 (2013), pp. 225–242.

[232] D. A. SNOW, *Framing Processes, Ideology, and Discursive Fields*, Blackwell Publishing, Oxford, United Kingdom, 2004.

[233] D. A. SNOW, R. D. BENFORD, ET AL., *Ideology, frame resonance, and participant mobilization*, International social movement research, 1 (1988), pp. 197–217.

[234] S. SOHAL AND H. KAUR, *A content analysis of youtube political advertisements: evidence from indian parliamentary elections*, Journal of creative communications, 13 (2018), pp. 133–156.

[235] SPLC, *Alliance defending freedom | southern poverty law center*. https://www.splcenter.org/fighting-hate/extremist-files/group/alliance-defending-freedom, 2020. (Accessed on 10/15/2020).

[236] K. STARBIRD, *Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter*, in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, 2017.

[237] K. STARBIRD, A. ARIF, AND T. WILSON, *Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations*, Proceedings of the ACM on Human-Computer Interaction, 3 (2019), pp. 1–26.

[238] C. STEMPEL, T. HARGROVE, AND G. H. STEMPEL III, *Media use, social structure, and belief in 9/11 conspiracy theories*, Journalism & Mass Communication Quarterly, 84 (2007).

[239] A. STOJANOV, J. M. BERING, AND J. HALBERSTADT, *Does perceived lack of control lead to conspiracy theory beliefs? findings from an online mturk sample*, PLOS ONE, 15 (2020), pp. 1–18.

[240] J. STONE AND N. C. FERNANDEZ, *To practice what we preach: The use of hypocrisy and cognitive dissonance to motivate behavior change*, Social and Personality Psychology Compass, 2 (2008), pp. 1024–1051.

[241] S. STÜRMER, B. SIMON, M. LOEWY, AND H. JÖRGER, *The dual-pathway model of social movement participation: The case of the fat acceptance movement*, Social Psychology Quarterly, (2003), pp. 71–82.

[242] P. SUEDFELD AND P. E. TETLOCK, *27 conceptual/integrative complexity*, (1992).

[243] C. R. SUNSTEIN AND A. VERMEULE, *Conspiracy theories: Causes and cures*, Journal of Political Philosophy, 17 (2009).

[244] H. TAKIKAWA AND T. SAKAMOTO, *The moral–emotional foundations of political discourse: a comparative analysis of the speech records of the us and the japanese legislatures*, Quality & Quantity, (2019), pp. 1–20.

[245] Y. R. TAUSCZIK AND J. W. PENNEBAKER, *The psychological meaning of words: Liwc and computerized text analysis methods*, Journal of language and social psychology, 29 (2010).

[246] C. TAYLOR, *Modern social imaginaries*, Public culture, 14 (2002), pp. 91–124.

[247] N. TERKILDSEN AND F. SCHNELL, *How Media Frames Move Public Opinion: An Analysis of the Women's Movement*, Political Research Quarterly, 50 (1997), pp. 879–900.

[248] E. TIAN, *The qanon timeline: Four years, 5,000 drops and countless failed prophecies - bellingcat*.
https://www.bellingcat.com/news/americas/2021/01/29/the-qanon-timeline/, January 2021.
(Accessed on 03/16/2021).

[249] S. TIWARI, *Elections 2019: 45 million new, young voters could play a key role in 2019 ,Äì and they want jobs*.
https://scroll.in/article/913411/data-check-45-million-new-young-voters-could-play-a-key-role-in-2019-elections, February 2019.
(Accessed on 03/09/2022).

[250] R. TOPINKA, *The politics of anti-discourse: Copypasta, the alt-right, and the rhetoric of form*, Theory & Event, 25 (2022), pp. 392–418.

[251] R. H. TURNER, *The public perception of protest*, American Sociological Review, (1969), pp. 815–831.

[252] V.-P. TYNKKYNEN AND N. TYNKKYNEN, *Climate denial revisited:(re) contextualising russian public discourse on climate change during putin 2.0*, Europe-Asia Studies, 70 (2018), pp. 1103–1120.

[253] T. ULRICH, *How social media contributes to vaccine hesitancy | technology networks*.
https://www.technologynetworks.com/vaccines/news/how-social-media-contributes-to-vaccine-hesitancy-358752, February 2022.
(Accessed on 05/05/2022).

[254] J. USCINSKI AND C. KLOFSTAD, *New poll: the qanon conspiracy movement is very unpopular - the washington post*.
https://www.washingtonpost.com/news/monkey-cage/wp/2018/08/30/the-qanon-conspiracy-movement-is-very-unpopular-our-new-poll-finds/, 08 2018.

(Accessed on 05/24/2020).

[255] J. E. USCINSKI, C. KLOFSTAD, AND M. D. ATKINSON, *What drives conspiratorial beliefs? the role of informational cues and predispositions*, Political Research Quarterly, 69 (2016), pp. 57–71.

[256] I. VAGHEFI AND H. QAHRI-SAREMI, *From it addiction to discontinued use: A cognitive dissonance perspective*, (2017).

[257] S. VALENZUELA, *Unpacking the use of social media for protest behavior: The roles of information, opinion expression, and activism*, American behavioral scientist, 57 (2013), pp. 920–942.

[258] J.-W. VAN PROOIJEN, K. M. DOUGLAS, AND C. DE INOCENCIO, *Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural*, European journal of social psychology, 48 (2018), pp. 320–335.

[259] J.-W. VAN PROOIJEN AND N. B. JOSTMANN, *Belief in conspiracy theories: The influence of uncertainty and perceived morality*, European Journal of Social Psychology, 43 (2013), pp. 109–115.

[260] N. VAN RAEMDONCK, *The echo chamber of anti-vaccination conspiracies: mechanisms of radicalization on facebook and reddit*, IPAG Knowledge Series, (2019).

[261] J. VAN STEKELENBURG AND B. KLANDERMANS, *Social psychology of movement participation*, The Wiley-Blackwell Encyclopedia of Social and Political Movements, (2013).

[262] J. VAN STEKELENBURG AND B. KLANDERMANS, *The social psychology of protest*, Current Sociology, 61 (2013), pp. 886–905.

[263] ——, *Individuals in movements: A social psychology of contention*, in Handbook of social movements across disciplines, Springer, 2017, pp. 103–139.

[264] C. A. VERMEULE AND C. R. SUNSTEIN, *Conspiracy theories: causes and cures*, Journal of Political Philosophy, (2009).

[265] B. VIDGEN, T. YASSERI, AND H. MARGETTS, *Trajectories of Islamophobic hate amongst far right actors on Twitter*, pp. 1–20.

[266] J. A. VITRIOL AND J. K. MARSH, *The illusion of explanatory depth and endorsement of conspiracy beliefs*, European Journal of Social Psychology, 48 (2018).

[267] M. WAHLSTRÖM, A. PETERSON, AND M. WENNERHAG, *‚Äúconscience adherents‚Äù revisited: Non-lgbt pride parade participants*, Mobilization: An International Quarterly, 23 (2018), pp. 83–100.

[268] I. WALLER AND A. ANDERSON, *Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms*, in The World Wide Web Conference, 2019, pp. 1954–1964.

[269] Y.-C. WANG, M. BURKE, AND R. KRAUT, *Modeling self-disclosure in social networking sites*, in Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, 2016, pp. 74–85.

[270] Y.-C. WANG, R. KRAUT, AND J. M. LEVINE, *To stay or leave? the relationship of emotional and informational support to commitment in online health support groups*, in Proceedings of the ACM 2012 conference on computer supported cooperative work, 2012, pp. 833–842.

[271] D. WEBER AND F. NEUMANN, *Amplifying influence through coordinated behaviour in social networks*, Social Network Analysis and Mining, 11 (2021), pp. 1–42.

[272] J. WIEBE, T. WILSON, AND C. CARDIE, *Annotating expressions of opinions and emotions in language*, Language resources and evaluation, 39 (2005), pp. 165–210.

[273] J. WILSON, *Introduction to social movements*, Basic Books, 1973.

[274] M. J. WOOD AND K. M. DOUGLAS, *Online communication as a window to conspiracist worldviews*, Frontiers in psychology, 6 (2015), p. 836.

[275] M. J. WOOD, K. M. DOUGLAS, AND R. M. SUTTON, *Dead and alive: Beliefs in contradictory conspiracy theories*, Social psychological and personality science, 3 (2012), pp. 767–773.

[276] J. K. WOOLLEY, A. M. LIMPEROS, AND M. B. OLIVER, *The 2008 presidential election, 2.0: A content analysis of user-generated political facebook groups*, Mass Communication and Society, 13 (2010), pp. 631–652.

[277] E. WULCZYN, N. THAIN, AND L. DIXON, *Ex Machina: Personal attacks Seen at Scale*, in Proceedings of the 26th International Conference on World Wide Web - WWW '17, New York, New York, USA, 2017, ACM Press, pp. 1391–1399.

[278] D. YIN, Z. XUE, L. HONG, B. D. DAVISON, A. KONTOSTATHIS, AND L. EDWARDS, *Detection of Harassment on Web 2.0*, in Proceedings of the Content Analysis in the WEB 2.0 Workshop at WWW2009 - CAW2.0, vol. 2, 2009, pp. 1–7.

[279] W. YUAN, Y. HAN, D. GUAN, S. LEE, AND Y.-K. LEE, *Initial training data selection for active learning*, in Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, 2011, pp. 1–7.

[280] M. N. ZALD AND R. ASH, *Social movement organizations: Growth, decay and change*, Social forces, 44 (1966), pp. 327–341.

[281] X. ZHOU, X. LIANG, H. ZHANG, AND Y. MA, *Cross-Platform Identification of Anonymous Identical Users in Multiple Social Media Networks*, IEEE Transactions on Knowledge and Data Engineering, 28 (2016), pp. 411–424.

[282] Y. ZHOU, E. REID, J. QIN, H. CHEN, AND G. LAI, *Us domestic extremist groups on the web: link and content analysis*, IEEE intelligent systems, 20 (2005), pp. 44–51.

[283] M. ZONIS AND C. M. JOSEPH, *Conspiracy Thinking in the Middle East*, Political Psychology, 15 (1994), pp. 443–459.