GOVERNANCE, COMPETITION, & EXTREMISM: HOW THE STRUCTURE OF SOCIAL MEDIA PLATFORMS RADICALIZES COMMUNITIES

By

Colin Michael Henry

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Political Science

December 16, 2023

Nashville, Tennessee

Approved:

Jennifer Larson, Ph.D.

Cassy Dorff, Ph.D.

Emily Ritter, Ph.D.

Anita Gohdes, Ph.D.

ACKNOWLEDGMENTS

No one goes anywhere alone. We all carry and are carried by a mycelial web of acquaintances, colleagues, friends, mentors, and family. It is through this network that any one of us is made whole.

First, a boundless thank you to Dr. Cassy Dorff, without whom I wouldn't be within a Texas mile of a PhD. I am grateful to have found a mentor and friend like Cassy, and lucky to have worked alongside such a kind and brilliant person. Dr. Dorff has taught me many things in the nearly 8 years we have worked together: how to work with humility and care for the communities and people that we study, to walk thoughtfully through these spaces and do our best to leave them better than we found them. How to build productive and serious scientific community through empathy and joy, rather than at their expense. How to be brave, adamant in the defense of our earnest and poetic selves in a profession that often threatens to flatten our many dimensions. Also binders and binders of important facts about horses and birds. Thank you.

I was also supported through this long process by an incredible dissertation committee. Few of us are fortunate enough to work with their heroes as I did. Thank you to Drs. Jennifer Larson, Emily Hencken Ritter, and Anita Gohdes, who endured drafts of questionable quality, presentations full of incomprehensible Internet jargon, and many, many, many emails. Without the generosity of their time and feedback, none of this would be possible. I want to especially thank Jenn for not putting me on an ice floe when writing had slowed to a snail's pace. And Emily for her guidance and patience with an unorthodox transfer student of questionable quality. And, finally, Anita, with whom I'd dreamed of collaborating long before I got to Vanderbilt, for donating many Berlin evenings to our meetings and discussions.

I owe much to the wider political science community—in particular my colleagues in the Conflict Workshop at Vanderbilt University. Specifically Drs. Brenton Kenkel and Brad Smith, who are some of the finest teachers I've had the pleasure of learning from in my long student career. The Workshop and the conflict community have provided endless opportunities for feedback and professionalization for which I am very grateful. I also want to acknowledge the Political Science Department at the University of New Mexico. In particular, Drs. Christopher Butler, Jessica Feezell, Jamie Nelson-Nunez, and Kendra Kouivu, who gave me far more of their time, advice, and expertise than I deserved.

Graduate school can be an isolating and totalizing experience without excellent grad worker colleagues. I've had the privilege of befriending and working with some of the best: Lauren Chojnacki, without whom I would quite literally not have survived our tornado-pandemic-insurrection run through grad school. Thank you for many hours spent over tiny cups of coffee, pushing me to accept all things Taylor Swift, and for keeping Harper company when I could not. Fiorella Vera-Adrianzen, my mentor and friend, for years of love, support, and Double Whites. And Alex Lawhorne: one of the finest people I've met, from your fact-filled tour of Nashville on my prospective student visit to an endless font of Star Trek wisdom.

I owe everything to my family: Joel, Kelley, Matt, Jess, Rowan, and Asher Henry. My parents, who have supported every wacky and insane decision I've ever made, helped move me across country damn near a dozen times, and celebrated my achievements as their own: all of the best parts of myself came from you, and I am so grateful to be your son. Thank you to Matt and Jess for your strength and support for many years and across many parts

of the country.

Thank you to my wife, Dr. Jessica Jones, my one and only person. We tried to finish two PhDs in two years: not easy for anybody. But we did it, amidst some of the more challenging circumstances one could imagine. Thank you for your many sacrifices to get us both where we are today. Thank you for moving across the country and into a blizzard of uncertainty so I could be a part of Vanderbilt. And thank you for being steadfast for me through the challenges of this dissertation. I'm constantly in awe of your strength, resilience, and bravery. I love you, and I look forward to the rest of our lives together outside of graduate school.

P.S. Shout-out to my dog Harper, who is a very good girl.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In the summer of 2015, the new CEO of the social media and link aggregation website Reddit banned five communities or “subreddits” for violating the platform’s anti-harassment and “revenge porn” policies. Ellen Pao, with law and MBA degrees from Harvard, joined the company in 2013 and assumed the interim CEO position in 2014. Pao is also a woman and second-generation Taiwanese immigrant. The communities her administration banned that summer included subreddits that organized harassment of women, fat people, transgender people, and Black, brown, and Asian online communities. Reddit administrators had shut down other subreddits, which operate semi-autonomously under volunteer moderators, in the past, and banning just five of the nearly 650,000 communities seemed relatively benign. But in the months that followed, users of the banned subreddits flooded other communities with violent content aimed at their original targets—and Pao. Over the summer, the interim CEO received thousands of racist and sexist messages, many containing personal information and death threats. The “front page” of Reddit—the default landing for the site containing the top user content as ranked by an algorithm—regularly featured violent threats towards and crude images of the platform’s top executive. Pao was conciliatory and diplomatic to Redditors, posting daily announcements explaining policy changes, describing the direction of future policies, and soliciting feedback from users. Nonetheless, Pao resigned by the middle of July in what may be the first user-led regime change of a major social media platform. Her replacement, Reddit co-founder Steve Huffman, would preside over one of the fastest growing hate-speech and harassment communities on the internet, “The_Donald” (158). Composed of many users and moderators from the subreddits banned in 2015, The_Donald would play a key role in organizing ”swatting” and offline harassment campaigns, fabricating responses to online Lucid and Pew Research polls during the 2016 US election, praising the 2016 Orlando nightclub shooting, coordinating the 2017 Unite the Right rally, and spreading videos of the Christchurch mosque shooting in New Zealand (176). This subreddit would not be banned until late 2020.

In an interview just before assuming the CEO role, Pao spoke at length about building infrastructure to reward positive participation on Reddit, arguing that ”Reddit is about communities” and that her role was to “build a community that lasts for many, many years” rather than engage in short-term monetization (102). Users of communities banned in 2015 and their supporters, however, complained that her actions in establishing and executing new policies against harassment and revenge porn were authoritarian and exclusionary. They charged Pao with favoring the preferences of advertisers over their freedom to create and exchange online content. Pao's Reddit administration did not have any recognizably democratic method of aggregating preferences on policies. But the truth is more complex. Post-hoc studies showed that the communities banned under Pao's new policies were outsized contributors to a huge harassment and incitement problem on the platform (32); users from these spaces regularly intruded on other communities, drowning out expression with spam, threats, and violent imagery. What Pao's administration identified was a problem with how attention was allocated on Reddit. Users from a few communities exploited gaps in the political institutions and processes of the platform to gather more than their share of public goods; and when Pao closed those gaps, they reacted by turning to even greater levels of political extremism and violence.

In this project, I argue that social media platform governance drives variation in online political extremism at the community level. My argument makes two claims: first, social media platforms are a kind of private governance, a new iteration of non-Weberian political authority specializing in control over information and access to public goods like attention and engagement. Whereas much of the work on platform governance has focused on interaction with existing state (primarily Western) legal systems or new transformations of human-centered design and social organization, the *political* institutions and processes of digital platforms are best understood within the framework of competitive state-building. This approach re-frames platforms as the result of progressive consolidation around and monopolization of information and attention. This, I argue, is the source of authority, legitimacy, and sovereignty for social media platforms. Technically, platforms have absolute sovereignty over information: control over code means control over bits. However, their claims to authority and legitimacy vary widely based on the political institutions and practices through which they exercise sovereignty. It is this political environment–and the authority and legitimacy that it generates–that can determine the network structure and level of extremism in online political communities. Second, I argue that the mechanism that connects platform political regimes and extremism in online spaces is *community competition for public goods.* When platforms govern through the algorithmic imposition of small, angry micro-identities and pit them against each other in a contest for information, attention, and engagement, political communities tend towards extremist ideological commitments and behaviors.

Chapter 2, "Varieties of Platform Governance," takes on the dearth of systematic observation platform regime types. How does platform governance work, and how does it vary—conceptually and empirically—between platforms? To systematically study the development and effects of platform regimes, we need systematic and reliable classifications. Cross-national methods of comparison have long been a useful tool for concept formation, measurement, and limited causal inference in the international relations and comparative politics subfields in political science (see, e.g., (126; 103; 37)). Platform studies and political communication work on the political internet, however, has thus far lacked systematic measurement and data collection across many cases, making cross-case comparison difficult.

This chapter makes progress on solving this problem by developing a typology of platform regime and building a cross-platform dataset. I collect data on on content policies and processes, product design and deployment, and investor disclosures for four contemporary social media platforms (Twitter, TikTok, Reddit, and Youtube). I show that platforms vary according to how they make moderation decisions, the level of autonomy afforded communities, and the normative framework of discourse. These three principles of platform governance provide the conceptual basis to compare how different types of platform governance affect the radicalization pathways of online political communities.

Chapter 3, "Community Sorting Across Platforms", considers how platform regime type and the structure of the social media space interact to sort online communities. This chapter solves the puzzle of why online political communities pursue strategies of radicalization across platforms. Many existing approaches to studying radicalization and extremism adopt individualized or whole-of-platform frames. I argue that communities—those loosely organized, basic units of political movements—offer another level of analysis to understand how political extremism grows online. This chapter focuses on the strategic interaction between communities and platforms across the social internet system.

The setup of "Community Sorting" adapts a classic "voting with your feet" model of public good distribution to explain how community forum shopping happens. My computational variation of Tiebout sorting shows that different types of platform regimes differ in their ability to deliver utility to non-extremist and extremist political communities alike. In particular, I show that governance types with limited community autonomy, exclusionary moderation systems, and competition for allocating attention perform poorly when there are few platforms available, and that this poor performance has consequences for community extremism. I also show that the system of platforms, and the composition of platform types available to communities, influences the strategic behavior of extremist communities.

Chapter 4, "Community Competition and Political Extremism," tackles the microfoundations of this model. In this paper, I situate the dissertation in historical context of developing telecommunications architecture, from early Usenet message boards to contemporary multi-modal social platforms, and the emergence of online political extremism as a modern form of political violence. I explain how legalistic frameworks of content moderation and the relationship between platforms and states dominates the literature on platform governance. As an alternative, I suggest a theory explaining where power, authority, legitimacy, and sovereignty come from and how they function in private governance. I look to recent research from a wide spectrum of substantive areas that articulates new forms of political authority outside of (but not necessarily in competition with) the traditional Weberian state. Research on non-state armed actors and militias, for example, shows that non-state public service provision can generate alternative sources of legitimacy, authority, and sovereignty (192; 130). Indigenous communities reclaiming their autonomy from corrupt or ineffective state structures offer real, practical cases of non-state authority in action (57; 135). Traditional states facing collapse or threat of collapse from exogenous forces like climate change face gaps in legitimacy and authority that are often filled by non-Weberian actors. I draw on this long tradition of competitive state-building work in political science and sociology to understand social media platforms as objects of political contestation for their users.

This chapter also clarifies how political extremism is a distinct from civil conflict, terrorism, one-sided violence, and other similar types of political violence, providing clear definitions and examples. I review the past two decades of terrorism literature, showing how typical theoretical approaches in political science towards non-state violence are inadequate to understand political extremism. Since 2017, so-called "post-organizational" extremism has largely overtaken organized terror by non-state groups, especially for "domestic" political violence in Western states (45; 205). New groups with traditional membership structures are rare, short-lived, and rapidly adopt similar aesthetics or ideological commitments to past incumbents. Unlike traditional terror groups, these extremist movements are often ideologically complex, with contradictory political claims and identities. Spatial models of conflict or negotiation are inadequate or incomprehensible when imposed on this new online class of aesthetically and ideologically motivated actors.

Using relational data from users participating in political communities on four social media platforms, I show that communities express different levels of network-wide extremism when subject to different types of platform governance. When platforms encourage conflict between communities to generate attention and engagement, those political communities adopt more extremist rhetoric, ideological claims, and political identities. In particular, platform institutions and processes that are less inclusive, autonomous, and transparent produce, network-wide, more radical and violent political communities. To demonstrate this, "Community Competition" uses an empirical strategy that combines months of digital fieldwork in the white power, neo-sexist, and Qanon conspiracy extremist movements with big data collection on four social media platforms. I show that communities associated with these movements express a greater share of extremist content and identity affiliation when platform rules force out-group users into competition and conflict with in-group users.

In short, the way platforms develop authority, legitimacy, and sovereignty matters for which communities radicalize. How platforms try to make users intelligible by sorting them into angry, competing micro-identities determines what communities express greater levels of political extremism. It is not enough to focus on the legal frameworks that inform content moderation policies or the interactions between multi-national tech companies and nationstates; and it is not enough to understand individual radicalization pathways. Rather, to understand the proliferation of extremist politics online, we need to understand how communities move and change in response to the political institutions and processes of platforms. When platforms deny online communities access to these institutions and processes, they turn to extremism.

CHAPTER 2

Varieties of Platform Governance

Globally, more people experience day-to-day coercive interactions with social media governance than they do with state agents. Meta Platforms, the parent company of Facebook and WhatsApp, commands the attention of nearly 3 billion and 2 billion users, respectively. And yet, our understanding of the political nature of these quasi-sovereign corporate entities is surprisingly limited. Nearly two decades ago, Barbara Geddes argued that political science's predominant focus on democracy left us with "few shoulders of giants on which to stand" for the study of autocracy and authoritarianism (Geddes). We similarly lack foundational texts for conceptualizing and measuring differences between private governance broadly and social media platform regimes more specifically. Classifying platforms along classic dimensions of democracy is fairly easy: no corporate platform has recognizable democratic governance for users.[1] But the lack of systematic study of interior political institutions and processes on platforms make more nuanced distinctions difficult. Is Facebook more autocratic than Twitter? Does Reddit have more exclusionary governance now than in 2016? Do communities on Tumblr have more autonomy in some ways, but less in others?

This paper presents a new dataset that aims to conceptualize, measure, and categorize platforms as intrinsic political entities. Because few efforts exist to build a kind of typology for platform regime type, this project draws on platform studies, state development, and the political theory of corporate governance literatures to impose three principles of governance on the landscape of social media platforms. These definitions are not meant to represent the consensus on how these entities operate; rather, this effort is the first sentence in a new conversation on conceptualizing and measuring the political institutions and processes that comprise the interior functions of social media platforms.

As a preliminary (but not comprehensive) attempt at operationalizing these principles,

[1] Federated and open-source platforms do present a more diverse set of regime types, ranging from benevolent dictatorships to anarchic collectives. These entities are understudied because of their size and position in the information landscape, and deserve more attention in the political science and media studies literature. However, they are not the focus of this paper.

I collect primary source information on the structure and function of institutions within a handful of Western social media platforms. This data comes from three sources: security, privacy, and terms of service policies; product announcements and release statements; and financial disclosures filed with the Security and Exchange Commission or sent to investors. This dataset covers the lifespans of Facebook, Reddit, Twitter, YouTube, and TikTok. From this corpus, I trace the development of content moderation, community autonomy, and discourse architecture institutions and processes.

2.1 Three Principles of Platform Government

Conceptual frameworks support measurement schemes. As Sartori argues, "concept formation stands prior to quantification" (167, 1038). In this section, I suggest a conceptual foundation for developing measurement and data on varieties of platform governance. There are two central challenges: first, adapting taxonomies devised from studying states and state institutions to private governance; second, comporting conceptual primitives primarily based on material interaction—physical violence, economic exchange, and so forth—to a world that is primarily based on informational transactions. A typology of platforms resists classical categorization on the "ladder of generality" (167), largely due to the many concerns raised by interpretivists skeptical of a true comparative method (see, for example, (81) or (162)). Classical categorization emphasizes moving between specific categories with limited extension and high intension to more general categories with high extension and limited intension (38). However, this framework requires relatively homogeneous political and social phenomena, detailed case knowledge, and clear boundaries between categories. Social media platforms fail on many of these requirements. Despite isomorphism between platform product offerings, their actual political institutions and processes remain relatively heterogeneous. So-called "big data" approaches focus on individual interactions of users, making direct systematic observation of platforms themselves rare. And lack of consensus over defining "platform" makes clear conceptual boundaries difficult to come by. Rather, my approach draws from the idea of "family resemblance" categories, in which we can recognize categories of similarity based on an array of commonalities even when there is no single shared attribute among cases (38; 211).

To that end, I offer three different dimensions: moderation, autonomy, and discourse. These high-level measures capture concepts of core government functions and should be recognizable to political scientists. The concept of *moderation* refers to the exclusivity and responsiveness to governed communities of political institutions that control moderation decisions. The ability of constituents to participate in the political process or hold elites accountable is fundamental to identifying different types of political rule, and reflects concepts captured in the study of both polyarchy (44) and anarchy (80). The principle of *autonomy* captures the degree of centralization or decentralization of political power on the platform. While "decentralization" in platform and technology studies often refers to where code is executed and who has control over computational power, I use it here to reflect the ability of communities to control political institutions and processes within the platform. Finally, the dimension of *discourse* captures the normative framework constructed by the technological environment. This concept reflects work on the quality of deliberation in democratic states (40; 43; 93) as well as norms of political discussion (74).

2.1.1 Moderation

Can users participate in moderation decisions, and are there mechanisms of accountability for decision-makers? How are moderation decisions implemented, and how visible are they to users? The *moderation* principle of platform governance is perhaps the most studied component of platform governance. The institutions, processes, and mechanisms of moderation vary widely over time and from platform to platform. However, they are all chiefly concerned with the practice of screening user-generated content (UGC) (165). Moderation requires an agent of the platform—a direct agent, third-party, automated system, or some combination of these—to enforce rules about what information is allowed on the platform, typically by interacting directly with the UGC. Thus, moderation has three components: a collection of informal and formal rules; an enforcement agency; and a process for executing rules on UGC.

Moderation is central to a platform's sovereignty and authority over the communities that fill their digital spaces. Proscription of legitimate information and enforcement of those prohibitions is how the codified computational infrastructure of a platform—and its absolute control over bits—is translated into control over more ephemeral forms of information (and thus users). This form of control is not isolated to social media platforms, and is an intrinsic feature of any networked computational architecture. See, for example, the "platform wars" between competing personal computers in the 1980s, competition between search engines and link aggregation pages in the 1990s and early 2000s, and early social media platforms before the introduction of the smartphone (82).

Why do online communities and users give up this control to platforms? And why do platforms—from hardware architecture to search to social media—perched atop the protocol layer provide services in return for power over users? One answer is profit or capital accumulation. Like the so-called "stationary-bandit" (152), platforms may derive greater rents from monopolizing information services for the communities that inhabit these spaces. Communities, meanwhile, enjoy the convenience and consistency of the "rational monopolization of theft"—of their information, control over content, and access to audiences—rather than the chaotic nature of uncoordinated competitive theft.

But platforms are not directly profitable. Early pre-smartphone social media platforms like MySpace and Friendster struggled to generate profit even at peak usage. Google loses money on YouTube; renting cloud services and serving advertisements across the web through affiliate programs provide most of their revenue (3). Twitter has been famously unprofitable for most of its existence (201). Meta, meanwhile, has generated billions of dollars of profit through the Facebook, Instagram, and WhatsApp social platforms (144). Regardless, operating costs for platforms are huge: Meta, again, reported $90-95 billion in expenses "primarily driven by investments in data centers, servers, [and] network infrastructure" (145).

Instead, big social media platforms occupy a crucial place in the broader internet system by granting their owners control over information that can be used to make other digital spaces profitable. As Meta CEO Mark Zuckerberg famously told the U.S. Congress, "Senator, we run ads" (News). Like Olson's bandit, which uses the monopoly over taxation and subsequent provision of public goods to dramatically increase the productivity of subjects, platforms convert their monopoly over user information into highly efficient forms of attention: advertisements. The encompassing interest of the tax-collecting autocrat, after all, is not just to extract the maximum possible surplus from subjects but also to grow the possible surplus itself. Providing a minimal level of political order through content moderation creates order and security for users, who can safely turn a greater share of their attention to advertisements without worrying about inefficient, predatory theft of their information. Thus, content moderation is at the root of platform development and the origin of public good provision online.

To operationalize the principle of moderation, I break the concept down into three subcomponents: the agents executing moderation rules, the codification of rules themselves, and the mechanisms available to users to appeal decisions by agents or review rules.

Moderation agents. Is moderation performed by volunteers, platform officials, third-party proxies, algorithms, or some combination of these?

Moderation rules. How specific, extensive, and visible are moderation rules?

Appeal mechanisms. When users are subject to moderation, what is the mechanism for appealing the decision?

How *inclusive* or *exclusionary* are moderation agents, rules, or appeals mechanisms? An inclusive platform might let users volunteer to engage in content moderation, or allow users direct access to a professionalized moderation enforcement institution. Exclusionary platforms do not; instead, users are subject to arbitrary and opaque moderation by algorithms or distant, third-party moderators. They may also completely lack mechanisms for appeal, or adjudicate appeals using the same automated systems. Finally, vague moderation rules might be especially exclusionary, as it creates discretionary space for enforcement.

2.1.2 Autonomy

What sort of autonomy do communities enjoy? Can users self-sort into group identities, or are they imposed on users by the platform itself? Finally, are communities given or able to develop software tools for community administration? The principle of *autonomy* for online communities reflects three aspects of platform governance: the degree of centralization, the pattern of political exclusion, and the means by which platform impose administrative order on complex and disordered processes of identity and community formation. All three derive from the extensive political science literature on distributive politics.

If politics is the study of “who gets what, when, how” (31), then the system of distribution of attention—the chief public good of social media platforms—is central to understanding platform politics. Distributive institutions, processes, and policies typically involve diffuse costs through taxes and concentrated benefits through transfers (209). However, social media platforms present a special case. “Taxation” of attention is often opaque, executed through the use of dark design patterns that rely on users’ subjective emotions and psychology to derive greater usage of the various platform user interfaces (30). “Transfers,” meanwhile, may occur through black-box algorithmic distribution systems that reward communities or even individual users with attention. However, transfers may also be simply overt partnerships with preferred users and communities, such as advertisers. Constituencies are rigorously micro-identified using elaborate data collection schemes (208). The “localities” that receive benefits from attention transfers may not reflect actual community formation, but rather the view of platform data aggregation systems—or even the deliberate imposition of politically advantageous community boundaries on users.

How centralized, exclusive, and artificial these distribution systems are has major consequences for the authority and legitimacy of platform regimes. Literature on distributive politics theorizes that accountability is a key outcome of distribution systems, suggesting that elected officials allocate specific public goods to specific constituencies during election cycles (58). Similarly, selectorate theory argues that autocratic regimes seek to build legitimacy and support among a smaller winning coalition of elites (46). The constituencies that platforms prioritize point towards those communities or users it considers important for maintaining authority and legitimacy. Policy responsiveness is also an outcome of interest for distributional regimes. Both centralized and decentralized systems are vulnerable to political capture by elites and preferred communities, but which elites are able to capture distribution and which communities are targeted for transfer differs (11; 56). Platforms with highly centralized distribution systems may be poorly monitored, degrading legitimacy, but may be able to provide more intra-community equity in attention. In contrast, platforms with more decentralized distribution systems may offer more transparent and accountable delivery, but are vulnerable to inter-community capture by elites.

To operationalize the principle of autonomy, I break the concept down into three subcomponents: inter-community control over attention, intra-community tools for distribution, and coercion of community boundaries.

Audience control. How much control over attention do communities have?

Community tools. How widespread are tools for community administration? Do platforms allow open-source or third-party solutions, or are tools proprietary?

Community boundaries. Are community boundaries drawn by users, communities, or platforms?

These subcomponents are defined by their degree of *centralization* or *decentralization*. Communities in decentralized platform power structures have greater autonomy to control what information escapes the community to the broader platform, more tools to perform administrative tasks themselves, and more effective membership boundaries. Highly centralized platforms, meanwhile, impose community boundaries on users—often invisibly, forcing users to guess which "area" of the platform they are on at any given moment. They also provide limited or proprietary tools for community administration that are less effective at administrative tasks or shape community administration to the interests of the platform. And centralized platforms can also force community or user virality, or block the audiences available to community members.

2.1.3 Discourse Architecture

What is the dominant medium of communication on the platform? How broad is the connective surface of this media to other users inside or outside the platform? Does the platform include or exclude conversation context when users share or respond to content? The principle of *discourse architecture* tackles the quality and form of attention and information on the platform. Discourse architecture refers to the distinct bundles "technological characteristics" (74) that constrain the use and exchange of information and attention in communities. The architecture of communication on the platform in turn determines how identities emerge, ties form, and community networks evolve.

It is important to remember that most online communities are "imagined": members will never meet or interact, but nevertheless construct similar identities around shared interests and ideologies (5). Mutual production and reproduction of information by audiences and authors, the defining feature of social media, play a key role in the construction of these identities. Identities and their content are inert without interaction (105; 19), and social media provides many different forms of interaction.

The content and normative quality of that interaction can vary widely. Expression can, for example, contain just a few words (e.g., “lol”, an initialism for “laugh out loud” originating on Usenet message boards in the 1980s indicating a user laughing (42)), a small emoticon image (e.g., 😂, a small illustrated glyph of a face crying with laughter proposed by Google and Apple employees in 2009 (168)), an image of someone laughing, audio of laughter, or a video of the user laughing–all conveying a similar meaning. These parallel vernaculars bind community members together and emphasize differences with out-groups (101; 127; 90). As the emergence of a state-wide nationalism depended on the “mass ceremony” of reading printed media (5, 35), online communities create their own ceremonies of identity through frequent discussion and mutual exchange of shibboleths unique to their space. Studies of online cultural fandom in sociology have pioneered exploration of these mass ceremonies, particularly in exploring how the “horizontal creativity” of fan fiction communities establish and promulgate community norms and values (157; 86). Theories of so-called “politicized fandom” show that the productive and interactive aspects of fandom—particularly those that produce strong emotional commitments to community and community norms—are analogous to those present in online political communities (202; 47).

The design of discursive spaces helps determine these parallel vernaculars and shape these mass ceremonies. The interaction between platform interfaces and communities generates both epistemic convergence around ways of interacting unique to the platform and specific to an array of Balkanized communities. The consequences of design choices are myriad and well-studied. For example, the synchronicity of “chat-like” spaces leads to less coherent and more abusive standards of interaction compared to asynchronous “forum-like” spaces (183; 182). Stripping content of conversational and user context can also diminish the quality of interaction (87). Discursive spaces that mix together many topics with little structure or sorting similarly diminish the quality of substantive exchanges (151).

Topic definition. Is content well-structured by topic, or does it appear unsorted in the platform interface?

Context exposure. Does the platform flatten or expand content context?

Synchronicity. Are discursive spaces asynchronous or synchronous?

Subcomponents of discourse architecture are defined by their degree of individualism or communitarianism. On one end of this scale, *individualism* refers to discursive spaces where users have a wide latitude for expression (106; 28; 204). This discourse architecture places few restrictions what and how much can be written; users are atomized and pursue self-expression with little concern for replies or deliberation. Asynchronous communication rules, and users can easily dismiss or ignore context. On the other end, *communitarianism* refers to the degree in which users are encouraged by design to engage in community engagement and foster community identity construction. More communitarian discourse architectures promote bonding between members and exclusion of non-members, building on principles of so-called "enclave deliberation" (188). Communication may be more synchronous and immediate, allowing more deliberate interactions with in-group members.

These three principles and their dimensions lay the building blocks for a typology of platform regime types, giving us the conceptual language to distinguish between platforms on the basis of their internal political functions. It also creates the initial framework for tracking changes to platforms over time across multiple dimensions. While significant work exists that tracks just one component of one principle—legalistic content moderation policies, for example—this typology gives us the tools to describe more nuanced historical trajectories of platform governance.

2.2 Turning Three Principles of Platform Governance Into Data

These subcomponents are arranged along how they contribute to the norms of communication on the platform and within communities. Does the architecture of discourse push discursive spaces towards a more communitarian norm, which emphasizes in-group collaboration and community myth-building, or towards individualism, in which users are atomized and express themselves without the expectation of shared civility or responsiveness? Early community blogs are the an example of a highly communitarian discourse architecture (109, 73-75). These spaces had strict topical structure, high levels of context, and asynchronous exchanges, all features that contribute to maintaining group boundaries and encourage bonding between members. On the other hand, boundary-less spaces with significant context collapse and the expectation of rapid, near-synchronous conversation—such as the Twitter main feed—are highly individualistic.

2.2.1 Case Selection

My initial version of the Varieties of Platform Governance (V-PG) dataset covers five major platforms (Facebook, Twitter, TikTok, YouTube, and Reddit) over their lifetime as social media entities. I use two main sources to build indicators: first, demographic platform data gathered from authoritative sources; second, evaluative indicators generated from primary source data. Demographic indicators include size of the platform in daily active users (DAUs), size of the platform in employees, ownership structure, and reported revenue. Evaluative indicators are coded from primary source data for each subcomponent and aggregated into a variable for each principle of governance. Finally, I construct categorical indicators from these top-level principle indicators.

The world of social media platform cases is richer than first glance. Most platform research has focused on big, politically significant entities like Twitter or Facebook that have relatively accessible data (typically in the form of application program interfaces, or APIs). However, the broader universe consists of legacy protocols and platforms going back to Usenet in 1980 (a distributed form of bulletin board system (BBS) built at Duke University and widely regarded as the first proto-social computing network) and Talkomatic in 1973 (a social chat system built at the University of Illinois that pre-empted peer-to-peer social platforms like Telegram) (64). This universe is also inclusive of a wide spectrum of platforms with narrower ambitions than Facebook or Twitter, such as Nextdoor (local issues) or Pinterest (moodboarding), and distributed protocols, like vBulletin or Mastodon. Future versions of the Varieties dataset should incorporate this wider set of cases.

However, initial case selection focuses on a cross-section of contemporary social media sites with significant DAUs and, crucially, primary source data that is either hosted live by platforms or accessible through archives. Facebook is a massive, multi-modal general platform that allows users to post text, audio, photo, and video content. Users also have access to community building tools, like Pages and Groups, and broadcasting tools, like Live.

Data on Facebook covers the platform from creation in 2006 to 2020. YouTube is similarly massive, but focuses exclusively on video. User communities are built around "channels" that are topical or branded. Data on YouTube ranges from 2006 to 2020. Twitter and TikTok offer similar bite-sized chunks of media in text and video, respectively. Communities and content are algorithmically generated and served. Data on Twitter starts in 2006 and TikTok in 2016; both end in 2020. Reddit, finally, is a multi-modal link aggregation site aggressively focused on community building and maintenance. Users self-select into communities with rigid boundaries. Reddit data starts in 2006 and terminates in 2020.

2.2.2 Data Selection

Visibility into these platforms is notoriously difficult, as they rigorously control access to computational data and organizational information. Of course, this is not dissimilar from attempting to measure the *de facto* functions of autocratic polities, as popular indices of democracy often do (see, for example, Varieties of Democracy, Polity IV, or Freedom House). These projects frequently draw on expert surveys, asking subject-matter experts to respond to questions about democratic performance or quality of governance based on their unique insight into opaque institutions or "on the ground" experiences. Ideally, the V-PG project would conduct similar surveys. Instead, I use observational data collated from platforms themselves. First, I gather financial filings in the form of quarterly statements (Form 10Q) to the United States Securities and Exchange Commission and, in the case of privately owned firms, public investor statements. These documents contain not just information about the financial health of the company, but also rich information about the strategic goals of the platform's administration, plans for new product releases or community updates, and any perceived threats to the platform's integrity. Second, I collect information on product creation and release through the platform's development blogs aimed at open-source developers and potential employees in the software engineering field. Because platforms are competing in the same labor pools of highly skilled programmers, they have an incentive to periodically release accurate information about the technical details of computational systems. I collect these documents through both official sources on platform websites and through the Internet Archive, which catalogs historical snapshots of websites. Finally, I collect privacy, safety, and Terms of Use rules posted by platforms towards users. Often, these documents start as simple statements to protect platforms from misuse, and evolve into complex, legalistic tracts that mirror rulemaking in state bureaucracies. Because these are also changed in-place on platform websites, I collect these documents through the Internet Archive.

These data are collected from 2006 to 2020, although the temporal coverage per platform is necessarily constrained by the platform's own history. I begin collecting data on Twitter and Facebook (via parent company Meta), for example, in 2006, even though Twitter's parent company emerged in 2004 and Facebook was technically founded in August 2005. In the summer of 2006, however, both formed into a recognizable social media entity: Facebook becomes available to anyone with an email address meeting age requirements in September (Facebook 2006), and Twitter use balloons around the August earthquake in Sonoma County, California (Twitter 2006). Others, like Reddit and TikTok owner ByteDance, are private companies that do not report financial information to US financial regulatory agencies. Instead, we have to rely on sporadic and semi-public information released to potential investors. Reddit, for example, has completed twelve rounds of investment (thirteen if their postponed initial public offering is included) since 2013. During each of these rounds, investor letters are sent to venture capitalists, retail and institutional investors, angel investors, and other financiers. These documents are typically available on data aggregation websites like PitchBook for a monthly fee. Another data collection problem arrives in the form of rebranding and reorganization efforts by companies. YouTube, for example, operated independently for a year before purchase by Google in 2006. Information from the first year of independent operation is largely unavailable. A major reorganization of Google into a holding company called Alphabet in which YouTube is a subsidiary separate from Google itself again shuffled access to SEC and other financial data. Similarly, Facebook became a subsidiary of the holding company Meta. In both cases, financial information in the dataset is split between different holding companies. Finally, changes to user-facing platform websites means that product release blogs, privacy and security policy, and Terms of Service (hereafter TOS) are often overwritten. Fortunately, I recover a significant history of these documents through the Internet Archive, which automatically collects public website information using web crawlers. Although I am able to show variation in these data sources over time, the Archive's Wayback Machine that collates website snapshots is necessarily incomplete. I cannot, for example, guarantee that these platforms did not make changes to TOS or other policy documents more rapidly than Archive web crawlers could catalogue web pages, or that early versions of these pages were promptly indexed by the Wayback Machine. The Archive crawlers are also biased towards popular sites: they may include crawlers seeded by Alexa, a traffic ranking system; volunteer crawlers; and partner crawlers from libraries. Thus, more frequent changes to TOS, policy documents, or public-facing blogs are likely correlated with popularity of the platform.

Platform	**Start year**	**SEC 10Q (count)**	**Investor statements**	**Product blogs**	**Privacy & safety**
YouTube	2006	42 (Google) 17 (Alphabet)	N/A N/A	545	15
Facebook	2006	32	N/A	135	8
Twitter	2006	24	N/A	37	5
Reddit	2007	N/A	15	120	7
TikTok	2018	N/A	2	12	2

Table 2.1: Platform data collection, count by document type

2.2.3 Aggregating Methods and Procedures

Although the data conceptualization model above has many antecedents in data projects like Polity, V-Dem, and others, the Varities of Platform Governance departs considerably in operationalization. Where indices of democracy or autocracy are focused on determining the distance of a particular case across time from a particular idealized state—say, Dahl's institutional conception of polyarchy (1971)—the V-PG codebook is more exploratory and less determinitive. The three principles of platform governance offered above capture core functions and operations of platforms along seemingly independent dimensions: moderation from inclusivity to exclusivity, autonomy from decentralization to centralization, and discourse from individualism to communitarianism. Indices measuring core aspects of democracy or autocracy, on the other hand, can be assured of (or at least can theoretically justify) the correlation between and substitutibility of different aspects or principles of democratic or autocratic rule. We simply have more theoretical and empirical reason to think these attributes hang together in predictable ways. We know, for instance, that if elections are fraudulent or most political parties are outlawed then universal suffrage is not, by itself, an indicator of "democracy." Or we know that if suffrage is restricted but those without access to the franchise are able to freely associate or express themselves then regimes with restricted voting rights are not necessarily "autocratic." Rather, these dimensions are mutually reinforcing in a coherent manner that admits to aggregation, substitution, and weighting.

Do aspects of platform governance work the same way? Frankly, we do not know, as the social internet is young and we rarely study platforms from this perspective. Does a more autonomous, communitarian, and deliberative community compensate for an exclusionary moderation system? We do not have the theoretical framework or empirical data to say one way or the other. It is possible that, because social media platforms are mostly concerned with governing information and identity, different combinations of the three principles of platform governance produce unexpected outcomes for unanticipated reasons. Because of this limitation, I only specify an aggregation procedure for each subcomponent. And while these aggregated indices are useful for some limited empirical modeling purposes, the primary data product from the V-PG is series of categories for building a typology of platform regime types.

Indicators for each subcomponent are built by a trained team of undergraduate coders. For each platform-year, a corpus of documents associated with each subcomponent is generated from data collected through scraping EDGAR and the Internet Archive. For example, the moderation principle is composed of three subcomponents: agents, rules, and the appeals mechanism. Any discussion of these topics from 10Q forms, financial disclosures, investor letters, product blogs, TOS, or privacy and security documents is compiled into a corpus for moderation agents, rules, and appeals. Each document in the corpus is evaluated as more or less inclusive of users and communities, or exclusive of those governed by the platform. Inclusion or exclusion for each document in the corpus associated with one of these subcomponents is estimated along a five-point Likert scale by coders (see Figure 2.1).

These ratings are aggregated into an indicator for each subcomponent and platform-year. Because each platform-year corpii contains a different number of documents related

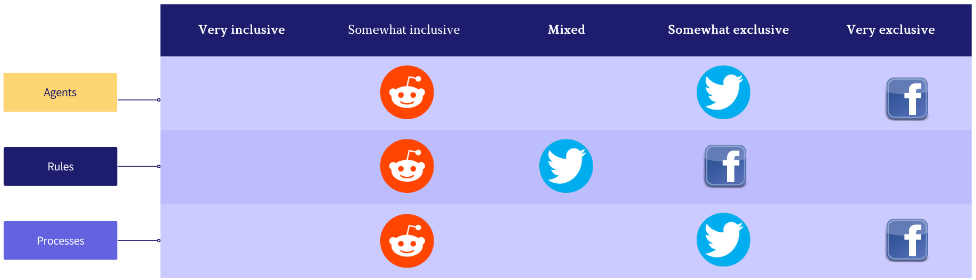


Figure 2.1: Moderation scale, sample

to each subcomponent, the aggregated ratings are weighted by corpus size. The formula for the agent subcomponent looks like this:

$$R_{ma} = \sum_{i=1}^{N_{ma}} ma_i$$

where N_{ma} is the number of document ratings and ma_i is an individual rating. Thus, the full index for moderation is this formula:

$$M = \frac{R_{ma} + R_{mr} + R_{mm}}{N_m}$$

where N_m is the size of the moderation corpus and each R term is the summed ratings of the subcomponents. This same formula is used for all of the subcomponents and all three principles of platform governance.

2.2.4 Platform Regimes

What does the landscape of platform regime type look like? Figure 2.2 below describes modifications to platform architecture over time, as proxied by the number of product release blogs collected in the sample. Of note here are the similar patterns of increasing counts of product releases for most platforms throughout the 2010s for Facebook, Reddit, and Twitter.

This is consistent with the institutional isomorphism evident in the dataset, particularly in discourse architecture. The so-called “pivot to video” starting in 2015 saw all of these platforms try to add short-form video content and tooling, deemphasizing text-based

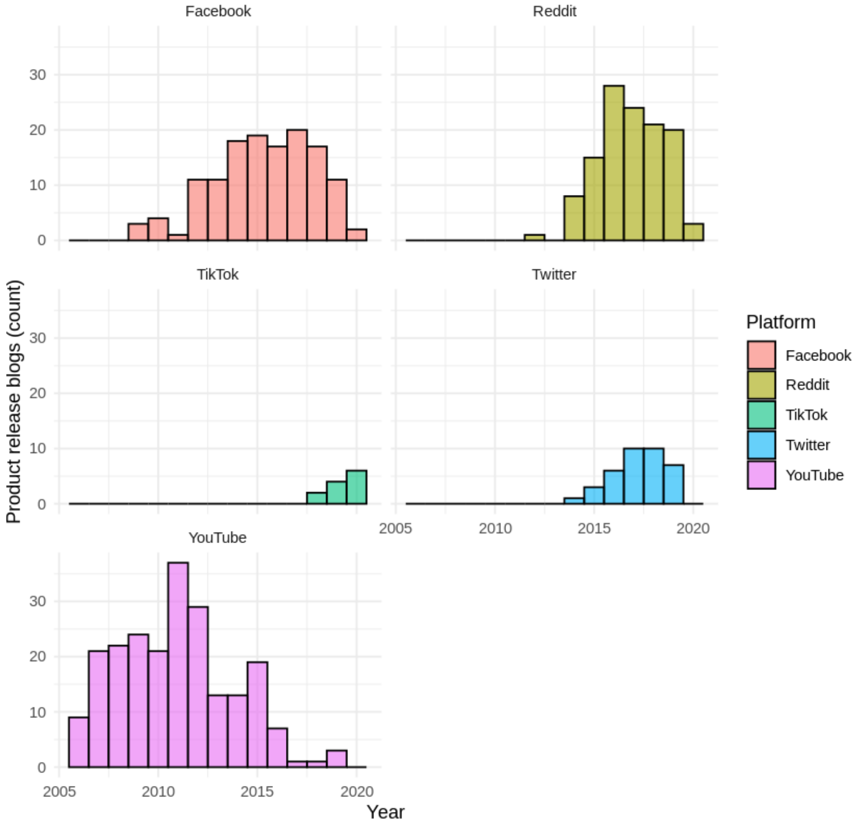


Figure 2.2: Count of product release blog documents collected, by platform

content. A more individualized discursive medium, short video interfaces, recommender systems, and live streaming tools became common place on Twitter, Reddit, and Facebook. A similar but smaller rush to adopt the endless-scroll-video architecture of TikTok happened in 2020, although it is more difficult to detect here, as Meta pushed these changes to Instagram rather than Facebook. Another obvious trend in this Figure is the so-called "feature trim" era of YouTube. Starting in late 2014, the platform steadily began removing discourse features like video annotations, direct messaging, video categorization, and other community-focused tools. The overall effect was a persistent individualizing of the platform's discourse architecture, especially for channels where content creators had previously had significant control over community development.

Another trend across all platforms is the increasing size and specificity of privacy, security, and service rules. Early rules were short and vague: Twitter famously had a 538-word terms of service that was focused on concerns from a pre-social media internet, like spam and "name squatting." Facebook rules were similar. Yet, as the challenges of governance expanded on these growing platforms, explicit rules for content, conduct, and community expanded as well. In Figure 2.3 below, we can see the size of platform privacy, safety, and conduct rules growing for Facebook, Twitter, and Reddit. Part of Facebook's growing ruleset can be traced to data collection by the platform. The controversial February 4, 2009 update that granted Facebook irrevocable license to all user content, in particular, generated expansive changes to platform rules in response to user outcry. Twitter, meanwhile, sees a massive spike in 2014 and 2015 as the platform struggled to contain the Gamergate harassment campaign. Twitter added policies banning hate speech, revenge porn, and violent threats, and rigorously pursued content by dangerous groups like the Islamic State during this time period.

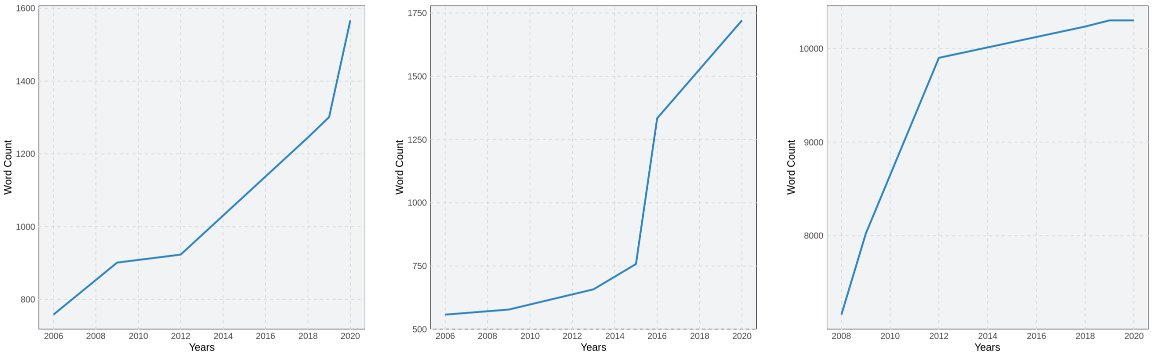


Figure 2.3: Word count by year: Facebook, Twitter, Reddit

[figure: Word count by year, TOS: Facebook, Twitter, Reddit]

For contemporary versions of the platforms in 2020, we see a few distinct categories. Facebook pioneered the use of algorithmic feeds—in fact, the now-ubiquitous term "news feed" is generalized from the News Feed product introduced in 2006. And while this main feed that users interact with is highly individualistic, the ability of users to create and join Groups or Pages provides an alternative discourse architecture that is more communitarian. Groups are also self-moderated by users, in contrast to the highly exclusionary moderation scheme applied to the Timeline. Here, algorithms triage most of the banned content and overflow is handled by an army of third-party workers. Users do not have access to algorithmic criteria or decision-makers themselves, and appeals are also handled algorithmically. Facebook occupies a unique category here, with highly exclusionary moderation, moderately centralized community autonomy, and below average individualizing discourse architecture.

Reddit, meanwhile, is steadily moving towards a more individualistic discourse architecture, adding in atomizing features from other platforms like scrolling videos, text and video statuses, and more. The unpopular 2018 redesign of the site deemphasized text-based discussion, disrupting deliberative context within and across thread. However, "subreddit" communities remain the foundation of the platform, and the interface is still the most asynchronous and text-dense of the included cases. An early regime choice to open-source the site's code, allow users to customize subreddits, and give volunteer moderators most of the enforcement power on the site makes Reddit more decentralized and communitarian than most major platforms. Moderation can still be somewhat exclusionary, as small groups of users can monopolize moderation power across many major subreddits. However, user mobilization against moderators—and the ability to simply create a different subreddit within the site—gives users a degree of inclusivity in the moderation and enforcement unheard of on the other studied platforms. If trends each subcomponent continue on the site (for example, the forced removal of community moderators protesting changes to API access rules), it will soon occupy a similar category as Facebook.

TikTok, the newest platform, is blisteringly individualistic, with few community management tools, communitarian discursive interfaces, or community boundaries. What makes the platform attractive to the newest cohort of social media users—the ability to reach large audiences without building a local social graph—is the result of an enormous connective surface between user content. Users are connected through audio "sounds," visual trends or templates, "remixes," comments, automated comment search and discovery, and more. All these pieces of connective tissue run roughshod over developing community boundaries, which essentially do not exist as a matter of designed digital space. Instead, they are entirely replaced by algorithmically-generated micro-identities to a degree unobserved in other cases. The success of this platform at the far end of each subcomponent spectrum has undoubtedly generated some of the movement we see from other platforms towards these extremes.

Twitter is slightly less individualistic and centralized than TikTok, although actual real-world usage of the communitarian Community, Circles, and Spaces features may be quite limited. Like Facebook it occupies a unique category, but the inverse of the Meta product: somewhat less exclusionary moderation in the 2020 iteration of the platform, similarly centralized autonomy of communities, but much more individualized discursive spaces. It shares a commitment to individualized, algorithmically-generated micro-identities with TikTok. Both platforms offer similar governance structures conducive to content moving rapidly across communities and rewarding community competition for public goods.

YouTube, finally, is a somewhat special case. During the data collection process, coders noticed that feature releases, community tools, and moderation rules were bifurcated: most focused on video creators, who run channel communities, and commenters, who interact below posted videos. The discourse architecture for these two spaces were strikingly different. For video creators, discourse here is relatively communitarian, with video posted asynchronously, occupying strict topical categories, and offering broad context. The comment sections, however, are not and bear a closer resemblance to Twitter's automatically generated feeds around curated trending topics: disconnected, flattened context and highly synchronous.

2.3 Discussion

Although this platform governance project is in the early stages, these data support a promising research agenda. These findings show significant variation between platforms on key dimensions. Not all platforms are built the same, and even those that occupy similar spaces on the three principle measures of platform governance differ on the composition of subcomponents.

More research is needed, specifically in two areas: first, meso- and micro-theoretical work on how different aspects of content moderation, community autonomy, and discourse architectures interact, as well as empirical testing of these interactions. This research is in its infancy, and platforms change faster than state regimes. However, archival and historical work is urgent, as access to this data—and the data itself—is disappearing quickly. If we want to contend with the dramatic rise in online extremism and political disinformation in the past decade, we need to understand how decisions about platform design and choices about how information moves through these multi-platform systems affect the structure and emergent properties of communities. Political extremism varies widely across communities and platforms; what explains why some communities tip over into violent radicalization, while others do not despite similar user composition and material political incentives? Expanding our view of private governance in digital spaces can help answer these tpes of questions.

Second, future versions of this dataset should significantly broaden the case selection. While this initial iteration of the project focused on big platforms central to the social media system, there are many more platforms on the periphery who may occupy different spaces in the regime typology or play outsized roles by virtue of their position in the structure. Focusing only a few major social media platforms limits the reach of the dataset. Big, general platforms governed by public companies may be the most visible platforms, but other, smaller spaces occupy significant positions within the internet media system. Legacy systems like bulletin boards or forums moderated by volunteers are still widely popular with heavy internet users. What role do these smaller actors play in the interconnected platform system? How does variation in governance and moderation help explain the emergence of violent, extremist sites like Stormfront, the infamous online white supremacist community? Similarly, new, emergent forms of internet communication also require our attention. In just a couple of years since the first Varieties of Platform Governance data was collected in 2020, changes to internet governance—mostly to Twitter, now “X”—and platform action on extremism and disinformation have further fractured the online media system. The so-called “alt-tech” space, where politically motivated actors seek out preferred moderation, community, and discourse policies, is ascendant. And many users have gravitated away from public sites like Facebook or Twitter to semi-private enclaves like Discord, where pre-existing social graphs are reproduced in tightly integrated networks. How communities handle this “splinternet” and decide where to move

CHAPTER 3

Community Sorting Across Platforms

How political communities move between social media platforms and other online spaces is an open question. In the last five years, the once consolidated social media landscape has begun to fracture; while Meta platforms like Facebook, Instagram, and WhatsApp still command more than two billion users each, niche platforms enjoying out-sized political and social influence have proliferated. Many of these spaces have built explicit ideological goals into content amplification, moderation, and management architecture. Right-wing extremists on Gab, for example, wrote antisemitic and racist content preferences into the site's mission statement, while federated instances on the decentralized platform Mastodon deploy zero-tolerance policies for transphobic or anti-Black hate. General platforms like Twitter or TikTok also have policies that ostensibly diminish or ban hateful content, but they are deliberately engineered to give content and moderation teams room to maneuver around politically fraught decisions (114).

How do online communities navigate these changing spaces? Sometimes we observe communities leaving platforms that do not adequately protect them from harassment; many Jewish, Black, and transgender communities left Twitter after the platform relaxed policies against hateful content in 2022 and 2023, for example. Similarly, we see the extremist movements that promulgate this violent content moving to alternative platforms that protect and encourage hate when threatened with widespread "deplatforming" by mainstream platforms (48). However, they do not always stay there. The same hate-friendly policies adopted by Twitter in 2022 attracted some of these deplatformed extremists back to the platform, and extremist watchers frequently observe "off-site" organizing to return to mainstream platforms for messaging, recruitment, or harassment (125; 75).

Migrations of political communities and the movements they embody across online spaces has enormous consequences for governments both democratic and autocratic. Widespread smartphone use and access to platforms like Twitter and Facebook are widely believed to have made the so-called Arab Spring of the early 2010s possible (200). The effectiveness of media usage by citizens and elites alike depends on *where* audiences are located. Leaders can shift focus from unpopular policies through social media use (10), surveille and impede political opposition (85; 154), or respond to democratic constituencies (9). Platforms are also effective spaces for mass politics, leveraging users' social graph to mobilize opinion (163), motivate protest participation (124), and give voice to disenfranchised groups (25; 155). All of these effects are mediated by where political communities interact and engage.

Do communities sort themselves onto social media platforms with favorable content moderation and attention distribution mechanisms, or do they pursue engagement with out-group communities regardless of hostile platform policies? In other words, do these political communities engage in "forum shopping"? And what are the consequences for the social media research agenda in political science?

To solve this puzzle and explore the strategic interaction between platforms and communities, I turn to a simple classic model of "voting with your feet." The Tiebout model of citizen sorting argues that local governments efficiently allocate public goods if voters can move freely between municipalities, effectively "voting with their feet" on preferred policies (195). The pure Tiebout hypothesis relies on a number of assumptions about voters and municipal governments. "Consumer-voters" have complete information and full mobility to move between municipalities. They have many options and moving is costless, both in terms of friction and obtaining housing or employment in a new home. Municipalities are constrained by the optimal community size necessary to provide services and will seek to attract or shed voters to settle on this optimum. With just these base assumptions, Tiebout argues that local governments can meet voters' preferences for public projects efficiently. The theoretical literature on public good distribution has extensively tested and refined the "pure" model. (22), for example, shows that the original hypothesis that local governments can find efficient distribution equilibria only holds when there are at least as many voters as community options. Relaxing some of the baseline assumptions can overcome this limitation: by assuming there are fractional costs to moving (213), particularizing preferences (116), or changing the nature of public goods (20; 166). The pure model and subsequent refinements focus on a political environment in which public goods are concrete, material things and municipal institutions are strictly majority-rule.

While the original model and Tiebout's central hypothesis were too simple to capture the dynamics of citizen sorting—perhaps exemplified best by the success of recent work on partisan geographic sorting (121; 138)—the Tiebout model fits another case better: movement between internet spaces. Applying the model to social media platforms as distributors of a public good creates an opportunity to plausibly retain some of the original model assumptions while varying others in unexplored ways. In this paper, I argue that social media platforms are quasi-governments offering actors—conceptualized as distinct online communities—a bundle of public goods in the form of attention to and engagement with user-generated content (UGC). Communities vary in their preferences for what kind of UGC platforms prioritize and for UGC reaching a particular audience as a function of their ideological commitments.

This strategic interaction drives community sorting—and community radicalization. Using computational simulations of agents sorting themselves among platform spaces, I show that some starting combinations of platform system structures, communities, and platform preference aggregation mechanisms are conducive to pushing communities towards more extreme viewpoints. As communities compete under these systems for public goods, they face incentives to radicalize, obtaining greater and greater utility from ideological commitments to identity-based political violence against competing communities. In particular, an online informational environment with few available platform options, vague or exclusionary content moderation policies, and limited community autonomy drives political communities to radicalize.

3.1 Simulation Experiments

The primary focus of this projects is on three questions. First, do communities efficiently sort themselves onto favorable platforms if allowed to freely move? Second, do different types of platform governance and structures of inter-platform interaction lead to more or less favorable outcomes for online communities? And, finally, do extremists sort themselves onto platforms with favorable bundles of public goods, or do they stay on platforms with "mainstream" communities?

To answer these questions, I construct three simulation experiments using an agent-based model (ABM). While the rich literature investigating the Tiebout hypothesis and other sorting questions are primarily focused on investigating fixed equilibria, simulation models offer an opportunity to study the *dynamics* of complex systems using researcher-controlled inputs. In particular, ABMs are powerful simulation tools for exploring problems centered on autonomous agents—often, but not always, individuals—interacting with each other and a socio-political environment (118). ABMs can be powerful exploratory tools for emergent behavior (66), hypothesis testing (21), or theoretical mechanisms (6). This project is chiefly concerned with the dynamics of community movements across platform spaces and the composition of platform-community clusters in the settled state of the simulation.

To put a platform sorting model into action, I build an ABM simulation environment using *python* and the *AgentPy* open-source modeling library (70). The simulation environment (hereafter "sim") starts with the minimal assumptions necessary to produce community sorting between platforms, and I make two experimental modifications to observe how additional complexity changes the state of the sim. In this section, I describe each of these sim variations and my expectations for them. The unmodified sim is the *base model,* which contains just enough minimal viable machinery to produce community sorting between platforms. In this model, several assumptions hold: an arbitrary number of communities are assigned to an arbitrary number of platforms, each with single-dimensions linearly separable preferences and policies, respectively. A Tiebout cycle, shown below in Figure 3.1, begins with communities moving (if applicable), computing their utility, searching for potential destination platforms, and selecting a strategy of "move" or "stay" for the next step in the sim. The sim concludes when no communities can improve their utility by moving and all have a selected strategy of "stay." I expand on the details of this model below I expect that, in line with the results of typical game theoretic models of the Tiebout hypothesis, this sim will efficiently distribute public goods across community agents and, consistent with other Tiebout simulation experiments (see, for example, (115)), conclude in less than ten steps.

The first modification adds preference aggregation institutions to platforms. This *institutional model* introduces three mechanisms for changing platform policies: direct ranking, coalition formation, and algorithmic recommendation. I describe these mechanisms in detail below. While these features resemble voting systems common to social choice theory,

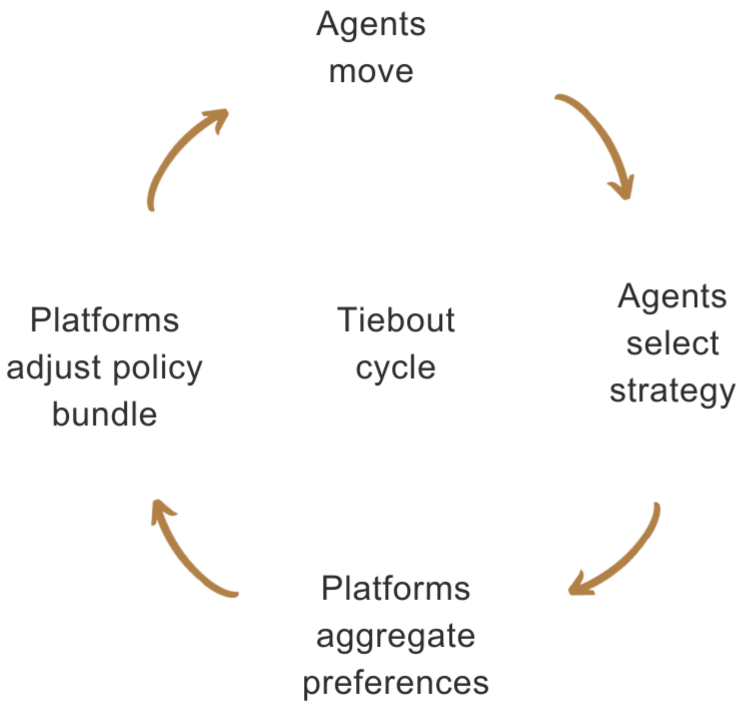

Agents move
Agents select strategy
Platforms aggregate preferences
Platforms adjust policy bundle
Tiebout cycle

Figure 3.1: Tiebout algorithmic cycle

it is useful to think of them both as decision mechanisms for changing policy as well as distribution mechanisms for distributing public goods directly. This model preserves the assumptions of the base model. The Tiebout cycle in this version allows communities to change platform policies before computing their utility and selecting a new strategy. Importantly, communities do not have visibility into the strategic interaction of preference aggregation on other platforms when searching for potential destinations; that is, they do not consider if they would be the “deciding voter” on a new platform before moving. Again, the sim concludes when no communities can improve their utility by moving. In this model, the sim can be configured to compare three different institutions, described below, or a mix of these institutions in the same simulation iteration.

Finally, the *extremism model* attempts to integrate different community types into the subgame model. While community sorting between different platform regimes is interesting by itself, what we really want to know is how extremist communities behave. Do they pursue sympathetic content distribution? Or do they look for "mainstream" communities to antagonize? This modification introduces community types that receive a portion of any utility taken from fellow tenant communities when platform policies change. In other words, "extremist" communities view the pool of utility available on a platform as zero-sum, while "mainstream" communities do not. Extremists can thus extract utility on a platform by making the distribution of public goods to every other community worse. In the sections below, I elaborate on the utility functions of political extremist communities and how they change the dynamics of the final simulation experiment.

3.1.1 Base Sorting Model

The base model is a modified version of the Tiebout sorting model with similar assumptions. First, assume that there is a set of N_c communities. Communities, $\boldsymbol{C}$, are the primary actors seeking favorable attention distribution systems from a collection of available social media platforms. Each community must join one platform from a set N_p. Platforms, $\boldsymbol{P}$, offer a fixed set of public goods given by a set of N_a.

Rather than representing a municipal expenditure on a public works project, as in the original Tiebout model, public goods here are a bundle of platform moderation rules, community autonomy, and discourse architecture design choices. Discourse architecture refers to anything in the designed function of the website, mobile application, or other user interface that influences how users interact. This could include media type (text, image, video, and so on), synchronicity of communication, or context collapse. Community autonomy refers to how users are able to sort themselves and build collective identity on the platform. Platforms may provide built-in tools for this, like community management or moderation processes for users; they might not, leaving users to self-govern and police their own community boundaries. Finally, content moderation rules are the legalistic rules that direct platform processes. These could be vague, allowing platforms a wide latitude of action, or highly specific, banning user activites.

In the base model, I specify this bundle as binary variables to amplify or suppress user generated content (UGC); let $b_{pa} \in 0,1$ represent the policy bit a of platform p. Platform policy bundles are arbitrary in size. The full bundle is given as $B_p \in 0,1^{N_a}$. Communities have linearly separable preferences across these bundles such that $B_c \in 0,1^{N_a}$, and obtain a unit of utility for each bit a on platform p that matches these preferences. Community preferences In the base model, assume that the utility function for a community, u_c is the sum of the matches between platform policy bits and community preference bits, where δ is a Kronecker delta function:

$$u_c(B_p, B_c) = \sum_{a=1}^{N_a} \delta(b_{pa}, b_{ca})$$

Each platform aggregates preferences for determining the final program of policies from among the communities currently on the platform. In the base model, this is a simple direct majority vote on each b_{pa} by every community N_c on platform N_p. After platforms determine the program of distribution, communities react by staying on the platform or migrating to other available platforms. The process of communities determining their utility, voting for a new program of policies, and then choosing a strategy of “remain” or “move” constitutes a single Tiebout cycle. The model reaches equilibrium when no communities can better their utility by moving to a new platform in the next cycle. Importantly, I constrain communities from computing the likelihood of their vote on a new platform overturning the current public good bundle; in other words, communities do not have knowledge of nor do they try to predict where other communities may move in the next Tiebout cycle.

3.1.2 Community Assumptions

For the base model, I rely on key set of assumptions about community actors. First, I assume that communities are unitary actors with linearly separable preferences on each policy. Significant work on networked groups shows that their ability to cooperate or engage in collective action is contingent on network structure. In a similar simulation example, (60) show that a competitive network—where armed groups are all at similar risk of attack—can drive civilian victimization. The key intuition here is that, although low-level incentives can push individual group members into violence against civilians, network effects by themselves can result in group-level decisions to victimize civilians. Moreover, variation in ties between group members can affect both preference aggregation as a function of information transmission and collective action, and not always in the same direction at the same time. For example, weak ties and fragmented networks can impede information transmission (123). But weak ties are not necessarily an impediment to collective action when peripheral members boost calls to action (178; 124) or are motivated by small clusters of strongly tied members (35).

Membership in online communities presents a similar problem. Typically, online communities do not have formal membership mechanisms; instead, they are complicated attachments subject to variation in roles (185), groups, and personal relationships. Users have multiple identities, and the salience of their community identities may vary according to situations and interactions (184; 173). Communities may grow or shrink in response to external events that provoke greater salience. Community memberships also vary in centrality (185) or prominence (179) as users actively construct and reconstruct identity hierarchies. While this assumption necessarily loses some precision, the focus of the project is the structure of interaction at the platform-community level.

In the same way, I assume that every member of a community obtains utility from attention and interactivity on social media. While there is room for variation in community type—as I will explore in modifications to the base model below—this assumption means that a standard utility function fits all communities and community preferences. In this model, the public good offered is the supply of audience attention to communities producing UGC. Communities gain utility from information reaching the intended audience and from that information *not* reaching audiences they would rather avoid. Similarly, platforms also offer the public good of deflecting attention—communities gain utility when they are not exposed to information that they would otherwise prefer not to see. Preferences and policies, as sets of binary variables, can represent both of these orientations towards information. Substantive interpretation of this mechanism can vary based on the dimensionality of the preference set. For example, a small policy bundle—say, three binary variables—could represent moderation rules that proscribe or allow pornographic, violent, and automated content. Higher dimensional variables, meanwhile, could be interpreted as capturing a greater extent of the discourse architecture and community autonomy decisions of platforms—such as structural choices to push users towards video rather than text content, providing moderation tools to volunteer moderators, or concealing the function of an algorithmic recommendation system.

Third, I assume that communities do not incur costs when moving between platforms. This reflects a similar assumption in the original Tiebout model. Moving between digital platforms is less costly than physical movement between municipalities. Physical relocation has highly heterogeneous effects, from major improvements in subjective well-being and lifetime earnings when low-income citizens move to more prosperous neighborhoods (133) to substantial wage losses when workers are involuntarily displaced (119). However, platform migration is not empirically *costless.* "Fandom" communities—users connected by shared enjoyment of cultural products—that move between platforms experience significant disruption, including loss of content and social fragmentation (68). Online communities may not "complete" migrations when displaced, leading to smaller, less stable communities spread across multiple platforms (33). Other communities—particularly those with violent extremist ideological commitments (146) or those organizing against state security forces—may deliberately spread themselves across platforms to build network redundancy. While some communities may experience significant disruption from movement between platforms, I argue that major political identities of interest are more likely to survive migration.

Finally, I assume that communities are given and exogenous to the model. Of course, we know empirically that communities may be produced by the platform itself. Strong parasocial relationships between influencers and users can generate communities built around this parasociality (99; 181; 214). Design choices intended for a an intended use may produce communities built around "hacking" these platform functions, like users building collaborative stories on TikTok's "remix" mechanism (50) or members of early forums building the image macros that blossomed into modern meme culture (62; 18). Users may even become members of identity-based communities constructed around use of the platform itself (95). A model that incorporate evolutionary combination, recombination, and dissolution of communities would be self-evidently valuable; however, the focus of this model is how existing communities move through the interconnected platform system.

3.1.3 Platform Assumptions

In this model, I assume that platforms are passive actors. Here, platforms simply provide a collection of public goods and a process for preference aggregation; they are not sensitive to tenant community preferences or the pool of off-platform communities. Platforms make significant institutional and procedural changes, often quite rapidly and in response to user behavior. They may also deliberately change to attract or repel certain communities. For example, Gab, a micro-blogging platform billed as an alternative to Twitter, explicitly sought to attract right-wing extremist communities by building antisemetic and racist policies into the platform ecosystem (104; 49). Following Elon Musk's takeover of Twitter and subsequent gutting of safety and privacy policies, contentious and violent communities enjoyed substantial increases in attention and engagement. This assumption does weaken of the model. It relocates responsibility for political extremism and other bad behavior to communities, rather than platforms—actors who undoubtedly possess more power than any one community or user. It also reifies the idea of the 'platform' as a 'discursive resting point' (15), a passive collection of infrastructure that is simultaneously crucial and fragile (82). This rhetorical strategy is often used by platform companies to justify friendly legislation, loose regulatory frameworks, and liability shields. It also provides political cover for platforms to ignore the preferences of underrepresented communities in favor of policies and design choices that benefit preferred communities. However, the focus of this model is on how communities behave and react to platforms; thus, variation between preference aggregation mechanisms is the only way to distinguish platform types.

I also assume that platforms are unitary actors with perfectly functioning processes for enforcing content moderation decisions, maintaining discourse architecture, and stabilizing community autonomy. This is, of course, not the case. Platforms perform these functions through a kind of private bureaucracy, which faces principal-agent problems that mirror those of public sector bureaucracies (39; 26; 159). Implementing platform policy, especially content moderation, is often out-sourced to third-party contractors, which may exacerbate agent shirking (196; 160), corruption, and lax enforcement (27).

In addition to the more general bureaucratic literature on obstacles to service provision, there is a rich literature on the difficulty of content moderation at scale. In a magisterial 2017 review, for example, Klonick identifies a litany of external influences to "legalistic" governance of the public goods provided by platforms (2017). These include media interference, nation-state requests or demands, third-party advocacy, and slippage between levels of content moderation. Klonick writes that "private platforms are increasingly making their own choices around content moderation that give preferential treatment to some users over others," adding in a footnote examples of *ad hoc* rule changes for United States President Donald Trump, North Korean officials, and Indian Prime Minister Narendra Modi (114)[1665]. Platforms also face a unique challenge with agent corruption. Moderators, engineers, executives, and anyone else with access to platform code can wield enormous power over regular users. Instagram, for example, has suffered for years from a black market of third-party scammers working with employees to extort banned users (34). Similarly, off-platform teams accept money for "ban-as-a-service," exploiting algorithmic enforcement mechanisms that disable user accounts receiving too many reports in a short time-frame.

Scholarship on the structure of the international state system often assumes actors are unitary, as does rational choice and game theoretic approaches in other area. The former case is supported theoretically by key assumptions and observations about the function of sovereignty in the Westphalian state system: that there is some final authority within a state that "stops at the waters edge" (97; 174). Despite differences in aggregating citizen interests—through a small selectorate in an autocracy or through mass elections in a democracy, for example—theories explaining the international system identify states as uniquely capable of acting on behalf of societies (117; 120).

Platforms may function similarly. Although theories of platform "sovereignty" stand on shakier (and more recent) theoretical ground than those of Westphalian statehood, platform authority within its "boundaries" is nonetheless absolute. TikTok moderation policies do not extend to Facebook users of content, for example. Platform claims to authority and sovereignty might be functionally stronger than some states, in fact, given that power over code does not require the manipulation of complex social systems; command over an algorithm is likely to be greater than command over a bureaucracy. In the latter case, I am also interested in a parsimonious model of inter-platform dynamics and assume unitary platform actors to that end. Early theories of realist international politics—and the rational choice approaches that inherited from them—argued that assuming states are unitary actors is a modeling choice that trades abstraction for explanatory power at a particular level of analysis (see, for example, (207). Structural models of internet platform systems have an even greater claim to this trade-off as we have less observational and theoretical knowledge about the inner workings of platform companies and content moderation bureaucracies than international relations scholars do about states and foreign policy establishments.

3.1.4 Experimental Parameters

The simulation environment consists of series of researcher-specified parameters for agents and a sequence of Tiebout cycles. I include the number of communities, number of platforms, and dimensionality of preferences and policies as structural parameters. The make-up of the community and platform cohorts is also specified: the proportion of "mainstream" and "extremist" communities and the composition of platform preference aggregation institutions, respectively. For the base model without institutional or extremism modifications, I set both of these parameters to null. I am broadly interested in testing whether community movement between platforms can produce an efficient allocation of public utility, how changing compositions of community and platform types affect this allocation, and how the dynamics of community movement between sim states change with parameter variation.

To understand these outcomes, the sim generates measures of sim-wide, per-platform type, and per-community type average utility; number and history of community movements as both single measures and bipartite graphs; and measures of sim stability, including the number of steps before completion, the average utility improvement per step, and number of policy changes per platform. Computational sim experiments give us great control over not just parameters, but the ability to conduct many trials. While the parameter space is quite large, *AgentPy* and *python* ensure the computational cost of re-running a specific sim specification is very low. To this end, I iterate across each parameter set for 100 trials, producing aggregate outcomes for each measure captured by the sim[1].

The first simulation experiment compares the efficiency—average utility per community—and stability—average number of community moves—between the base model with a single platform, the institutional model with a single platform, and the institutional model with multiple platforms. Across these three configurations, the number of communities is standardized at $N_c = 100$, the dimension of public good bundles at $N_a = 10$, and the number of sim steps at $t = 50$. The second compares the efficiency and stability of sim configurations with a mixed set of institutions across inter-platform structures with few and many platforms. This experiment has three configurations, with $N_p \in [3, 9, 27]$, the number of communities standardized as a function of the number of platforms with $N_c \in \{100, 300, 900\}$, and $N_a = 10$.

Finally, the third experiment examines the dynamics of extremist and non-extremist communities. It compares the efficiency and stability of difference sim configurations and also looks at the underlying states of the model during simulation. First, this experiment compares the performance of each type of platform institution with extremists comprising five, ten, and fifteen percent of all communities. These sims each have a standardized configuration of $N_c = 100$, $N_p = 5$, and $N_a = 10$. Second, I use network graphs reconstructed from simulation states to describe how extremists move through the platform, examining whether they settle on platforms with favorable goods bundles comprised of other extremist communities, stay on platforms with unfavorable bundles to consume utility from mainstream communities, or continuously move through the platform system. Each experimental configuration is run using the *AgentPy* Exploratory Modeling and Analysis (EMA) Workbench, a suite of tools for reasoning through complex systems with significant uncertainty (8). The EMA Workbench generates 100 iterations for each experiment using the same random seed. All outcomes are averages across these iterations.

[1] The parameter space is technically infinite, as N_c and N_p can be any positive real number and the tensor dimensions of preferences or policies is limited only by the availability of system memory. To limit the sim to empirically reasonable parameters, I set the range of communities to $[100 - 1000]$, platforms to $[3 - 100]$, and preference dimensions to $[10 - 250]$. These parameters ensure that experimental iterations will conclude in hours rather than days or weeks, and can be replicated on sufficiently parallelized consumer personal computer hardware. Future work can explore how the simulation behaves with arbitrarily large parameters.

3.1.5 Comparing Institutional Models

Direct voting: platforms surface or hide content based on community input on each bit of UGC. After communities have relocated at the start of a new Tiebout cycle, platforms aggregate preferences by allowing communities to "vote" on each element of the public good bundle. Bundles that receive a majority vote remain, while those that fail to reach this threshold are inverted. Tied votes are inverted randomly. Table 3.1 below shows that this form of preference aggregation is the most stable and delivers the highest per-capita utility across all platform types for sim configurations with a single platform. Agents can earn a maxium of 10 utility, and direct voting deliver slightly more than half that on average. The possibility of any one agent's vote changing the platform *status quo* is low; longer tenure on the platform without the ability to move elsewhere means average utility among tenant communities goes up. As Table 3.2 shows, this advantage diminishes when there are multiple direct platforms. Although direct voting delivers similar average utility to communities, agents still move an average of 23.6 times before reaching an equilibrium.

Institution	**Average utility**
Direct	5.07
Coalition	4.77
Algorithmic	4.16

Table 3.1: Single platform, $N_c = 100$

Coalition formation: while "direct voting" does resemble some content visibility schemes—such as Reddit's "up" and "down" voting system—these are almost always aimed at aggregating the preferences of *users,* rather than communities. We know from social choice theory that aggregated group preferences may not reflect the best preference ordering of citizens. Thus, choosing to model communities as a democratic referendum-style institution is an awkward fit. Instead, it may be useful to consider more movement-focused aggregation mechanisms. For example, Kollman et. al. build a Tiebout model that examines how "adaptive" parties might form through a variety of search algorithms (1997). In their computational model, "platforms"—in this case, a party platform—form *ex post* in municipalities by surveying the preferences of agents from the previous Tiebout cycle. Current agents can then vote on the public expenditure promised by the parties.

Platforms do not have parties or, aside from pseudo-voting systems like Reddit or similar "rating" mechanisms, actual democratic processes in which they can participate. However, we do observe some coordination efforts among different types of communities to influence platform policies, enforcement, or architecture. For example, campaigns by trans communities with support from Korean pop music fans, lawyers, and human rights advocates led to Twitter banning "deadnaming" as a form of harassment; similarly, a campaign of hate and harassment by many varieties of transphobic extremists led to Twitter quietly repealing this same prohibition in 2022. Movements in online spaces can form and vanish quickly, and communities can move—or split—between them just as quickly.

The "movement formation" mechanism tries to capture this dynamic. In this version of the preference aggregation institution, movements can propose a bundle of changes to the public good set to communities, and slates that receive a plurality vote replace the current platform *status quo*. Crucially, movements are not randomly assigned, but respond to the composition of community tenants on a platform. In this version of the Tiebout cycle, a number of movements on each platform—specified by the researcher at the initial setup of the sim environment—have the chance to survey communities and adapt their public good bundle to maximize vote total. After this adaption process is complete, communities have an opportunity to vote on bundles before performing their utility calculations.

How do political movements adapt to the preferences of the communities they represent? Much of the work on change and adaptation in political movements focuses on organizational characteristics, tactics, and change (see, for example, (140; 190; 156)). The theoretical frameworks of this research tradition emphasize the strategic behavior of *groups:* political entities with hierarchy, some degree of formal membership, sanctioning mechanisms, and so forth. In this case, movements adapt and change in response to resource pressures, state repression, and other features of the political environment. However, the movements of interest here are the product of emergent behaviors. In other words, they are the result of decentralized cooperation and the informal governance of many interconnected actors. Actors in these networks are individuals, and their interactions—whether they are by chance (122), shared kinship (59) or ethnicity (123), or friendship (67)—sustain cooperation through sharing information, sanctioning misbehavior, or rewarding pro-social behavior. Repeated interactions create shared norms, define the range of acceptable behaviors or expressions, and proscribe the limits of group boundaries.

In this case, actors are whole communities, and ties between actors vary in type and strength. The interactions between them that produce shared norms, transfer information about preferences and ideological commitments, and affirm boundaries between them are generated by chance encounters. “Movements” that result from these loose ties and varied interactions are likely to be ephemeral even when successful at changing the bundle of public goods offered by the platform. There are many ways to model tie formation between actors that have these properties and no one model is likely to be accurate.

For the adaptation mechanism, I tested a variety of metaheuristic algorithms that try to solve local search problems, incuding hill-climbing, nearest neighbor, and genetic algorithms. I chose a simple implementation of a genetic algorithim using mutation genetic operator in order to capture the intuition of repeated chance encounters between commnuity members. At the start of a Tiebout cycle, communities randomly generate g number of coalition public good bundles; g is a dynamic parameter in the sim that ranges from $2-10$ across simulation iterations. Each coalition polls tenant communities to calculate the fitness of the bundle. Then coalitions are allowed to “mutate”: three randomly chosen bits from the bundle of public goods is flipped, and communities are re-polled for fitness. This mutation process is repeated for ten iterations. Finally, communities are allowed to vote on coalition policy bundles, and the bundle with a plurality vote replaces the platform *status quo.*

Coalition platforms deliver slightly less average utility to communities in single platform simulations, as Table 3.2 shows. They also are less stable, as shifts in the *status quo* are more dramatic. However, these platforms perform poorly in multiple platform simulations, providing the lowest average utility and requiring communities to move nearly 40 times. This reflects expectations about the ephemeral nature of online movements described above. Inside of this simulation configuration communities experience large shifts in utility as bundles change in response to newly formed movements. Agents move, seeking out new coalitions that might deliver better public goods or looking for more stable platforms where significant shifts are less likely. But each agent relocation to a new platform changes the coalition formation inside platforms, making it difficult for coalition platforms to efficiently deliver utility to communities.

Institution	**Average utility**	**Average moves**
Direct	5.68	23.6
Coalition	4.74	39.5
Algorithmic	6.75	1.88

Table 3.2: Multiple platforms, $N_c = 100$

Algorithmic recommendation: while coalition formation and, to a lesser extent, direct voting capture the attention distribution bundles of early social media platforms, the contemporary landscape looks much different. The popularity of TikTok and its recommendation algorithm generated a cascade of platform isomorphism. Now, incumbent platforms like Facebook, Twitter, YouTube, and others all have discursive spaces with short video and aggressive recommendation algorithms that break users down into smaller and smaller micro-communities. On these systems, group-specific recommendations obscure traditional community building; unlike coalition formation, preference aggregation algorithms sort users and existing communities into categories largely hidden from the user. Users may not know how attention is being allocated or understand how to express themselves within the range of acceptable behaviors for this ascribed shadow community. This can have deleterious effects on the utility of users individually and existing communities at large.

There are many recommendation algorithms and systems in theory and practice. The most popular one may be collaborative filtering, in which attention is allocated to content that has been rated by many members of a community (either acquired through user choice or ascribed by the algorithm itself). It largely functions by comparing past user behavior within some grouping or clustering parameter, using these repeated interactions to shift platform attention to new content that closely resembles previously popular content. While TikTok and other big platforms use computational collaborative filtering, automation and computation is not necessary; for example, Wikipedia's distributed editorial and contribution scheme is also a form of collaborative filtering. Thus, there are many ways to perform this kind of filtering using human, computer, or even naturally occurring algorithms. The most famous computational implementation is almost certainly matrix factorization, specifically single value decomposition, a set of methods popularized by Simon Funk's submission to the Netflix Prize[2].

SVD recommendation works by decomposing high dimension user preferences into latent factors that are then used to recommend new content. Crucially, they depend on users engaging with and “rating”—either deliberately through a rating mechanisms or approximately through various attention-capture methods—many pieces of content. Without this repeated activity, SVD algorithms face the so-called “cold start” problem, where there is not enough rating data to produce recommendations. One way to overcome this problem is through *group-specific SVD*(23). This variation on SVD groups together users with similar preferences, generating latent factors on a per-group basis. Then, when a new user with too few interactions seeks a recommendation, the algorithm can more effectively push their attention towards content based on their assigned group category.

Beyond solving the technical “cold start” problem, this meta-heuristic also provides a more intuitive example of how recommendation algorithms build ascribed shadow communities—in this case by literally assigning an invisible category to new users. To approximate this preference aggregation institution in the sim, I build a simple toy implantation of the Bi et. al. group-specific SVD algorithm (2016). At the start of the Tiebout cycle, platforms randomly generate a collection of possible policy bundles and polls communities on their preferences for these bundles. Based on these ratings, the platform constructs k groups of similar communities within r range of utility; these parameters are supplied by the researcher and static for the duration of the sim. From these sets of communities, the algorithm constructs latent factors using SVD and stores these. In the first cycle, tenant communities are offered the most preferred policy set by each group. In future cycles, new communities joining the platform are assigned to a group with similar preference sets and offered the group policy set. After group utilities are calculated, the latent factors for each group are updated with the addition of new communities and a new policy bundle for each group is prepared for the next cycle.

This preference aggregation institution is unique as, unlike direct voting or coalition formation, communities on the same platform may experience very different attention al-

[2] The Netflix Prize was a competition to build a collaborative filtering system that surpassed Cinematch, the Netflix-developed algorithm. It started in 2006, and ended in 2009, when the $1,000,000 prize was awarded to a team named “BellKor’s Pragmatic Chaos.”

location bundles. It is also worth noting that in the simulation the link function that generates "ratings" from policy bundles is known to the researcher. It is simply the utility function of an agent. In practice, SVD and similar algorithms use latent factors to discover or approximate this unknown function. Users may have both widely different preference and different ways of obtaining utility from applying these preferences. SVD works by reducing uncertainty over these differences. In this case, we are actually *adding* uncertainty, perturbing the simple mapping of an agent's utility function onto the policies offered by a platform, in order to describe the imperfect information space platforms exist in.

Algorithmic platforms performed the worst on the single platform simulation. And, contrary to my expectations, algorithmic platforms perform much better than other platform types on multiple platform configuration. In the former, the average utility delivered to communities is just slightly less than *movement* platforms. However, in the latter, results are dramatically better: communities get substantially higher average utility per capita and move less than twice on average. What explains these disparate results? Observational studies of algorithmic news feeds also have mixed outcomes. Some large-scale surveys show that users derive greater enjoyment out of algorithmic architectures, preferring them to curation by editors or friends (194).

However, these same studies show wide variation among surveyed cohorts, with significant differences mediated by age, trust in online content, political preferences, privacy concerns, and so forth. This suggests that different communities might derive different levels of utility from algorithmic platforms for reasons exogenous to network structure, such as trust in platform governance and bureaucracy. In this case, however, it may simply be that multiple algorithmic options gives community agents the ability to select recommendation systems that work best for them. Inside of these multiple platform configurations, we observe communities moving between platforms across a wide temporal range. Some move early in the simulation and quickly develop personalized recommendations that deliver preferred public good bundles; other communities stick around on platforms providing low levels of utility, and only move later in the simulation when alternative platforms have successfully tuned recommendation systems to communities with a similar preference set.

If this observation of the internal dynamics of multiple algorithmic platforms is true, then these configurations might be sensitive to variation in the number of platform options available. To that end, I extend this first experiment to configurations with small ($N_p = 5$), medium ($N_p = 10$), and large ($N_p = 20$) numbers of platform options.

3.2 Mixed Platform Configurations and Extremist 'Utility Vampires'

While we can learn something important about how different platform types sort communities in homogeneous settings, we know that, observationally, the existing online platform system is heterogeneous. Many different types of platforms coexist. Communities move between them often; their structures and behaviors are shaped by varieties of platform governance. The second experimental configuration compares the performance of and dynamics within different types of platforms in a mixed setting, with $N_p = \{3,9,27\}$ and $N_c = \{100,300,900\}$. N_p is split equally between each of the three platform types, and, at $t = 0$, communities are equally distributed between all platforms. In addition to examining the average number of moves and average utility per capita by platform type, mixed platform configuration sims also record the final distribution of communities per platform type.

Institution	**Average moves**	**Average utility**	**Communities (count)**	**Communities (ratio)**
Overall	20.2	5.85	100	1
Direct	—	6.91	35	0.35
Coalition	—	5.13	15	0.15
Algorithmic	—	5.32	50	0.50

Table 3.3: Mixed multiple platforms, $N_c = 100, N_p = 3$

In the “small” mixed configuration in Table 3.3, we see an interesting divergence occur (table 3.3). The direct platform delivers a high level of utility per capita at almost 7, while the coalitionand algorithmic platforms linger around 5, half of potential community utility. Significant sorting also happens. Communities have largely evacuated the coalitionplatform for the direct and, mostly, algorithmic platforms. Recall that in the first experiment algorithmic platforms performed better in the multiple platform sim; in this case, that utility advantage disappears, despite attracting more communities. We see the same pattern in the “medium” mixed configuration.

However, as both N_c and N_p increase, the performance of the simulation gets better (Table 3.4). For the "large" mixed configuration, overall average utility has increased to 6.3, driven by gains to the direct and algorithmic platform types. Interestingly, the proportion of direct platforms has shrunk relative to coalition platforms despite communities in direct spaces enjoying a 2 point utility advantage. The distribution of utility across tenant communities is also interesting. In Figure 3.2, the algorithmic and coalition platforms offer a more condensed utility distribution centered around the mean, while the direct platforms have larger variance. The average number of moves increases as well, pointing to more instability across the system. As the *status quo* bundles on platforms change, communities are forced to move or discover better platform candidates.

Institution	**Average moves**	**Average utility**	**Communities (count)**	**Communities (ratio)**
Overall	42.27	6.29	100	1
Direct	—	7.19	174	0.19
coalition	—	5.12	192	0.21
Algorithmic	—	6.58	534	0.59

Table 3.4: Mixed multiple platforms, $N_c = 900, N_p = 27$

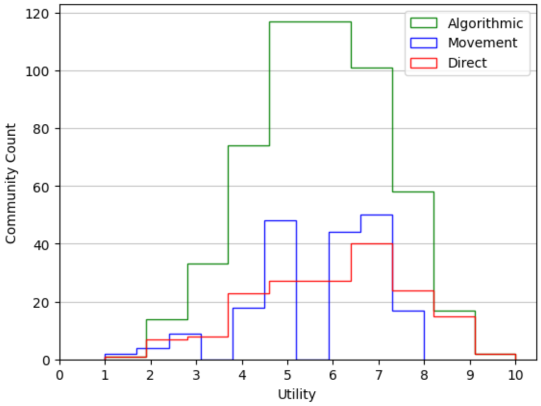


Figure 3.2: Histogram of community utilities, $N_c = 900$, $N_p = 27$

To answer the final question posed above, I modify the mixed platform configuration to build an *extremist* sim configuration. How should we model political extremist communities? One way is to simply vary their preference set a few standard deviations from the median. So, for example, some proportion of community types could have preference such that $B_c = 0^{N_a}$ or $B_c = 1^{N_a}$. This is the most common approach for defining "extremists" in spatial models of politics. For example, defining “extremism” according to DW-NOMINATE score or contribution-estimated ideology, placing candidates or parties on a sliding scale relative to some global distribution of ideology is widespread in studies of American electoralism (see (91; 191; 73) for widely cited examples). This approach has the advantage of easily fitting into existing models of median voters, lawmakers, or parties. Theories that cover typical policy deliberation and movement can be deployed to understand, for example, adherents to the Qanon conspiracy theory or a wide variety of centuries-old antisemitic lies. Unfortunately, this extension is mostly atheoretical; worse, it can actually normalize extremist and conspiracy ideologies by situating them as just another option along a continuum of “mainstream” political commitments. As a result, political commitments out of the mainstream—say, removing all export tariffs—are theoretically equivalent to political violence or genocide.

Instead, I turn to the burgeoning literature on contemporary political extremism, an emerging interdisciplinary area that draws on lessons from sociology, communication, and cultic studies. Here extremism is defined as the belief that in-group survival cannot be separated from hostile or violent action towards the out-group (19; 77). It consists of not just ideological beliefs, but a shared mythology, social identity, group threat, and commitment to a repertoire of action. In other words, it is multidimensional and theoretically complex. For this particular modification to the sim, I capture one dimension: the commitment to hostility towards out-group members. To model this, I introduce two different types of communities—mainstream and extremist—and add additional term to the community utility function to distinguish between them. Extremists online have a limited repertoire of cation compared to offline extremists, who can engage in physical harassment and violence. Instead, these communities turn to harassment, hate speech, and other actions to degrade the experience of out-group targets. In other words, they act as “utility vampires,” extracting utility from users and communities they interact with. The utility function for extremist communities is given as:

$$u_{ce}(B_p, B_c) = \sum_{a=1}^{N_a} \delta(b_{pa}, b_{ca}) + \sum_{cm=1}^{N_{cm}}$$

where N_cm is the number of neighboring mainstream communities on the platform. Extremist communities steal one unit of utility from each community on direct platforms, each coalition neighbor on coalition platforms, and each algorithmically determined grouping on platforms with SVD recommendation systems. Thus, the utility for mainstream communities is given as:

$$u_{cm}(B_p, B_c) = \sum_{a=1}^{N_a} \delta(b_{pa}, b_{ca}) - \sum_{ce=1}^{N_{ce}}$$

where N_ce is the number of extremist communities nearby. Neither extremist nor mainstream communities can observe the composition of platforms before moving; they decided to relocate based solely on the same decision metric as the first two experiments. Extremists also have "extreme" preferences for platform bundles, such that $B_c = 0^{N_a}$ or $B_c = 1^{N_a}$, randomly assigned. This incorporates elements of spatial political models described before as well.

How do extremist and mainstream communities interact? This final experiment looks at average utility per platform type *and* community type, the composition of community tenants on platforms, and the final count of communities per platform type. It also examines the internal dynamics of extremists moving between sim steps. The first iteration of this experiment uses three different single-institution configurations, testing $N_p = 5$ and $N_c = 100$ parameters across each institution type with 5%, 10%, and 15% of extremist community types. The central tension in the extremist's utility function is the difference between utility obtained from an unfavorable bundle of public goods and the utility parasitized from mainstream communities. On the one hand, extremists have distinct preferences over how platform governance should occur—they might prefer, for example, the lax policies against hate speech allowed on Gab or the lack of community boundaries on Twitter that amplify their content to non-extremists—but also prefer to have targets for out-group hostility. If extremists on a platform push the platform *status quo* too far towards their own preferred bundle, mainstream communities will desert it and they will lose opportunities to feed.

Table 3.5 above shows the results of single-institution, multiple-platform sim configurations. What is notable here is that, at the low- and medium-end of 5% and 10% extremists

5% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Direct	5.88	26.40	4.80	37.44
Coalition	4.71	23.80	3.71	44.77
Algorithmic	6.56	17.80	5.96	29.84

10% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Direct	5.65	23.00	3.72	40.69
Coalition	4.91	20.50	3.17	44.13
Algorithmic	6.43	20.40	4.87	34.72

15% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Direct	5.56	24.67	2.29	43.51
Coalition	4.32	20.60	1.45	43.31
Algorithmic	5.42	18.33	3.14	43.50

Table 3.5: Multiple platforms + extremists, $N_c = 100$

respectively, the overall efficiency of the platform system is unchanged from configurations without extremists. Average utility for each platform type is essentially unchanged from Table 3.2. The main difference is that much of the volume of utility across the system has been redistributed from mainstream communities to extremist communities. Only the algorithmic platforms manage to deliver over half of potential utility to mainstream communities, and then only at the 5% level. The distribution of mainstream community utilities is interesting as well. As before, direct platforms have high variance distributions; some communities in medium- and high-extremist configurations even obtain negative utility with direct voting. This suggests that some communities in these configurations are enduring targets of extremists, and end up on platforms with favorable bundles of public goods but are nonetheless subjected to considerable vampirism by their extremist neighbors.

In the second iteration of this experiment, I extend the extremist model to include mixed platform configurations. Recall above that we observe empirically that the social media system often contains many different types of platform governance that shape community experiences. And platform types that perform well in single-type systems, such as algorith-

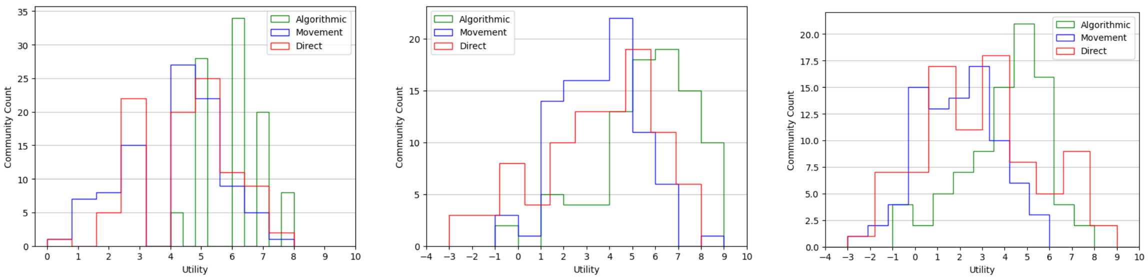


Figure 3.3: Histogram of mainstream utilities, $N_c = 100$, $N_{ce} = \{5\%, 10\%, 15\%\}$

mic recommenders, may experience performance declines in mixed-use systems. To examine this, I repeat the second experiment above with different levels of extremism, combining configurations with $N_p = \{3, 9, 27\}$ and the proportion of extremists $N_{ce} = \{5\%, 10\%, 15\%\}$. Because the utility functions of extremist and mainstream communities are sensitive to the number of communities in the simulation, I standardize these sim configurations with $N_c = 900$.

Strong patterns emerge from these iterations of the experiment. First, configurations with high platform consolidation perform much worse at delivering utility to mainstream communities than configurations with more platform choice when extremists communities are introduced. Second, Algorithmic platforms still command a larger share of community tenants, while direct and coalition platforms comprise a smaller, peripheral space in the system. However, contrary to the second experimental configurations above, coalition platforms outperformed direct and algorithmic platforms for mainstream communities as the proportion of extremists increase. And finally, two extremist behavioral dynamics emerge: first, I observe platform enclaves that shift public good bundles towards extremist preferences where extremists congregate; and second, extremists move around the system, disrupting platform *status quo* and siphoning off utility from mainstream communities.

Platform consolidation: fewer platforms and platforms with a higher proportion of communities generate the lowest utility per capita on mixed platform configurations where extremists are present. In Table 3.6 and Table 3.7 below, utilities are, in general, lower for $N_p = 3$ than $N_p = 27$. With few alternatives for mainstream communities the sorting mechanisms in coalition and algorithmic platforms work in favor of extremists, giving them excess power over not just public good bundles but also access to the utility of mainstream users. This result is driven by two dynamics. First, more platforms means a wider distribution of extremist communities across the system. Lower specifications of platforms means the minimum count of extremists per platform is higher. In these scenarios, higher number of extremists siphon more utility from mainstream communities; however, mainstream communities largely control changes to public good bundles, generating lower output for the left half of the extremist utility function. Second, because extremists are so integrated into mainstream spaces, the preference aggregation institutions on coalition and algorithmic platforms are more likely to construct groups that include both mainstream and extremist communities, lowering the value of the bundle offered to both. In other words, in a more consolidated system, mainstream and extremist communities are served content from one another that they prefer not to see, are denied access to preferred audiences, and forced to engage with discourse architectures that they do not like.

5% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.41	287.40	-9.95	48.50
Direct	5.34	189.10	-10.14	—
Coalition	7.37	191.23	-6.79	—
Algorithmic	6.65	503.33	-9.93	—

10% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.51	269.60	-23.82	46.00
Direct	5.20	172.70	-23.52	—
Coalition	8.21	162.22	-20.52	—
Algorithmic	7.56	496.93	-29.93	—

15% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.17	257.67	-39.39	43.50
Direct	4.81	172.12	-33.84	—
Coalition	9.21	169.50	-38.98	—
Algorithmic	2.51	437.00	-39.23	—

Table 3.6: Mixed platforms with extremists, $N_p = 3$

Coalition platform performance: Platforms with coalition aggregation mechanisms performed poorly in the first two experiments. However, once extremists are introduced, they occupy a more interesting space in the simulation system. For one, they outperform direct and algorithmic platforms across each configuration of extremists (see Table **??** and Table **??**). Coalition mechanisms deliver higher overall utility; but more importantly, mainstream communities obtain their highest per capita utility on coalition platforms and extremists their lowest in all but one case ($N_p = 3$ and $N_{ce} = 5\%$). This result is explained by looking at the final bundles offered by platforms at the end of each simulation iteration. Movements become dominated by either extremist or mainstream communities fairly early in each run of the experiment; this has the effect of rapidly shifting the *status quo* bundles offered by platforms towards extremist or mainstream preferences. With extremists obtaining high levels of utility from the right side of their utility function, $\sum_{a=1}^{N_a} \delta(b_{pa}, b_{ca})$, they consistently choose the "stay" strategy for the duration of the simulation. In other words, once extremists find platforms with favorable public goods that allow them draw explicit community boundaries, they stay there. And the same is true for mainstream communities with similar preferences over possible bundles of public goods.

Extremist behavioral dynamics: But what about direct voting and algorithmic platforms? Although per capita utilities aggregated from across many iterations of a simulation configuration can tell us a lot about the efficiency and distribution of utility within a system, we have to look closer to answer questions about how extremist and mainstream communities interact. By tracking community movements across platforms and expressing them as a relational dataset—where nodes are agents and ties are agent movements between nodes—we can recover information about the historical paths communities take before sorting.

Coalition platforms harden into enclaves for mainstream agents with low Hamming distances betewen preference sets or extremist agents. Direct and algorithmic platforms, meanwhile, form a sort of symbiotic relationship. When I track the relocations of communities across the simulation, we see that informal groups of extremists tend to move back and forth between the same direct and algorithmic platforms. For example, Figure 3.4 depicts platform-community networks from the first cycle of on iteration of a configuration with $N_p = 27$. On the left, direct platform 902 has 18 extremist communities, marked in red, to 16 mainstream communities in green; on the right, algorithmic platform 906 has only

5% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.73	40.13	3.76	46.25
Direct	5.97	27.92	3.92	—
Coalition	7.49	19.56	4.90	—
Algorithmic	5.55	60.84	3.71	—

10% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.38	36.51	1.92	45.43
Direct	5.37	25.97	2.23	—
Coalition	7.38	24.37	3.62	—
Algorithmic	5.39	57.21	1.93	—

15% Extremists

Institution	Average utility	Avg extremist utility	Avg mainstream utility	Average moves
Overall	5.74	32.49	1.01	43.14
Direct	6.50	25.56	0.90	—
Coalition	8.04	23.00	2.09	—
Algorithmic	4.87	54.55	1.03	—

Table 3.7: Mixed platforms with extremists, $N_p = 27$

6 extremist communities, marked in purple, to a majority of mainstream communities in green. At $t = 1$, 902 has already had one cycle to change the bundle of public goods offered to the platform in the direct of the extremist preference set, and can be thought of as an "extremist" platform. 906, meanwhile, has a large number of mainstream communities, and has grouped the few extremists on the platform into algorithmically shaped spaces with lots of mainstream neighbors. In other words, 902 is providing extremists with utility through preferred policy outcomes but limited access to mainstream communities to siphon off utility. But 906 is the opposite, delivering many mainstream neighbors for the few extremist agents to browse from, but mostly mainstream public goods bundles.

At $t = 2$, the relocation of extremists primarily from the direct platform to the larger, mainstream-dominated algorithmic platform begins. New extremist communities move from the direct platform to the algorithmic one, occupying the recommendation-shaped spaces created by the few existing extremists. A few extremists from 906, marked in purple, leave

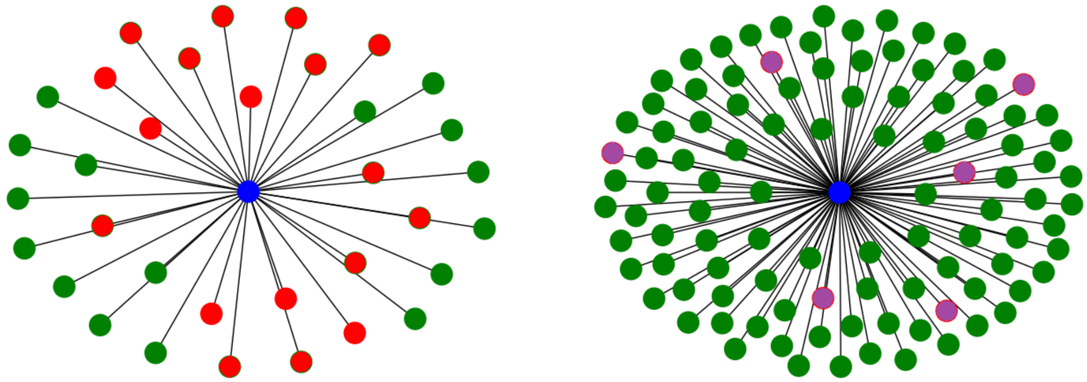
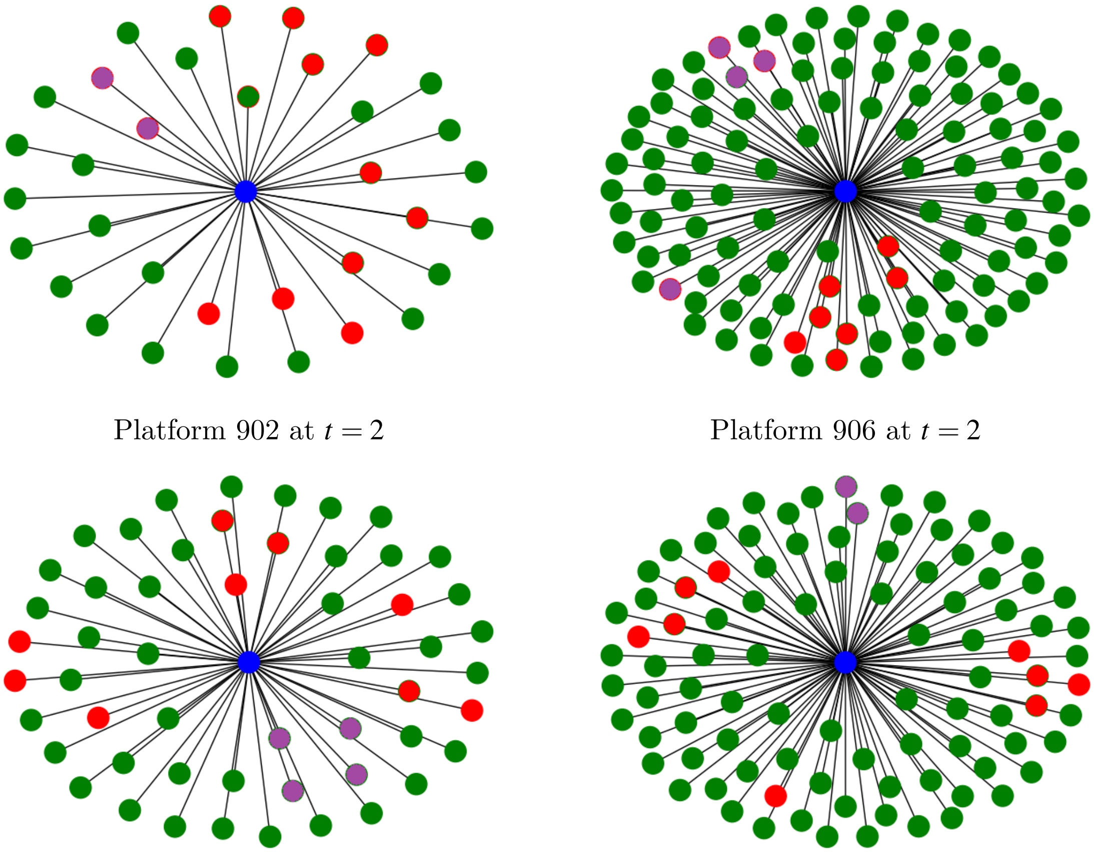


Figure 3.4: Direct platform 902 at $t = 1$ (left); algorithmic platform 906 at $t = 1$ (right). Extremists (red and purple), mainstream (green).

for better policy outcomes on 902. More extremist communities, 7 in total, move from 902 to 906, grouping in with large numbers of mainstream communities and obtaining higher utility from the right side of their utility function. Similarly in $t = 3$, another 2 extremist communities from 902 move to 906 (see Figure 3.5 below). During these two steps, the extremists in red moving from 902 to 906 are sorted across categories by the recommendation algorithm—largely due to the "cold start" problem, in which recommendation systems perform poorly when new users join. From within these artificial groupings, 902 extremists can obtain higher utility stealing from mainstream communities. We can imagine this as targeted harassment or, perhaps, producing extremist content that is repugnant to mainstream users. 906 extremists, meanwhile, move to the direct platform for better public good bundles. However, there are fewer mainstream users to harass or combat.

By the conclusion of the third step, however, the algorithmic platform has sorted extremists into spaces with other extremists, and their utility from public goods bundles is not enough to compensate for fewer mainstream users to steal from. In the final step (see Figure 3.6), most of the 902 extremists who spent a couple simulation steps in 906 return back to the direct platform. This four step cycle of extremist movement between smaller, peripheral platforms onto larger platforms seeking mainstream users to shake down for utility repeats for the duration of the simulation. Each cycle sees the concentration of extremists on direct platforms growing, and each "raid" of extremists from these peripheral spaces to larger,



Platform 902 at $t = 2$

Platform 906 at $t = 2$

Platform 902 at $t = 3$

Platform 906 at $t = 3$

Figure 3.5: Community movement across platforms. Extremists (red and purple), mainstream (green).

mainstream dominated spaces gets bigger. The net result, as we see from the tables above, is overall lower utility for mainstream communities on direct and algorithmic platforms.

3.3 Discussion

Results from the modified Tiebout simluation experiments show that platform dynamics are more complicated than previously suggested. Community interactions, especially across platforms with different regime types, have the potential to radicalize or provide substantial utility gains to extremist communities. These results from the simulation experiments should be interpreted as a device for building theory and directing empirical inquiry. We are left with the intuition that communities are competing for public goods on these platforms, different types of platforms deliver goods in more or less efficient methods, and communities

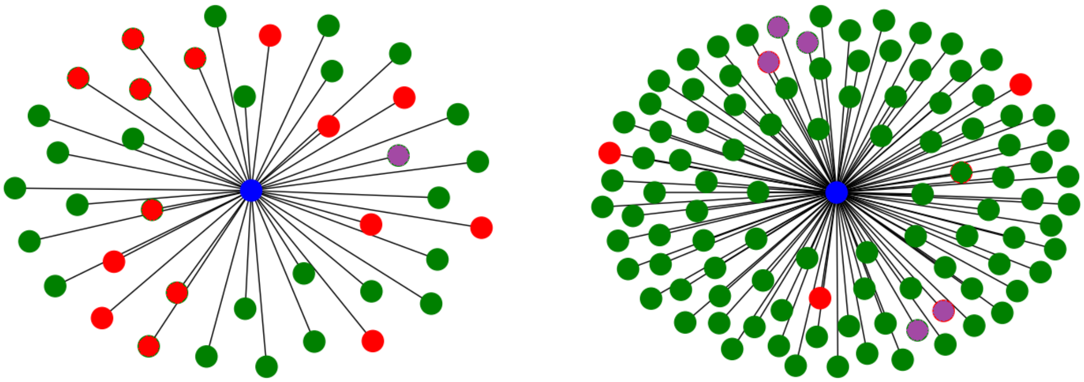


Figure 3.6: Direct platform 902 at $t = 4$ (left); algorithmic platform 906 at $t = 4$ (right). Extremists (red and purple), mainstream (green).

obtain utility from competition in different ways. Integrating these intuitions into existing theories to produce testable implications is our next task.

First, these results suggest that more platforms—and a wider variety of platforms—improve the overall utility of political communities. Consolidation into huge platforms leaves utility on the table for users. And there is certainly empirical evidence to support this. The contemporary landscape of social media platforms is not characterized by variety, but rather consolidation. Meta alone gobbles up billions of users on just three platforms; TikTok commands one-third of all social media users on the planet. Single platforms acting as gatekeepers to community social graphs may be more susceptible to interference by states interested in disrupting challengers at the application layer. Interfering with internet access at the infrastructure level, including complete blackouts, can be a double-edged sword for repressive regimes]; when monopolistic platforms act as pseudo-infrastructure, it may reduce those costs. Variety in media systems is also important for viewpoint and news diversity in democracies. Algorithm-shaped information spaces flatten community boundaries and can drive viral misinformation or homogeneous political information into platform-created bubbles.

Second, it suggests that extremist communities may profitably pursue strategies of frequent movement between platforms. Journalists working on the online extremist beat have long published stories on harassment or hate campaigns organized off-platform—see, for example, the case of Kiwi Farms, a small “direct”-type platform that became famous for organizing violent harassment campaigns in mainstream spaces. Many possible motivations for this behavior have been suggested: recruitment, “trolling,” persuasion, and so forth. Simulation results suggest that the motivation or strategy for extremist movement between platforms may depend on the social media structure itself.

Finally, the Tiebout simulation model highlights the importance of community competition within a larger structural system of many platforms. Community competition for public goods is not necessarily an inefficient means of distributing things like attention, engagement, monetization tools, community management features, and other aspects of discourse architecture. However, the online media system in which competition happens may produce more costs than benefits. Communities competing within just a few massive platforms generates lower utility for mainstream users, and makes it easier for extremists to find and target their hated out-group. It may also make extremism worse, as extremist communities find themselves competing with each other for the spoils of both victimizing out-group communities and shifting public good bundles towards their preferred outcome.

CHAPTER 4

Community Competition and Political Extremism

Online extremist political movements are on the rise. The internet has removed or weakened barriers to entry for anyone seeking out organized extremist groups, manifestos, propaganda, and content. While this has made it easier for extremists to radicalize internet users, it has simultaneously fragmented extremist movements into a loosely coordinated international network of many communities (45). These communities vary widely in their structure, leadership, ideological commitments, and level of extremism. What explains this variation in extremism at the community level? Why do some communities endorse more violent calls to action, while others embrace a less violent spectrum of beliefs and strategies?

Interdisciplinary scholarship on radicalization and extremism is still relatively new. This nascent subfield provides an alternative set of explanations for extremist violence to mainstream theories derived mostly from studying a narrow set of Islamic extremist organizations[1]. What falls under the umbrella of extremism, particularly the so-called "mixed, unclear, or unstable" (MUU) ideologies that comprise the bulk of terrorist violence committed globally, is an under-theorized space. This mix of seemingly incoherent ideological communities within broader movements has been characterized as "composite violent extremism" (76), "idiosyncratic terrorism" (149), or, in the words of FBI Director Chris Wray, "salad bar extremism" (172).

In this paper, I argue that variation in community-level extremism is driven by competition between online communities for the public goods supplied by social media platforms. To do this, I build a theoretical framework that identifies communities as the meso-level unit of interest for understanding online extremism. I bring together literature on social media platform governance, social movements, and competition between armed actors in conflict spaces to offer a system-level explanation for how competition for audience attention in

[1] Islamic extremism in the Middle East has also fractured and diversified beyond the traditional hierarchical, bureaucratic, foreign-based terrorism organization. However, religiously-motivated terrorism also declined by 82% and was overtaken by ideologically-motivated violence in 2021 (69). Therefore, this paper focuses primarily on ideologically-motivated extremism.

digital spaces could push political communities to increasingly extreme ideological commitments. To examine testable implications of this framework, I build a data collection pipeline that incorporates an overlapping snowball chain sampling algorithm and a human-assisted transformer classification mechanism to detect communities within the Qanon conspiracy and neo-sexist (or so-called "manosphere") political movements on major social media platforms.

Specifically, I argue that competition between communities within a movement drives extremism through two mechanisms. A more competitive platform makes attention and engagement more valuable. On platforms designed with community boundaries in mind, users can signal in-group commitment simply by joining the community. Communities can spend less time policing and hardening group identities. To claim attention and engagement goods, communities instead turn to more extreme claims and actions towards out-groups. However, on platforms with weak boundaries, communities do not survive without audience resources—both from sympathetic and oppositional audiences. Communities in these spaces need to constantly police boundaries between in-group and out-group identities, and try to claim platform goods by signalling more extreme commitments to the in-group ideology. Through this mechanism, extremist content is focused more on strengthening and purifying in-groups than sharpening threats against out-groups. To test this mechanism, I offer a relational model of audience competition between communities that shows that the share of extremist content within communities is higher on platforms with a more competitive environment. Specifically, evidence from these data show that communities that overlap, sharing more users, express a higher ratio of extremist content. I also find evidence that communities on platforms with strict community boundaries by design, like Reddit's "subreddits" or YouTube's channels, have higher levels of out-group-focused extremism.

These findings have important implications for how we conceive of online extremism. For one, it shows that platform design *by itself* can incentivize community extremism simply by pitting communities against each other for resources. Moreover, it differentiates between out-group-focused extremism, which is more likely on platforms with strong algorithmic recommendation systems, and in-group-focused extremism, which is found more often on platforms with strong community boundaries. The paper is organized as follows. First, I describe how the theory of community competition is situated within the growing field of extremism research and within traditional political violence work. Second, I expand on the theoretical argument above, explaining how the incentive structure of communities changes in response to platform design. Then, I introduce the data collection and coding pipeline, explaining the overlapping snowball chain sampling algorithm and human-assisted transformer classification process, followed by case selection strategies. Finally, I use a simple beta regression model to show strong support for the relationship between competition and community extremism.

4.1 Understanding Community Extremism

Political extremism is a set of beliefs, norms, and behaviors built around hostility towards a hated out-group. Identification with the extremist in-group is inseparable from violent acts—from harassment to killing—towards the out-group (19). This definition has two equally important components. First, extremism is preoccupied with identity and ideology, which sets it apart from other frameworks of political violence. A focus on group membership and belonging means extremists are constantly policing boundaries, sorting individuals into acceptable in-group or hated out-group. Extremist identity, like any collective identity, is reproduced through interaction, negotiation, and exchange (100; 143; 175). Social identities are adopted or attributed to others in order to situate individuals in social or political space, and are often grounded in social or political roles. Extremism relies on hardening a political identity, making identification with the extremist in-group the primary personal identity and defense of the in-group sense of "we" the primary role of adherents.

The second component of this definition is a call to action. In order to defend the in-group identity, extremists must be ready to engage in hostile action towards members of the out-group. Success for the group is contingent on not just holding extremist beliefs—such as disapproving of a particular religion—but actively harming targets of the belief—such as arresting or deporting members of the disapproved religion. I conceptualize these two components below as *in-group-focused extremism*, which is aimed at rigorously defining the boundaries of extremists identities and belief through interaction and reproduction of extremists texts, norms, and behaviors; and *out-group-focused extremism*, which aims to persecute targeted out-groups with hostile and, occasionally, violent acts.

In this paper, I focus on the increase of extremism through community-level radicalization. Typically, "radicalization" is a concept used to refer to the socio-psychological process of an individual developing extremist ideologies and beliefs (113; 12). Often, this means the person is adopting or constructing belief systems that justify the use of violence against an out-group, or actively supporting violence for political purposes against an out-group (137). Here, I propose a distinct process in which the socially-constructed and collectively agreed upon ideological commitments, norms, and identity of a community become increasingly extreme. While it is true that community-level extremism may require at least some individual-level radicalization among community members, it nevertheless is the result of a different mechanism.

Worsening community-level extremism may bear the closest resemblance to so-called "push" factors for radicalization (203). In this category are structural mechanisms for individual radicalization, like cycles of state repression, poverty, lack of economic opportunity, and so forth. These mechanisms overlap considerably with conflict research on why individuals join rebel groups or criminal organizations. Studies of structural causes of recruitment or often rely explicitly on Becker-style models of economic replacement, in which profit elasticity between licit and illicit activity pushes individuals to join violent black markets or engage in violence themselves (16; 51; 14). Or the underlying theoretical intuition implicitly relies on this logic, where deprivation or replacement occurs through non-economic means. For example, real or perceived political exclusion or power shifts between ethnic groups make accepted avenues of political change less "rewarding" and the costs of violent action less "costly" (see, for example, (215) or (150)).

I characterize an online political community as collection of users who share a common identity or set of identities and, through frequent interaction, construct a set of behavioral norms, ideological commitments, and aesthetics. This definition draws together insights about how online communities create social cohesion through feelings of camaraderie, empathy, and social support (96; 164). It also incorporates how social identity construction can situate users within a community or multiple communities. Community identities are complicated attachments subject to variation in roles (185), groups, and personal relationships (180). The salience of a community identity may vary according to situations and interactions (184; 173). Communities may grow or shrink in response to external events that provoke greater salience. Community identities also vary in centrality (186) or prominence (78) as users actively construct and reconstruct identity hierarchies. Thus, community definition centers around interaction. Without user interactions, the community and the text of social norms, ideologies, and aesthetics is inert and unobservable.

While I focus on the online component of these communities and restrict observation to a handful of platform spaces, it should be noted that they rarely exist purely online (161). Interactions—and therefore the construction and reproduction of norms and ideas—are not typically restricted to a single platform medium, either, but spread across platforms and private channels. Membership is fluid, unlike organized and hierarchical groups; although platforms themselves have near-real-time insight into user categories, unless they opt to make these categories public as a matter of design, community boundaries are opaque and in constant negotiation. These characteristics make the community a challenging theoretical concept and empirical subject.

4.2 Community Competition

Studying communities is a worthwhile avenue to understanding how online political movements develop and elevate extremist ideology and identity. Studies of political and social movements based around resource mobilization theory (141) have mostly focused on social movement organizations as the central component of movements (189; 53; 29). In this work, social movements are composed of hierarchical organizations that collaborate and compete for scarce resources (55). The resources available to organizations drive their repertoires of action—and the decision to embrace repertoires of violence. “Radicalization,” in this sense, can occur through competitive escalation during protest cycles (52).

The primary mechanism by which inter-organization compete in traditional social movements functions is through actual use of violence. Groups competing for recruits and support from radicalized constituencies want to acquire a reputation for success through violent action, for example (41). Or to differentiate themselves through violent tactics and progress towards goals (142). Competition through outbidding requires both large changes to organizational structure and constant adaptation through engagement with adversaries. These actions are meant to appease more radical audiences within the movement support structure or demonstrate the strength of competing organizations. In civil conflict spaces armed groups similarly compete over support from civilian populations. Both state and non-state actors attempt to extract resources from civilian populations (210). Resources in this case could be information, materials, geographic access, or even recruits. And armed groups may decide whether to use violence and coercion to obtain these resources, shift support towards their cause, or punish audiences that support opponents (108; 212; 60).

However, online communities are not competing over material resources, but rather attention and audience support directly. "Attention" has always been a key resource for social movements and political actors like rebel groups (198). Attention from audiences is a key lever for recruitment, mobilization, persuasion, ideological construction, and so on. However, attention in social movement and conflict studies is rarely conceptualized and explored directly, but rather as an instrument for some other more important frame (see, for example, (84) or (17)). Typically it is transmuted into "support"—voting for a candidate, or supplying information to a rebel group—or some other more direct resource. In digital spaces, however, "attention" takes on a more complex and multidimensional role. It behaves as a scarce and fluid commodity that can be measured, shaped, and harvested by the social media platforms that control the supply. But "attention" is also constitutive of engagement in social media spaces, as the primary innovation of this form of media is continuous interaction between users who occupy the role of both author and audience simultaneously.

This is important for understanding political communities online. Attention and interaction are crucial for building and sustaining identities, reproducing community behaviors, and generating the ideological commitments that tie members together. Social media platforms supply attention and engagement as public goods to communities and users. The institutions, processes, and norms of platforms determine the structure of public good provision. Control over the design of discourse architectures, the operation of content moderation systems, and the autonomy of communities means control over the flow of these goods. Platform "governance," insofar as the institutions and agents of social media companies play a political role in the internal functions of platforms, is built on sovereignty and power over information and interaction. Rules and policies on the platform emanate from this "sovereignty of silence" (92), commanding a monopoly on noise or silence from subjects—and challengers. Communities (and, in the aggregate, users) want to reach preferred audiences with content and be reached by preferred content. In other words, they have preferences about the volume and direction of public goods that platforms supply, as these resources are crucial for building and sustaining communities and community identities.

We can understand extremist communities in relation to the functioning of platform governance and the supply of platform goods. Recall that extremist communities are preoccupied with identity and ideology. Without interaction and attention, extremist texts and ideologies are inert; they require constant maintenance and reproduction to function. These communities seek to maximize their consumption (and production) of platform goods through two strategies: first, the development and spread of the community identity. Bigger communities capture more attention and engagement, and, because of the effects of algorithmic recommendation systems, command greater production of attention and engagement. Competing communities will turn to greater levels of extremism to harden identities, radicalize in-group ideological commitments, and pursue more extreme repertoires of action against out-groups.

Hypothesis 1 *More competition between communities leads to an increase in community-level extremism.*

Variation in platform structure and governance mediates this effect, however. One visible way platforms can do this is through the design and drawing of strict community boundaries. Twitter, for example, has few community boundaries to speak of. Users can signal their commitment to a particular community through their profile bio, username, or profile picture. Members of the Qanon community often have Qanon-related *shibboleths* listed (e.g., "where we go one we go all" or the acronym WWG1WGA; "watch the water"; "the storm is coming"; etc). This is in contrast to Reddit, where users self-sort into "subreddit" communities with distinct identities, rules, and significant autonomy. Membership in the community is less costly, and users do not have to spend attention or engagement performing in-group signalling. On platforms with stronger community boundaries and better tools for community autonomy, I expect that community-level extremism will be more out-group focused. A greater share of extremist content will target out-groups with increasingly violent calls for and repertoires of action.

Hypothesis 2 *On platforms with stronger community boundaries, community-level extremism will be more out-group-focused.*

However, on platforms like Twitter and TikTok, weaker community boundaries mean a greater share of user interactions must be focused on continuously building and policing the in-group. Communities on these platforms are more loosely integrated networks, where efficient information transfer is made difficult by community members unable to recognize each other without significant signalling. In this case, I expect that a greater share of extremist content will be focused on hardening in-group identities, making increasingly extreme ideological claims, and seeking more costly commitments from fellow members.

Hypothesis 3 *On platforms with weaker community boundaries, community-level extremism will be more in-group-focused.*

4.3 Extremism in Two Online Political Movements

To examine the effects of competition between communities on community-level extremism, I focus on two online political movements from the past decade: the Qanon conspiracy movement and the neo-sexist movement. While both of these political movements have historical antecedents, it is useful to recall how and why these two cases are distinct from the universe of conspiracy theorizing and the broader category of misogyny or patriarchy, respectively.

The Qanon conspiracy movement grew out of the syncretic overlap of a genre of online live action role-playing (LARPing), a 2016 viral conspiracy centered on a pizza restaurant in Washington, DC, and the so-called “Q drops”—content posted to the 4chan and 8chan (now 8kun) websites. LARPing as well-placed sources in various state governments as a form of recreation, to spread conspiracies or extremism, or both has long been a feature of online forums. Online versions of this role-playing date back to at least the bulletine board systems (BBS) of the early 1980s, where users could exchange messages by "posting"—the origin of the term—to virtual message boards fashioned after those found in college coffee shops and student union buildings. The most famous of these is likely the "John Titor" or "TimeTravel_0" character, a pseudonym that appeared on The Time Travel Institute BBS in 2000 that claimed to be an United States military time traveler from 2036 (171). Positive forum response to this user, who continued posting through 2001, has generated many similar LARPs across many platforms, including the anonymous imageboard 4chan. In the aftermath of the hacking and subsequent leaking of emails from John Podesta, the campaign manager for American presidential candidate Hillary Clinton, many such politically-oriented LARPs sprang up on 4chan including the so-called "Qanon." These included "FBIanon" (who claimed to be a high-level official in the US Federal Bureau of Investigation), "High Level Insider" (who made no specific claims about government affiliation), and "White House Anon" (who claimed to be a high level official in then-President Donald Trump's administration).

Canonical or "mainstream" Qanon likely began with the October 27, 2017 post to 4chan and continued until November of the same year; "Q" posts after this were migrated to the 8chan imageboard and are often considered "second Q" by conspiracy researchers (see, for example, (4)). The user or users who authored these posts claimed to be a high-level official in the US government with "Q" clearance, an access authorization in the Department of Energy that grants access to highest-risk information, such as Critical Nuclear Weapon Design Information (CNWDI) (2). Early posts were cryptic and esoteric, and the goal of these statements seemed to be leading users to decode and discover secrets hidden inside of them. Q's most viral claim was that then-President Trump and a team of "white hats" (read: good guys) were waging a global war in secret against a cabal of current and former government officials in many countries (but centered around former Secretary of State Hillary Clinton), celebrities, international organizations, news companies, and others. Core to these beliefs were a series of prophecies: the "Great Awakening," where non-believers would be forced to acknowledge the truth of Q's claims; "The Storm," in which many of the high-level political opponents of President Trump would be killed; and the "Great Reset," where the financialized global economy would crash, revert to a metallic monetary system, and eliminate more than half of all human life on Earth.

Central to the Qanon movement is the desire for mass violence against perceived political opponents in the United States, who are cast as members of an out-group engaged in existential warfare against the in-group. This ideological commitment fits easily in my definition of extremism. However, the communities that formed under the Qanon movement—especially during the early stages of the COVID-19 pandemic in 2020 and 2021—are varied and international. In the first half of 2020 alone international Qanon pages on Facebook grew by 5,700% and United States Qanon pages grew by a staggering 22,000% (136). And for many of these communities, the Qanon conspiracy theory serves as a "theory-of-theories" from which they pick and choose elements of prophetic belief, mixing in their own collection of ideologies. Wellness and alternative health communities, for example, have adopted some of the less violent aspects of Qanon, often without naming the movement itself (110). On the other hand, more extremist communities—in particular those in the white nationalist, Christian nationalist, and antisemetic spaces—have co-opted parts of Qanon for recruitment and propaganda purposes (71). This has produced a chimera of mixed and often self-contradictory ideologies under the umbrella of the Qanon movement.

While the new religious aspects of Qanon allow us to trace liturgical and ideological elements of the movement back to a single source (7), the neo-sexist movement is a more amorphous phenomenon. "Neo-sexism" may not even be the most appropriate terminology for the rise of popular violent misogyny in online spaces; coined in the 1990s, neo-sexism refers to the assertion that gender equality has already been achieved, gender-based discrimination does not exist, and traditionally dominant forms of gender—masculinity—are victimized (139). Online this movement expresses itself mostly as antifeminism, a reactionary and hostile response to gendered out-groups entering male spaces (111). Offline attention is often paid to especially egregious versions of this digital antifeminism, such as calls to strip women of the right to vote or human trafficking. However, antifeminist politics online have a distinct form of masculinity, recruitment, content and persuasion, and ideology that differentiates it as a digital political movement. The collection of communities that contribute to this movement is colloquially known as "the manosphere."

Misogyny in computing has been around at least since Ada Lovelace, widely regarded as the first computer programmer in history, published her algorithm for Bernoulli numbers on the Babbage Analytical Engine (112). However, the modern online version of the neo-sexist movement and "manosphere" communities can be traced to the so-called "Gamergate" controversy in 2014. Gamergate started as a misogynistic revenge blog published by a man angry at his former partner, a game designer, and spiralled into a violent harassment campaign against women and gender non-conforming game developers, journalists, critics, academics, and others in the gaming community (187). Although not the first targeted violence against women in gaming—the primary subject of Gamergate, Zoe Quinn, had been previously harassed and threatened after publishing her game *Depression Quest* in early 2013—Gamergate was notable because of the way many disparate communities of the nascent "manosphere" rallied around the multi-platform campaign. Communities include: "incels", or involuntary celibates, an identity formed around mostly white, heterosexual men unable to find a romantic or sexual partner (98); the so-called "red pill" communities, a classic neo-sexist identity that claims to liberate men from the misandry of modern society (83); men's rights activists (MRAs), a more straightforwardly political identity that is concerned with eroding legal rights for men, particularly in child custody and divorce; "pick-up artists" and similar identities built around sexual encounters with women who are often found, paradoxically, alongside nascent "men going their own way" (MGTOW) communities who seek to remove women from public life entirely.

Similar to the Qanon movement, the manosphere that emerged from this storm of targeted harassment in 2014 is a chimeric stew of contradictory and often incoherent ideologies. Neo-sexist communities struggle against each other to define the out-group—should it include gender non-conforming people and transgender women, or just cisgender women?—and to calibrate the appropriately hostile response. In general, though, this loose group of communities, spaces, and subcultures are united by misogyny and hostility to non-male out-groups. Both movements have had outsized and violent offline effects. Qanon communities and loosely affiliated groups that share some Qanon beliefs were crucial for organizing and mobilizing the January 6th, 2021 attack on the United States Capitol. Men inspired by neo-sexist ideas and manifestos are responsible for a sizeable share of so-called “lone wolf” terrorism, including the 2014 Isla Vista shooting, the 2015 Umpqua Community College shooting, two separate Toronto attacks in 2018 and 2020, and the 2021 Atlanta spa shootings. Hostility towards women and modern feminism also forms a pillar of belief for Islamic extremist groups like Boko Haram, and a trend in integrating religious beliefs into contemporary manosphere communities—see, for example, violent misogynist YouTuber Andrew Tate’s conversion to Islam—is drawing these two movements closer together. Due to both the wide variation in commitments to extremist ideologies within these movements and the growing threat of political violence from the most extreme communities, Qanon and neo-sexism are appropriate cases for this study.

4.4 Data Collection and Measurement

I turn to a three part data collection and measurement strategy to understand variation in community-level extremism. First, I scrape data from Twitter, TikTok, Reddit, and YouTube using the official API of each platform. I employ a modified snowball chain sampling strategy to detect communities and capture social graphs within each platform. Then, my team of coders and I fine-tune a foundational large language model (LLM) to perform pre-processing and classification on this dataset. The final dataset contains social graph data, where nodes are users and ties are subscription-style connections (variously, “follows,” “subscriptions,” or “friends”), for communities within the Qanon and neo-sexist movements on four platforms.

Observations are on the community-movement-platform level. To capture community-level extremism, my key dependent variable, I construct measures of the *share of extremism*, *share of out-group-focused extremism*, and *share of in-group-focused extremism* in each community observation using the LLM classification tool. To capture community competition, my key independent variable, I measure the *degree of overlap* between sampled communities and the raw *number of communities.* To control for variation in platform governance, I incorporate measures of community autonomy for each platform. I summarize each step in this process below before moving on to modeling and results.

4.4.1 Community Datasets

For the analysis of community competition within the online Qanon and neo-sexist movements, I turn to four major social media platforms: Twitter, TikTok, Reddit, and YouTube. Each of these are mainstream, generalist platforms with global audiences. Twitter and Reddit are primarily text-based platforms of comparable size, boasting 35 million daily active users (DAUs) and 52 million DAUs in 2020, respectively. TikTok and YouTube are video-based, although the former is limited to just 10 minutes in length. Although YouTube had nearly five times the DAUs in 2020 with around 230 million to TikTok's 40 million, TikTok was also the fastest growing mobile app in history that same year. Twitter and TikTok treat user communities similarly, with few, if any, community boundaries. Users construct their social graphs by subscribing (in both cases, "following") to other users directly; they primarily interact with their network through continuous, algorithmically curated feeds. Reddit and YouTube, meanwhile, give community members more tools for erecting and enforcing community boundaries. Both feature designed walls between communities through Reddit's "subreddit" system and YouTube's "channel" system. In both cases, users subscribe (both platforms use "subscribe" rather than "follow") to subreddits or channels, actively affirming a community identity and ensuring interaction with other community members. This is in contrast to Twitter and TikTok, where endless-scroll algorithmic feeds ensure that content from out-groups can careen across the site and reach users outside of communities of origin.

Selecting for these differences across potential platform cases allows me to focus on concomitant variations. For example, if we observe more out-group-focused community-level extremism on two platforms with weak community boundaries, but observe different levels of community competition, we can conclude that community competition is driving some of the variation in community-level extremism. I explain how I measure community competition below.

I acknowledge that focusing only a few major social media platforms has some weaknesses. First, there are issues of scale: although these platforms are large, they are dwarfed by Meta's biggest products, Facebook, Instagram, and WhatsApp, which collectively boasted more than 2.6 billion DAUs in 2020. However, accessing data on these platforms with the granularity needed for this project was not possible. Scraping data on

Meta products is prohibited by the user agreement unless performed through an official API endpoint provided by the company. Semi-public tools available to researchers like CrowdTangle offer limited access to the detailed social graph and user-generated content required. Facebook, in particular, offers a useful test case of community competition. Unlike the platforms studied here, Facebook is designed to provide users both endless-scroll algorithmic feeds (through the Newsfeed) and communities with strong and active boundaries (through Groups). Future research should explore the theoretical implications of community competition in a platform offering hybrid public goods like Facebook.

On the other end of the scale, heavyweight generalist platforms governed by public companies like Meta or Alphabet obscure the fat-tail of the internet participation distribution. Significant political interaction occurs on smaller social media platforms built on different principles than big social networking sites. Internet forums, for example, are legacy discussion sites where users hold asynchronous conversations across many topics; they resemble bulletin boards more than the so-called ”public square” of Twitter or Facebook. Forums are often moderated and funded by volunteers, not for-profit. Although most are small and focused on hyperspecific topics, many hold outsized cultural or political influence. BlackPlanet, a forum for Black users started in 1999, was the first social media site that then-Senator Barack Obama joined in early 2007 (1). Others occupy a more malign space in the online ecosystem. The account posting the so-called “Q drops” that inspired the Qanon conspiracy movement originated on the site 4chan before spreading to 8chan (now 8kun); the forum Kiwi Farms, already infamous for targeted harassment campaigns that resulted in the suicide of at least one victim, published a livestream and manifesto of the 2019 Christchurch mosque shooter. Communities on these sites can play important political roles. Focusing on bigger sites with more readily available data risks a “model organism” problem that threatens representativeness and validity when generalized to the rest of the internet (199).

4.4.1.1 Sampling Strategy

To collect data from communities on Twitter, TikTok, Reddit, and YouTube, I turn to a series of primary and third-party APIs that provide direct, programmatic access to platform data. Twitter, TikTok, and YouTube all have enhanced API access for researchers accessible through off-the-shelf ‘python‘ tools. Pushshift, a user-built API interface, provided access to Reddit data (13). Each of these API endpoints provide similar access to user profiles and their social ties; for Twitter and TikTok, ties are user-to-user, while on Reddit and YouTube ties are user-to-subreddit or channel.

The data collection process has three steps to produce a relational dataset of user social graphs: pre-processing, snowball sampling, and community detection. These steps are built around a modified version of the ‘SbChain‘ community detection algorithm. ‘SbChain‘ is a community detection process, which takes as input a complete social graph and identifies communities around core nodes using a maximum common neighbor criteria (89). In plain language, the snowball chain implements an algorithmic version of the Friendship Theorem: in any group of at least three people, if any pair of individuals have precisely one common friend, then there is always a person (the so-called “politician”) who is everybody’s friend (129). ‘SbChain‘ works by, first, identifying those politicians in the full social graph by computing the local clustering coefficient for each node; in other words, finding those people who are most likely to have friends that know each other. From this initial set of seed nodes, the algorithm builds “snowballs,” or sets of nodes built by combining a politician with their best neighbors. Crucially, snowballs may overlap, with multiple communities claiming neighbors—a hyperparameter supplied by the researcher determines how many overlapping nodes should be absorbed into each snowball. The output is a set of crisp, distinct communities.

I re-purpose ‘SbChain‘ as a sampling and community detection algorithm with two modifications. Rather than supplying a full social graph, I begin with a set of known user accounts as seed nodes; snowballs are built from discovery of seed node follower graphs. And I discard the non-redundant node strategy used in ‘SbChain‘, allowing nodes to be part of multiple snowball chains. Chains may be combined, but only if the clustering coefficient of the union of the two chains is greater than the clustering coefficient of each chain individually. I describe the process in greater detail below. The resulting dataset is a collection of community subgraphs for each platform and movement clustered around a few “politican” user accounts connected by ties representing subscriptions or ”follows,” in which users may have membership in multiple communities.

First, in pre-processing, I build a set of seed nodes from known user accounts on each platform. In this case, we do not know the structure of the full social graph yet; rather, I use case knowledge to construct a list of nodes to start each snowball chain. I begin with a set of 30 accounts on each platform based on visible size of followers or subscribers and observation from digital ethnographic work in each political movement. I expand on my approach to digital ethnography in extremist digital communities in the appendix. I provide an abbreviated sample of seed accounts for each platform and movement below; the full list is withheld in accordance with data ethics policies. Again, more information about ethical guidelines followed while gathering data on users and content in extremist spaces is available in the appendix.

From this preliminary set of seed nodes, discovery and construction of snowball chains and communities proceeds in two steps. The first processing step starts by collecting the followers of the seed users (level 1) and the followers of nodes on level 1 (level 2). A stylized illustration of this first iteration is shown below. This step generates the initial seed graph, given as $G_s(V,E)$ where V is the set of seed nodes such that $V = v_i, v_j, ...v_n$, and E is the set of edges such that $E = e_{ij} = (v_i, v_j)$. From this step one subgraph, I compute the normalized degree of each node, given as $\lambda(v) = \frac{k(v)}{K}$ where $k(v)$ and K are the degree of the node and the maximum degree value in G_s respectively. The best neighbor of each seed node, the level 1 node with the highest normalized degree value tied to the seed, is added to the snowball chain S_n. Finally, for discovery I specify a new set of seed nodes from the level 1 subgraph, selecting the 30 users with the highest normalized degree value. This new set of seed nodes becomes the input for the next iteration of the first processing step, departing significantly from the ‘SbChain‘ community detection procedure.

In the second step, chains formed in step one, S_n where n is the number of snowball chains formed, are combined into communities based on the chain clustering coefficient (CCC) of each. The CCC is nominally the global clustering coefficient computed for just the subgraph S_n, and is defined as the ratio of closed triplets to the number of all triplets in the chain. If the CCC of any two combined chains is higher than the CCC of each chain, then the two are combined into a community. Otherwise, they are allowed to remain as separate chains. Combination and community detection continue until this criteria fails. Any chain that does not find a combination partner become a community itself. Crucially, this allows nodes to belong to multiple chains—and therefore multiple communities—without being combined during this step. This final step prevents communities from forming that are too similar, but also allows community overlapping.

I apply this sampling procedure using a set of 30 seed nodes for each platform-movement combination. These samples are temporally constrained, as each platform's respective API does not provide historical social graph data. In other words, we cannot track additions and subtraction to a user's subscription list over time without access to internal platform data or integration of archived data. Thus, the size and structure of a community discovered here is limited to the year in which it was collected: Twitter, Reddit, and YouTube were each sampled in late 2020, while TikTok was sampled in early 2022.

4.4.1.2 Community Competition Measures

Applying this sampling strategy to each platform with initial seed users from both Qanon and neo-sexist movements produce four relational datasets composed of unique user account ids, user profile information (typically just a few sentences), ties to other users in the form of subscriptions or "follows," and a set of community ids. From each platform-movement dataset, I construct two measures that capture community competition using the ids generated by the snowball chain sampling algorithm.

First, I measure the *number of communities* in each movement-platform sample overall. Cases with a greater volume of communities may be an indication of an overall level of competition across all communities on the platform. Second, I measure the *degree of community overlap* as a proportion of overlapping nodes for each community observation. Overlapping nodes—nodes with multiple community memberships—represent users that communities are competing over. This measure treats all overlapping nodes as homogeneous, although users that fall into this category may vary widely in their position within the network.

As we can see from Table 4.1, platforms vary widely in the number of communities they support within the Qanon movement. Twitter and TikTok, two platforms with few designed community boundaries, have many communities compared to nodes captured and

Platform	**Communities (count)**	**Average overlap (ratio)**	**Nodes (count)**	**Edges (count)**
Twitter	857	0.57	95,420	381,347
TikTok	257	0.33	45,612	100,346
Reddit	134	0.28	75,221	135,397
YouTube	221	0.15	102,666	122,932

Table 4.1: Community Dataset, Qanon

significantly more community overlap. Some Qanon communities share as many as three-quarters of their users on Twitter. The neo-sexist movement, meanwhile, is more integrated on platforms like Reddit and YouTube than the Qanon movement. See, for example, the community overlap ratio for YouTube in Table 4.2. Significant competition between neo-sexist communities on YouTube is consistent with the observed rise in influencers like Andrew Tate, who use pay-for-follow schemes to purchase dense networks of followers that spread misogynyst messages and boot antifeminist content.

Platform	**Communities (count)**	**Average overlap (ratio)**	**Nodes (count)**	**Edges (count)**
Twitter	400	0.25	65,705	95,985
TikTok	345	0.37	75,100	140,753
Reddit	144	0.27	71,337	130,121
YouTube	313	0.39	100,055	281,012

Table 4.2: Community Dataset, Neo-sexism

4.4.2 Community-level Extremism

In order to measure my key dependent variable, extremism at the community-level, I construct measures of the *share of extremism*, *share of out-group-focused extremism*, and *share of in-group-focused extremism* from each platform-movement dataset. Building these measures is a four step process: first, I sample user-generated content from each user present in the community-platform datasets generated above. Second, a coding team evaluates a small share of this textual content, scoring each document according to a domain-specific codebook measuring extremism and extremist focus. Third, using the codebook generated by the coding team and this small training set, I fine-tune a foundational large language model (LLM) to receive instructions from a codebook and extend the coding schema across the rest of the corpus. Finally, I apply this instruct-tuned LLM to the sampled corpus and conduct accuracy and reliability checks with other known extremism datasets.

As noted above, the community social graph datasets from each platform are constrained to the time period in which they were collected (2020 for Twitter, Reddit, and YouTube, 2022 for TikTok). Limitations in the availability of data means it is extremely difficult for a typical researcher to track historical changes in following or follower lists without access to internal platform data. I similarly constrain sampled user content—in this case, text-only content—to the years in which social graphs were sampled. I gather single, undirected pieces of user content: for Twitter, tweets but not quote tweets, retweets, or replies; on YouTube, comments under videos but not replies to other comments or videos themselves; on Reddit, top-level comments on posts but not posts themselves or responses to other comments; and on TikTok, comments on videos but not replies to other comments, videos, or audio content. These constraints limit what could be an intractably large dataset to merely big. Summary details for each dataset are described in the below table.

Classifying extremists or extremist content is a challenging exercise. Detection and sorting of extremist content grew out of hate speech detection research (see, for example, (72) or (134)) and into extremism, radicalization, and hate speech (ERH) detection. Many of these use text-only natural language processing and traditional machine learning algorithms, focusing on building supervised training sets—based on sentiment, pre-built lexicons or dictionaries, references to political entities, and so forth (e.g., (193)).

There are multiple difficulties with using NLP approaches for extremist classification. First, supervised learning is difficult to apply to NLP in general because labeling is expensive, both in time and coder experience. Manual coding by teams in political science—typically teams of undergraduates or graduate students, often with limited domain expertise—takes a long time and is vulnerable to intercoder unreliability (128). And achieving classification results with an acceptable level of test dataset accuracy and precision often means labeling a significant proportion of a given corpus. For example, (61) hand-labeled 1,000 documents from a corpus of 11,120 articles on drug-related violence; although no standard norm exists across the many fields that use text-as-data, this so-called “10 percent rule” is relatively common. Manually labeling 10% of the large dataset gathered from each platform's API was simply infeasible even with an undergraduate coding team.

Another difficulty with extremist content is that extremist communities frequently use coded language or ideological *shibboleths* to signal in-group status (94). Some of these linguistic shifts are so-called "algospeak"—changes in language used to evade algorithmic content moderation systems (177). This phenomenon is not limited to extremist communities; sex workers, LGBTQ, and bilingual communities frequently use high affinity terms for both algorithmic evasion and in-group signalling. It may also be used to rhetorically conceal more extreme ideological commitments. A classic online example of this is the antisemetic use of multiple parentheticals to identify Jewish users. Rather than make explicit antisemetic statements, extremist users may "bracket" out-group usernames like so: (((username))) (153). Meanwhile, some Jewish communities online have adopted a strategy of counterspeech by bracketing their own usernames to signal solidarity and resilience. A subset of shibboleth and affinity language, counterspeech also makes classification of extremist content problematic, as differentiating between extremist speech and counterspeech requires significant context.

The third difficulty with extremist content is that much of it is odious, violent, and harmful for consumption. Content moderators employed by platforms, mostly as third-party contractors, have well-documented psychological and health problems from viewing hate speech and violence on a daily basis (see, for example, (148) or (24)). These exploited workers bear the brunt of the horrific content uploaded to social media sites on a daily basis, keeping most of it from our sampled dataset. However, moderation policies, human error, and algorithmic decisions mean that extremist content can still contain quite a bit of hate and violence. Research-related trauma for social scientists working on issues of political violence and death, particularly issues dealing with sexual violence or violence against vulnerable communities of which the researcher is a part, is a real and understudied phenomenon. Personal risks are not limited to physical safety in the field, but also the psychological harm that comes from indirect exposure to violence (131). Team leaders and principle investigators can implement some safeguards against this, such as informed consent agreements with team members, limiting the length of coding sessions or exposure to violent content, mandatory breaks between sessions, or post-session debriefs as means of building a community of care (169). While I was able to implement all of these with my coding team, the volume of content that needed be processed to achieve the "10 percent rule" for typical machine learning classification models made them ineffective in the long-run. Rather than traumatize coders without the resources to provide them with adequate mental healthcare, I decided to seek alternative methods.

Using large language models (LLMs) for supervised classification tasks such as this one is a relatively new use-case. Mostly this is due to the expensive nature of LLMs; light-weight linear classifiers like fastText (107) or even larger pre-trained transformer-based language models like BERT (54) offer better performance per computational cycle. Transformer-based language models do offer advantages over linear classifiers for a complex task like classifying extremist content, however.

These models are likely to be better at dealing with highly context-dependent text common to extremist communities. Because transformer models are trained with bidirectional representations, they can overcome the so-called "unidirectionality constraint"—a limitation for linear language models that read train on text only from left-to-right. This makes sentence-level tasks where incorporating context from both directions is crucial sub-optimal. Transformer-based models are better at these tasks, and thus better at distinguishing between coded language, *shibboleths*, and counterspeech.

Most importantly, LLMs offer much higher performance with smaller training sets. This is especially true with domain-trained models that have been fine-tuned on instructions and text from the corpus of interest. BERT, for example, offers competitive performance on large classification tasks—more than 100,000 documents—with a training set of just 500 (63). Due to this flexibility and performance, I turn to LLMs to construct measures of extremism from the community dataset.

I then use instructional fine-tuning on the foundational 13 billion parameter LLaMa model released by Meta to researchers in February 2023 (197). LLaMa is an ideal model choice for this task for a few reasons. First, the LLaMa-13b model is performant on consumer hardware. It can theoretically run on an over-the-counter CPU, but excels on the highest-end GPUs. Second, the LLaMa 1 foundational models are trained on publicly available data sources including the CommonCrawl. The CommonCrawl corpus is well-known for containing a considerable amount of odious content, including hate speech and extremist speech (132). While this is undesirable for public-facing instructional uses, such as chat-based assistant services, this means that the foundational model has already been trained on the text captured in the content dataset built above. A non-exhaustive search of the corpus, for example, shows that content from each of the 30 seed users used to build the community dataset is present in the CommonCrawl corpus.

Accuracy, precision, and hallucination—an LLM-specific concern where the model produces inaccurate and nonsensical information—are of paramount concerns when using off-the-shelf models for new use cases. To explore the functionality of the instruction fine-tuning process, I also performed cross-validation testing of the LLaMa model on two existing datasets. First, the Sexual Violence in Armed Conflict (SVAC) dataset codifies a large corpus of human rights reports from the U.S. State Department, Amnesty International, and Human Rights Watch for the prevalence and intensity of war-time sexual violence (36). The project has a robust and well-developed codebook and consistent annotations for the labeled dataset. The second dataset is expert-annotated data classifying online misogyny from the European Chapter of the Association of Computational Linguistics (88). Data is generated from crowdsourced and expert labeled Reddit posts and comments, and is also accompanied by a detailed codebook. These datasets make ideal candidates for cross-validation. The SVAC data comes from lengthy documents written by expert observers and contains a considerable lexicon of domain-specific terms, which should challenge the contextual power of the LLM classifier. It is also likely that the corpus of human rights reports it is based on have already been consumed by LLaMa, as the plain text of these reports has been available online for the CommonCrawl to find since at least 2014 (65). The ECACL dataset, on the other hand, is concerned with very similar data as the community content dataset above; however, it is trained on posts from Reddit after 2020—data that is *not* available to LLaMA, which is updated only to early 2020. These selection allows me to see if the instruction-training process is sensitive to different linguistic domains or out-of-sample data. LLaMa instruction-tuned models of both of these datasets performed extremely well, labeling datasets with near-perfect accuracy after seeing just 500 documents from each labeled dataset. Summary statistics for these cross-validation tests are in Table

4.3 below.

Dataset	Instruction size	Accuracy	Precision
SVAC	250	0.65	0.67
ECACL	250	0.77	0.75
SVAC	500	0.94	0.94
ECACL	500	0.95	0.95

Table 4.3: LLM Performance

Extremist content is coded across four levels. First, at zero, content contains no extremist content. At level one, user-generated content contains identity-based abuse, identified by the use of pejorative expressions, negative connotations, harmful stereotypes, and insults on the basis of group membership. For example, a tweet that disparages women on the basis of a negative stereotype ("all women are...") would be coded a one. Level two extremism contains non-specific threats towards some target. Non-specific threats express action and intent to commit violence against a broader group and not by the poster themselves. Finally, level three extremism contains specific threats that express action and intent on behalf of the poster themselves. Expression reaches this level of extremism even if the target is a broad category (e.g., "Democrats") if it ascribes actions to the poster. In addition to level of extremism, I also code content for the target of abuse, non-specific threats, and specific threats. Because extremists often use violent or threatening language to police community boundaries or enforce norms and behaviors, I code content in which the subject of abuse or a threat is an in-group member as "in-group focused" extremism. The opposite, in which a target is a hated out-group, is coded as "out-group focused" extremism.

A coding team of undergraduate researchers coded 500 documents from each movement-platform corpus for a total of 2,000 pieces of content. They followed two separate codebooks, one for each movement. These plain-text codebooks were then reformatted and used for the instructional fine-tuning of the LLaMa model. The following is an example of an instruction-response pair in JSON typical of that given to the LLM:

> instruction: "Code the following as 1 if: text directs abuse towards an identity of group on the basis of gender; text contains any derogatory term which expressed

negative connotations on the basis of gender.”,

input: “don’t underestimate the cluelessness of a feminist. :femoid emoji:”,

output: “1”

We construct four final measures from this process. First, to measure the overall level of community extremism, we measure the share of all extremist content at all levels across the community; this is expressed as a proportion of community content. Second, we measure the score-normalized level of community extremism. This measure is the average score of all extremist content across the community. Finally, we measure the proportion of both the out-group focused extremism and the in-group focused extremism across the network.

4.5 Modeling

How should we model a dependent variable that is a proportion of outcomes within a multi-community network? One way to do this is to use a beta regression, which is very flexible and well suited for original proportions or rates. This simple regression assumes that outcome values are on the interval $(0,1)$, excluding 0 and 1. This assumption seems tenable, as it would be extremely unusual for *none* or *all* the content in a dataset with millions of posts to be classified as extremist. It also assumes that the dependent variable follows a beta distribution, $B(\mu,\phi)$ where μ is the mean and is expected to fall within the interval $(0,1)$, and are typically heteroskedastic. This assumption also seems reasonable, as heteroskedasticity is often observed when observation sizes vary widely; in the content dataset, community size varies significantly.

Thus, I estimate beta regression models for four dependent variables: the proportion of extremist content, the score-normalized level of community extremism, the proportion of in-group-focused extremist content, and the proportion of out-group-focused extremist content in a community. All models are run using data from the community and content datasets built above on samples from 2020 (Twitter, Reddit, and YouTube) and 2022 (TikTok). Each model includes, as measures of competition, the number of communities on each movement-platform combination and the proportion of nodes in the community that overlap with other communities. I also include movement and platform controls. Movement controls include

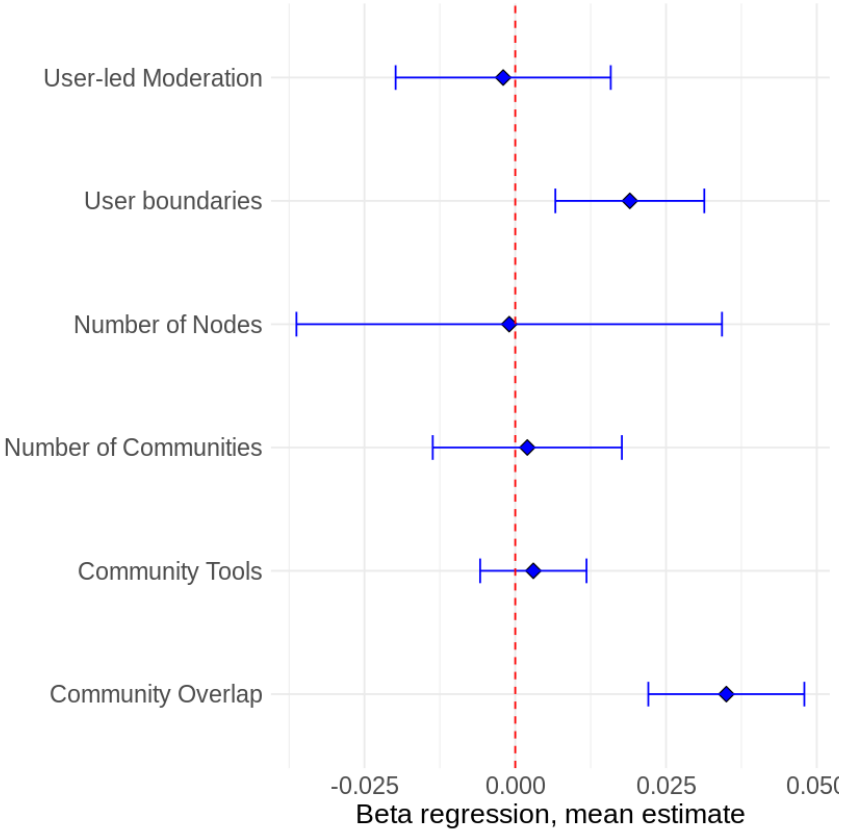


Figure 4.1: Model 1, overall share of extremism

the overall size of the movement on the platform in number of nodes; this controls for the dilution of extremist messaging as communities grow in size (206). Finally, I control for the level of autonomy platforms grant to communities, with binary variables indicating whether there is user-led moderation, the presence of community administration tools, and whether platforms allow users to choose their own community boundaries.

I begin by examining the relationship between community competition and overall levels of community extremism. Model 1 confirms my expectation in Hypothesis 1 that more competitive communities exhibit greater levels of community extremism. Figure 4.1 below shows the degree of community overlap corresponds to positive increases in the share of extremist expression within a community. The more users communities share and the more communities have to compete for platform public goods, the (1) more the share of extremist content spreads and (2) the higher the average level of that extremist content grows.

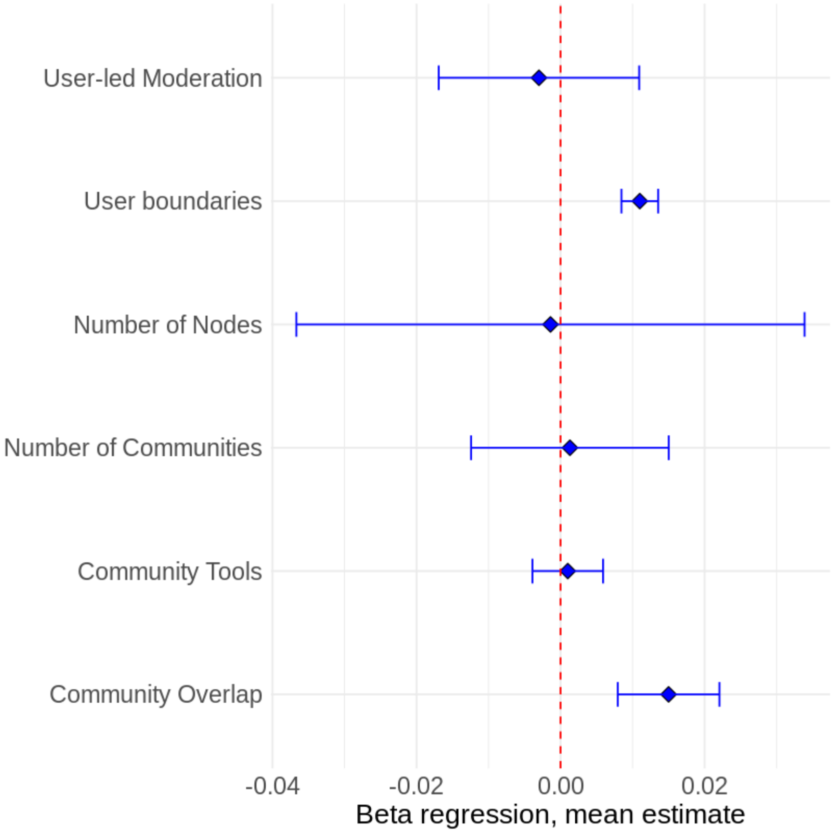


Figure 4.2: Model 2, share of out-group-focused extremism

Model 2 also confirms my expectations in Hypothesis 2: platforms with stronger community boundaries—in particular Reddit and YouTube—lead to extremist communities with a greater share of out-group-focused extremism. Where competitive communities on these platforms exist, we are likely to see more extremist content focused on threats to out-group members. Figure 4.2 shows these results below. Interestingly, these platforms also have higher extremist scores; not only is the share of out-group-focused extremism higher, but out-groups are targeted with more extreme ideological commitments and more violent calls to action.

Model 3, shown in Figure 4.3, produces null results for Hypothesis 3. On Twitter and TikTok, platforms with weaker community boundaries, communities with higher levels of extremism do not express more in-group-focused extremism. In fact, these observations have roughly the same among of in-group-focused and out-group-focused extremism overall.

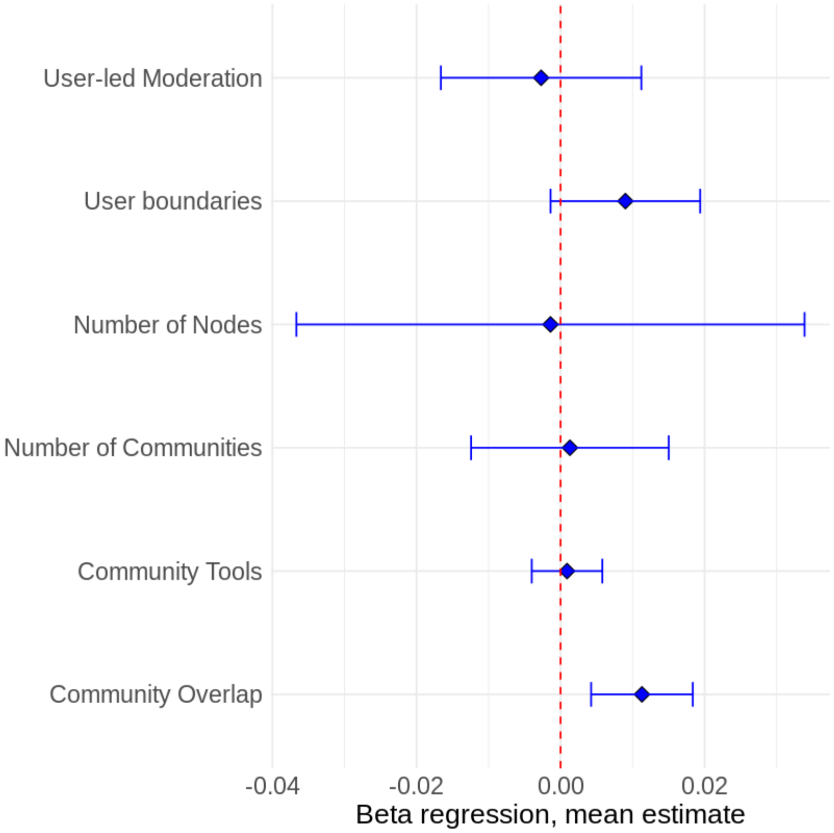


Figure 4.3: Model 3, share of in-group-focused extremism

Users in these spaces are just as likely to direct harassment and abuse at perceived opponents than they are to seek costly commitments from fellow members or make costly signals about their extreme ideological commitments.

Of the platform-level controls included, only the presence of community-drawn boundaries seems to have an effect on the level of extremist expression within communities. Neither user-led moderators nor the presence of community administration tools for users were significant in any of the specified models. This is an interesting and counter-intuitive finding, as studies of completely user-moderated and administered communities—like those on Telegram—often cite the user-controlled nature of these spaces as explanatory factors for radicalization (see, for example, (170)). It may be that hybridized platforms like YouTube and Reddit where platform regimes still have authoritative power over content decisions but cede some low-level, day-to-day powers of governance to users operate differently than fully user-controlled platforms like Telegram or private discussion forums. This merits further exploration in future research.

4.6 Discussion

Explaining variation within extremist political movements is crucial for understanding where future violent threats may emerge. While these movements appear monolithic or incoherent from the outside, the interior dynamics between and within different communities operate according to intelligible and familiar theories of competition and cooperation. It is tempting to see this complexity and argue that violent attacks perpetrated by members of these movements are stochastic or the actions of a "lone wolf." But we are beginning to understand that the identities, norms, behaviors, and, in some cases, aesthetics of these communities provide a fertile set of incentives to motivate acts of violent extremism.

I argue that competition over attention and engagement on social media platforms is one useful explanation for variation in violent extremism within movements. These findings suggest that platform design can be a driver for radicalization at the community level absent the usual "push" or "pull" factors. They also suggest structural changes that might stem the tide of rising ideological and political extremism online. If platform governance can be reformed to limit effects of competing micro-identities, we might slow community-level radicalization and introduce friction to the process of building new and powerful extremist identities.

Finally, this project finds more evidence for the central thesis of an emerging interdisciplinary field of contemporary extremism studies: the threat is the network, not any one group. When one community collapses or one traditional organized group disbands, the interconnected structure of online political communication rapidly replaces them. As long as platforms continue to provide attention and engagement in the form of commodities that can be easily converted into power, political extremists will continue to reorganize, reconnect, and rebuild.

CHAPTER 5

Conclusion

How do you join an online extremist movement? The simple answer is that you don't. The loosely affiliated communities that make up movements like Qanon do not have hierarchical structures. There is no introductory interview, fee to pay, oath to swear, religious conversion to undergo, or esoteric initiation. Nor do these communities threaten potential recruits with violence if they do not join, dig wells for your village, or fight a repressive government on your behalf. Rather, the community network grows around you, offering rewards for interaction with other members. "Membership" in the community is predicated on lived participation in the myth-making and construction of the extremist ideology and identity. It requires frequent maintenance, and users who go too long without interaction or signaling membership risk being left behind.

The first cults were Roman, the Latin "cultus" meaning "cultivation of the gods." To worship effectively meant active maintenance beyond simple veneration. Extremist communities demand the same of their adherents. Extremist identities are meant to be the primary sense of self, and everyday life is to be consumed by a desire to harm political opponents. But, with some exceptions, there is no central figure to organize around, no god to make offerings to. The community—and the movement itself—takes the place of the object of devotion.

The design and governance decisions that social media platforms make as governors of digital spaces helps explain which communities devolve into cults or cult-like networks, and which adopt increasingly violent repertoires of devotion. Across three papers, I describe a framework for understanding how private governance—or perhaps more accurately, an inattention to private governance by those running social media platforms—shapes these powerful actors. I argue that platforms are responsible for much of this radicalization, as the algorithmic imposition of many angry micro-identities pitted against each other in a contest for information and attention has unleashed a legion of small cults in our midst.

However, the fracturing of the social internet may well signal disruption of these trends and a new era of online extremism. Following the January 6th, 2021 attack on the US Capitol by a rich embroidery of different right-wing extremist communities, many of the mainstream social media platforms began aggressively banning and deplatforming extremists. This diaspora of Qanon and other communities generated a slew of so-called “alt-tech” spaces. Although some of these will fail, this “movement between lands” showed that big, generalist platforms need not occupy the central location of the social internet system. Unlike formal statehood, this form of private governance may not retain its status as the primary unit of political organization online. The possibility of future change to the most powerful communication system in human history highlights the importance of this work; before we can understand what comes next, we need to explain how the social internet has fundamentally shaped and changed our instinct for political community.

References

[1] (2012). Obama Networks on BlackPlanet.Com.

[2] (2023). Departmental Personnel Security FAQs.

[3] Alphabet, Inc. (2020). Form 10-Q. Technical report, U.S. Securities and Exchange Commission.

[4] Amarasingam, A. and Argentino, M.-A. (2020). The QAnon conspiracy theory: A security threat in the making. *CTC Sentinel*, 13(7):37–44.

[5] Anderson, B. (2006). *Imagined Communities: Reflections on the Origin and Spread of Nationalism.* ACLS Humanities E-book. Verso.

[6] Antelmi, A., Cordasco, G., D'Ambrosio, G., De Vinco, D., and Spagnuolo, C. (2022). Experimenting with agent-based model simulation tools. *Applied Sciences*, 13(1):13.

[7] Argentino, M.-A. (2022). Qvangelicalism: QAnon as a Hyper-Real Religion. In *Religious Dimensions of Conspiracy Theories*. Routledge.

[8] Bankes, S. (1993). Exploratory Modeling for Policy Analysis. *Operations Research*, 41(3):435–449.

[9] Barberá, P., Casas, A., Nagler, J., Egan, P. J., Bonneau, R., Jost, J. T., and Tucker, J. A. (2019). Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4):883–901.

[10] Barberá, P., Gohdes, A. R., Iakhnis, E., and Zeitzoff, T. (2022). Distract and Divert: How World Leaders Use Social Media During Contentious Politics. *The International Journal of Press/Politics*, page 19401612221102030.

[11] Bardhan, P. and Mookherjee, D. (2005). Decentralizing antipoverty program delivery in developing countries. *Journal of public economics*, 89(4):675–704.

[12] Bastug, M. F., Douai, A., and Akca, D. (2020). Exploring the "Demand side" of online radicalization: Evidence from the Canadian context. *Studies in Conflict & Terrorism*, 43(7):616–637.

[13] Baumgartner, J. M. (2019). Pushshift Reddit API. GitHub.

[14] Baysan, C., Burke, M., González, F., Hsiang, S., and Miguel, E. (2019). Non-economic factors in violence: Evidence from organized crime, suicides and climate in Mexico. *Journal of Economic Behavior & Organization*, 168:434–452.

[15] Bazerman, C. and Prior, P. (2003). *What Writing Does and How It Does It: An Introduction to Analyzing Texts and Textual Practices.* Routledge.

[16] Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of political economy*, 76(2):169–217.

[17] Benford, R. D. and Snow, D. A. (2000). Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology*, 26(1):611–639.

[18] Beran, D. (2019). *It Came from Something Awful: How a Toxic Troll Army Accidentally Memed Donald Trump into Office*. St. Martin's Publishing Group.

[19] Berger, J. M. (2018). *Extremism*. Mit Press.

[20] Berglas, E. (1976). Distribution of tastes and skills and the provision of local public goods. *Journal of Public Economics*, 6(4):409–423.

[21] Bert, F., North, M., Rovere, S., Tatara, E., Macal, C., and Podestá, G. (2015). Simulating agricultural land rental markets by combining agent-based models with traditional economics concepts: The case of the Argentine Pampas. *Environmental Modelling & Software*, 71:97–110.

[22] Bewley, T. F. (1981). A critique of Tiebout's theory of local public expenditures. *Econometrica: Journal of the Econometric Society*, pages 713–740.

[23] Bi, X., Qu, A., Wang, J., and Shen, X. (2017). A Group-Specific Recommender System. *Journal of the American Statistical Association*, 112(519):1344–1353.

[24] Biddle, S. (2020). Weeks After PTSD Settlement, Facebook Moderators Ordered to Spend More Time Viewing Online Child Abuse.

[25] Bonilla, T. and Tillery, A. B. (2020). Which identity frames boost support for and mobilization in the# BlackLivesMatter movement? An experimental test. *American Political Science Review*, 114(4):947–962.

[26] Brehm, J. O. and Gates, S. (1999). *Working, Shirking, and Sabotage: Bureaucratic Response to a Democratic Public*. University of Michigan Press.

[27] Brierley, S. (2020). Unprincipled Principals: Co-opted Bureaucrats and Corruption in Ghana. *American Journal of Political Science*, 64(2):209–222.

[28] Burnett, G. and Bonnici, L. (2003). Beyond the FAQ: Explicit and implicit norms in Usenet newsgroups. *Library & Information Science Research*, 25(3):333–351.

[29] Burstein, P. and Linton, A. (2002). The Impact of Political Parties, Interest Groups, and Social Movement Organizations on Public Policy: Some Recent Evidence and Theoretical Concerns*. *Social Forces*, 81(2):380–408.

[30] Cara, C. (2019). Dark Patterns in the Media: A Systematic Review. *Network Intelligence Studies*, VII(14):105–113.

[31] Carpenter, W. S. (1936). Politics: Who gets what, when, how. By harold D. Lasswell.(New york: Whittlesey house. 1936. Pp. Ix, 264.). *American Political Science Review*, 30(6):1174–1176.

[32] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017a). You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31:1–31:22.

[33] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017b). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–22.

[34] Chen, Tanya, L. S. (2021). Creators Are Turning To Shady Instagram Dealers To Restore Their Suspended Accounts. https://www.buzzfeednews.com/article/laurenstrapagiel/creators-instagram-dealers-restore-accounts.

[35] Chwe, M. S.-Y. (2000). Communication and coordination in social networks. *The Review of Economic Studies*, 67(1):1–16.

[36] Cohen, D. K. (2013). Explaining Rape during Civil War: Cross-National Evidence (1980–2009). *American Political Science Review*, 107(3):461–477.

[37] Collier, D. (1993). The Comparative Method.

[38] Collier, D. and Mahon, J. E. (1993). Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis. *American Political Science Review*, 87(4):845–855.

[39] Cook, B. J. and Wood, B. D. (1989). Principal-Agent Models of Political Control of Bureaucracy. *American Political Science Review*, 83(3):965–978.

[40] Coppedge, M., Gerring, J., Altman, D., Bernhard, M., Fish, S., Hicken, A., Kroenig, M., Lindberg, S. I., McMann, K., Paxton, P., Semetko, H. A., Skaaning, S.-E., Staton, J., and Teorell, J. (2011). Conceptualizing and Measuring Democracy: A New Approach. *Perspectives on Politics*, 9(2):247–267.

[41] Crenshaw, M. (1985). An organizational approach to the analysis of political terrorism. *Orbis-A Journal of World Affairs*, 29(3):465–489.

[42] Crystal, D. (2006). *Language and the Internet*. Cambridge University Press.

[43] Cunningham, F. (2002). *Theories of Democracy: A Critical Introduction*. Psychology Press.

[44] Dahl, R. A. (2008). *Polyarchy: Participation and Opposition*. Yale university press.

[45] Davey, J., Comerford, M., Guhl, J., Baldet, W., and Colliver, C. (2021). A taxonomy for the classification of post-organisational violent extremist & terrorist content. *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps*, page 78.

[46] De Mesquita, B. B., Morrow, J. D., Siverson, R. M., and Smith, A. (2004). Testing novel implications from the selectorate theory of war. *World Politics*, 56(3):363–388.

[47] Dean, J. (2017). Politicising fandom. *The British Journal of Politics and International Relations*, 19(2):408–424.

[48] Dehghan, E. and Nagappa, A. (2022a). Politicization and radicalization of discourses in the alt-tech ecosystem: A case study on Gab Social. *Social Media+ Society*, 8(3):20563051221113075.

[49] Dehghan, E. and Nagappa, A. (2022b). Politicization and Radicalization of Discourses in the Alt-Tech Ecosystem: A Case Study on Gab Social. *Social Media + Society*, 8(3):20563051221113075.

[50] Delfanti, A. and Phan, M. (2021). Rip it up and start again: Remix and co-option in the media industry. *AoIR Selected Papers of Internet Research*.

[51] Dell, M., Feigenberg, B., and Teshima, K. (2019). The violent consequences of trade-induced worker displacement in mexico. *American Economic Review: Insights*, 1(1):43–58.

[52] Della Porta, D. (2013). *Clandestine Political Violence*. Cambridge University Press.

[53] Della Porta, D. and Diani, M. (1999). Social movements. *European Studies*, page 365.

[54] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[55] Diani, M. (1992). The concept of social movement. *The Sociological Review*, 40(1):1–25.

[56] Diaz-Cayeros, A., Estévez, F., and Magaloni, B. (2012). Strategies of vote buying: Democracy, clientelism and poverty relief in Mexico. *Unpublished manuscript, Stanford University*.

[57] Díaz-Cayeros, A., Magaloni, B., and Ruiz-Euler, A. (2014). Traditional Governance, Citizen Engagement, and Local Public Goods: Evidence from Mexico. *World Development*, 53:80–93.

[58] Dixit, A. and Londregan, J. (1996). The determinants of success of special interests in redistributive politics. *the Journal of Politics*, 58(4):1132–1155.

[59] Dorff, C. (2017). Violence, kinship networks, and political resilience: Evidence from Mexico. *Journal of Peace Research*, 54(4):558–573.

[60] Dorff, C., Gallop, M., and Minhas, S. (2023a). Network Competition and Civilian Targeting during Civil Conflict. *British Journal of Political Science*, 53(2):441–459.

[61] Dorff, C., Henry, C., and Ley, S. (2023b). Does violence against journalists deter detailed reporting? Evidence from Mexico. *Journal of conflict resolution*, 67(6):1218–1247.

[62] Dynel, M. (2016). "I Has Seen Image Macros!" Advice Animals Memes as Visual-Verbal Jokes. *International Journal of Communication*, 10(0):29.

[63] Edwards, A., Camacho-Collados, J., De Ribaupierre, H., and Preece, A. (2020). Go Simple and Pre-Train on Domain-Specific Corpora: On the Role of Training Data for Text Classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5522–5529, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[64] Falk, J. K. and Drayton, B. (2015). *Creating and Sustaining Online Professional Learning Communities*. Teachers College Press.

[65] Fariss, C. J. (2014). Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review*, 108(2):297–318.

[66] Farmer, J. D. and Foley, D. (2009). The economy needs agent-based modelling. *Nature*, 460(7256):685–686.

[67] Ferrali, R., Grossman, G., Platas, M. R., and Rodden, J. (2020). It Takes a Village: Peer Effects and Externalities in Technology Adoption. *American Journal of Political Science*, 64(3):536–553.

[68] Fiesler, C. and Dym, B. (2020). Moving across lands: Online platform migration in fandom communities. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–25.

[69] for Economics, I. and Peace (2022). Global Terrorism Index 2022: Measuring the Impact of Terrorism. Technical report.

[70] Foramitti, J. (2021). AgentPy: A package for agent-based modeling in Python. *Journal of Open Source Software*, 6(62):3065.

[71] Forberg, P. L. (2022). From the Fringe to the Fore: An Algorithmic Ethnography of the Far-Right Conspiracy Theory Group QAnon. *Journal of Contemporary Ethnography*, 51(3):291–317.

[72] Fortuna, P. and Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

[73] Fowler, A., Hill, S. J., Lewis, J. B., Tausanovitch, C., Vavreck, L., and Warshaw, C. (2023). Moderates. *American Political Science Review*, 117(2):643–660.

[74] Freelon, D. (2015). Discourse architecture, ideology, and democratic norms in online political discussion. *New media & society*, 17(5):772–791.

[75] Gais, H. and Squire, M. (2021). How an encrypted messaging platform is changing extremist movements. *Southern Poverty Law Center, February*, 16.

[76] Gartenstein-Ross, D., Zammit, A., Chace-Donahue, E., and Urban, M. (2023). Composite Violent Extremism: Conceptualizing Attackers Who Increasingly Challenge Traditional Categories of Terrorism. *Studies in Conflict & Terrorism*, 0(0):1–27.

[77] Gaudette, T., Scrivens, R., Davies, G., and Frank, R. (2021). Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society*, 23(12):3491–3508.

[78] Gecas, V. (1982). The self-concept. *Annual review of sociology*, 8(1):1–33.

[Geddes] Geddes, B. Authoritarian breakdown: Empirical test of a game theoretic argument.

[80] Geddes, B., Wright, J., and Frantz, E. (2014). Autocratic breakdown and regime transitions: A new data set. *Perspectives on politics*, 12(2):313–331.

[81] Geertz, C. (1973). *The Interpretation of Cultures*, volume 5019. Basic books.

[82] Gillespie, T. (2010). The politics of 'Platforms'. *New media & society*, 12(3):347–364.

[83] Ging, D. (2019). Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities*, 22(4):638–657.

[84] Gitlin, T. (2003). *The Whole World Is Watching: Mass Media in the Making and Unmaking of the New Left.* Univ of California Press.

[85] Gohdes, A. R. (2020). Repression Technology: Internet Accessibility and State Violence. *American Journal of Political Science*, 64(3):488–503.

[86] Gray, J., Sandvoss, C., and Harrington, C. L. (2017). *Fandom: Identities and Communities in a Mediated World.* NYU Press.

[87] Gudowsky, N. and Bechtold, U. (2013). The role of information in public participation. *Journal of Deliberative Democracy*, 9(1).

[88] Guest, E., Vidgen, B., Mittos, A., Sastry, N., Tyson, G., and Margetts, H. (2021). An Expert Annotated Dataset for the Detection of Online Misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

[89] Gulati, J. and Abulaish, M. (2019). A Novel Snowball-Chain Approach for Detecting Community Structures in Social Graphs. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2462–2469.

[90] Hagen, L., Falling, M., Lisnichenko, O., Elmadany, A. A., Mehta, P., Abdul-Mageed, M., Costakis, J., and Keller, T. E. (2019). Emoji Use in Twitter White Nationalism Communication. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, CSCW '19, pages 201–205, New York, NY, USA. Association for Computing Machinery.

[91] Hall, A. B. (2015). What Happens When Extremists Win Primaries? *American Political Science Review*, 109(1):18–42.

[92] Han, B.-C. (2017). *In the Swarm: Digital Prospects*, volume 3. MIT press.

[93] Held, D. (2006). *Models of Democracy.* Polity.

[94] Hiaeshutter-Rice, D. and Hawkins, I. (2022). The Language of Extremism on Social Media: An Examination of Posts, Comments, and Themes on Reddit. *Frontiers in Political Science*, 4.

[95] Hillman, S., Procyk, J., and Neustaedter, C. (2014). Tumblr fandoms, community & culture. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW Companion '14, pages 285–288, New York, NY, USA. Association for Computing Machinery.

[96] Hiltz, S. R. (1985). *Online Communities: A Case Study of the Office of the Future*, volume 2. Intellect Books.

[97] Hinsley, F. H. (1986). *Sovereignty.* CUP Archive.

[98] Hoffman, B., Ware, J., and Shapiro, E. (2020). Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism*, 43(7):565–587.

[99] Horton, D. and Richard Wohl, R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry-interpersonal and Biological Processes*, 19(3):215–229.

[100] Hunt, S. A. and Benford, R. D. (1994). Identity Talk in the Peace and Justice Movement. *Journal of Contemporary Ethnography*, 22(4):488–517.

[101] Iorio, J. (2015). Vernacular literacy: Orthography and literacy practices. In *The Routledge Handbook of Language and Digital Communication*, pages 166–179. Routledge.

[102] Isaac, M. (2014). Reddit Execs Ellen Pao and Jena Donlin Get Serious About the Site's Business (Q&A). *Vox*.

[103] Jackman, R. W. (1985). Cross-National Statistical Research and the Study of Comparative Politics. *American Journal of Political Science*, 29(1):161–182.

[104] Jasser, G., McSwiney, J., Pertwee, E., and Zannettou, S. (2023). 'Welcome to #GabFam': Far-right virtual community on Gab. *New Media & Society*, 25(7):1728–1745.

[105] Jenkins, R. (2014). *Social Identity*. Routledge.

[106] Jenson, J. and Saint-Martin, D. (2003). New routes to social cohesion? Citizenship and the social investment state. *Canadian Journal of Sociology/Cahiers canadiens de sociologie*, pages 77–99.

[107] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification.

[108] Kalyvas, S. N. (2006). *The Logic of Violence in Civil War*. Cambridge University Press.

[109] Karpf, D. (2012). Social Science Research Methods in Internet Time. *Information, Communication & Society*, 15(5):639–661.

[110] Kelly, A. (2020). Opinion | Mothers for QAnon. *The New York Times*.

[111] Kelly, A. (2023). Alpha and nerd masculinities: Antifeminism in the digital sphere. *Patriarchy in Practice: Ethnographies of Everyday Masculinities*, page 25.

[112] Kim, E. E. and Toole, B. A. (1999). Ada and the first computer. *Scientific American*, 280(5):76–81.

[113] King, M. and Taylor, D. M. (2011). The radicalization of homegrown jihadists: A review of theoretical models and social psychological evidence. *Terrorism and political violence*, 23(4):602–622.

[114] Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. 131:1598.

[115] Kollman, K., Miller, J. H., and Page, S. E. (1997). Political institutions and sorting in a Tiebout model. *The American Economic Review*, pages 977–992.

[116] Konishi, H. (1996). Voting with ballots and feet: Existence of equilibrium in a local public good economy. *Journal of Economic Theory*, 68(2):480–509.

[117] Krasner, S. D. and Affairs, H. U. C. f. I. (1978). *Defending the National Interest: Raw Materials Investments and U.S. Foreign Policy*. Princeton University Press.

[118] Kravari, K. and Bassiliades, N. (2015). A survey of agent platforms. *Journal of Artificial Societies and Social Simulation*, 18(1):11.

[119] Lachowska, M., Mas, A., and Woodbury, S. A. (2020). Sources of displaced workers' long-term earnings losses. *American Economic Review*, 110(10):3231–3266.

[120] Lake, D. A. (2007). The State and International Relations.

[121] Lang, C. and Pearson-Merkowitz, S. (2015). Partisan sorting in the United States, 1972–2012: New evidence from a dynamic analysis. *Political Geography*, 48:119–129.

[122] Larson, J. M. (2017). Why the West Became Wild: Informal Governance with Incomplete Networks. *World Politics*, 69(4):713–749.

[123] Larson, J. M. and Lewis, J. I. (2017). Ethnic networks. *American Journal of Political Science*, 61(2):350–364.

[124] Larson, J. M., Nagler, J., Ronen, J., and Tucker, J. A. (2019). Social networks and protest participation: Evidence from 130 million Twitter users. *American Journal of Political Science*, 63(3):690–705.

[125] Lavin, T. (2020). *Culture Warlords: My Journey into the Dark Web of White Supremacy*. Legacy Lit.

[126] Lipset, S. M. (1959). Some Social Requisites of Democracy: Economic Development and Political Legitimacy. *American Political Science Review*, 53(1):69–105.

[127] Literat, I. and Kligler-Vilenchik, N. (2019). Youth collective political expression on social media: The role of affordances and memetic dimensions for voicing political views. *New media & society*, 21(9):1988–2009.

[128] Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4):587–604.

[129] Longyear, J. Q. and Parsons, T. D. (1972). The friendship theorem. *Indagationes Mathematicae (Proceedings)*, 75(3):257–262.

[130] Loyle, C. E., Cunningham, K. G., Huang, R., and Jung, D. F. (2023). New Directions in Rebel Governance Research. *Perspectives on Politics*, 21(1):264–276.

[131] Loyle, C. E. and Simoni, A. (2017). Researching Under Fire: Political Science and Researcher Trauma. *PS: Political Science & Politics*, 50(1):141–145.

[132] Luccioni, A. S. and Viviano, J. D. (2021). What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus.

[133] Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., and Sanbonmatsu, L. (2013). Long-term neighborhood effects on low-income families: Evidence from Moving to Opportunity. *American economic review*, 103(3):226–231.

[134] MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., and Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

[135] Magaloni, B., Díaz-Cayeros, A., and Ruiz Euler, A. (2019). Public Good Provision and Traditional Governance in Indigenous Communities in Oaxaca, Mexico. *Comparative Political Studies*, 52(12):1841–1880.

[136] Marc-André, A. (2023). *QAnon: A Survey of the Evolution of the Movement from Conspiracy Theory to New Religious Movement.* PhD thesis, Concordia University.

[137] Maskaliūnaitė, A. et al. (2015). Exploring the theories of radicalization. *International Studies: Interdisciplinary Political and Cultural Journal (IS)*, 17(1):9–26.

[138] Mason, L. (2015). "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American journal of political science*, 59(1):128–145.

[139] Masser, B. and Abrams, D. (1999). Contemporary sexism: The relationships among hostility, benevolence, and neosexism. *Psychology of women quarterly*, 23(3):503–517.

[140] McAdam, D. (1983). Tactical Innovation and the Pace of Insurgency. *American Sociological Review*, 48(6):735–754.

[141] McCarthy, J. D. and Zald, M. N. (1977a). Resource mobilization and social movements: A partial theory. *American journal of sociology*, 82(6):1212–1241.

[142] McCarthy, J. D. and Zald, M. N. (1977b). The trend of social movements in America: Professionalization and resource mobilization.

[143] Melucci, A. (2013). The process of collective identity. In *Social Movements and Culture*, pages 41–63. Routledge.

[144] Meta, Inc. (2020). Form 10-Q. Technical report, U.S. Securities and Exchange Commission.

[145] Meta, Inc. (2022). Form 10-Q. Technical report, U.S. Securities and Exchange Commission.

[146] Newhouse, A. (2021). The threat is the network: The multi-node structure of neo-fascist accelerationism. *CTC Sentinel*, 14(5):17–25.

[News] News, NBC. Senator asks how Facebook remains free, Zuckerberg smirks: 'We run ads'. *NBC News*.

[148] Newton, C. (2019). Three Facebook moderators break their NDAs to expose a company in crisis.

[149] Norris, J. J. (2020). Idiosyncratic Terrorism: Disaggregating an Undertheorized Concept. *Perspectives on Terrorism*, 14(3):2–18.

[150] Norris, P. and Inglehart, R. (2019). *Cultural Backlash: Trump, Brexit, and Authoritarian Populism.* Cambridge University Press.

[151] Noveck, B. S. (2009). *Wiki Government: How Technology Can Make Government Better, Democracy Stronger, and Citizens More Powerful.* Brookings Institution Press.

[152] Olson, M. (1993). Dictatorship, democracy, and development. *American political science review*, 87(3):567–576.

[153] Ozalp, S., Williams, M. L., Burnap, P., Liu, H., and Mostafa, M. (2020). Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society*, 6(2):2056305120916850.

[154] Pan, J., Shao, Z., and Xu, Y. (2022). How government-controlled media shifts policy attitudes through framing. *Political Science Research and Methods*, 10(2):317–332.

[155] Pan, J. and Siegel, A. A. (2020). How Saudi crackdowns fail to silence online dissent. *American Political Science Review*, 114(1):109–125.

[156] Pearlman, W. (2011). *Violence, Nonviolence, and the Palestinian National Movement*. Cambridge University Press.

[157] Pearson, R. (2010). Fandom in the digital era. *Popular communication*, 8(1):84–95.

[158] Peck, R. (2023). The Hate-Fueled Rise of r/The_Donald—and Its Epic Takedown. *Wired*.

[159] Pepinsky, T. B., Pierskalla, J. H., and Sacks, A. (2017). Bureaucracy and Service Delivery. *Annual Review of Political Science*, 20(1):249–268.

[160] Pierre, J. and Peters, B. G. (2017). The shirking bureaucrat: A theory in search of evidence? *Policy & Politics*, 45(2):157–172.

[161] Preece, J. and Maloney-Krichmar, D. (2005). Online communities: Design, theory, and practice. *Journal of computer-mediated communication*, 10(4):JCMC10410.

[162] Rabinow, P. and Sullivan, W. M. (1987). *Interpretive Social Science: A Second Look*. Univ of California Press.

[163] Reny, T. T. and Newman, B. J. (2021). The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests. *American political science review*, 115(4):1499–1507.

[164] Rheingold, H. (1993). A slice of life in my virtual community. *Global networks: Computers and international communication*, pages 57–80.

[165] Roberts, S. T. (2017). *Content Moderation*.

[166] Rose-Ackerman, S. (1979). Market models of local government: Exit, voting, and the land market. *Journal of Urban Economics*, 6(3):319–337.

[167] Sartori, G. (1970). Concept Misformation in Comparative Politics. *American Political Science Review*, 64(4):1033–1053.

[168] Scherer, M., Davis, M., Momoi, K., Tong, D., Kida, Y., and Edberg, P. (2009). Proposal for encoding emoji symbols.

[169] Schulz, P., Kreft, A.-K., Touquet, H., and Martin, S. (2022). Self-care for gender-based violence researchers – Beyond bubble baths and chocolate pralines. *Qualitative Research*, page 14687941221087868.

[170] Schulze, H., Hohner, J., Greipl, S., Girgnhuber, M., Desta, I., and Rieger, D. (2022). Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on Telegram. *Convergence: The International Journal of Research into New Media Technologies*, 28(4):1103–1126.

[171] Scott, A. C. (2007). *The Nonlinear Universe: Chaos, Emergence, Life*. Springer Science & Business Media.

[172] Security, C. o. H. and Affairs, G. (2020). *Threats to the Homeland*.

[173] Serpe, R. T. (1987). Stability and change in self: A structural symbolic interactionist explanation. *Social Psychology Quarterly*, pages 44–55.

[174] Smith, D. A., Solinger, D. J., and Topik, S. C. (1999). *States and Sovereignty in the Global Economy*. Routledge.

[175] Snow, D. (2001). Collective identity and expressive forms. *University of California, Irvine eScholarship Repository*.

[176] Sommer, W. (2019). Reddit 'Quarantines' Pro-Trump Forum Over Anti-Police Threats. *The Daily Beast*.

[177] Steen, E., Yurechko, K., and Klug, D. (2023). You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on TikTok. *Social Media+ Society*, 9(3):20563051231194586.

[178] Steinert-Threlkeld, Z. C. (2017). Spontaneous collective action: Peripheral mobilization during the Arab Spring. *American Political Science Review*, 111(2):379–403.

[179] Stets, J. E. and Burke, P. J. (2000). Identity theory and social identity theory. *Social psychology quarterly*, pages 224–237.

[180] Stets, J. E. and Burke, P. J. (2014). Social comparison in identity theory. *Communal functions of social comparison*, pages 39–59.

[181] Stever, G. S. (2009). Parasocial and social interaction with celebrities: Classification of media fans. *Journal of Media Psychology*, 14(3):1–39.

[182] Strandberg, K. and Berg, J. (2015). Impact of temporality and identifiability in online deliberations on discussion quality: An experimental study. *Javnost-The Public*, 22(2):164–180.

[183] Stromer-Galley, J. and Muhlberger, P. (2009). Agreement and disagreement in group deliberation: Effects on deliberation satisfaction, future engagement, and decision legitimacy. *Political communication*, 26(2):173–192.

[184] Stryker, S. (1968). Identity salience and role performance: The relevance of symbolic interaction theory for family research. *Journal of Marriage and the Family*, pages 558–564.

[185] Stryker, S. (2004). Integrating emotion into identity theory. In *Theory and Research on Human Emotions*, pages 1–23. Emerald Group Publishing Limited.

[186] Stryker, S. and Serpe, R. T. (1994). Identity salience and psychological centrality: Equivalent, overlapping, or complementary concepts? *Social psychology quarterly*, pages 16–35.

[187] Stuart, K. and @keefstuart (2014). Zoe Quinn: 'All Gamergate has done is ruin people's lives'. *The Observer*.

[188] Sunstein, C. (2003). *Republic. Com 2.0*. Princeton, NJ: Princeton University Press.

[189] Tarrow, S. (1996). Social Movements in Contentious Politics: A Review Article. *American Political Science Review*, 90(4):874–883.

[190] Tarrow, S. (2012). *Strangers at the Gates: Movements and States in Contentious Politics*. Cambridge University Press.

[191] Tausanovitch, C. and Warshaw, C. (2018). Does the Ideological Proximity Between Candidates and Voters Affect Voting in U.S. House Elections? *Political Behavior*, 40(1):223–245.

[192] Terpstra, N. and Frerks, G. (2017). Rebel Governance and Legitimacy: Understanding the Impact of Rebel Legitimation on Civilian Compliance with the LTTE Rule. *Civil Wars*, 19(3):279–307.

[193] Thelwall, M. and Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology*, 64(8):1608–1617.

[194] Thurman, N., Moeller, J., Helberger, N., and Trilling, D. (2019). My Friends, Editors, Algorithms, and I. *Digital Journalism*, 7(4):447–469.

[195] Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of political economy*, 64(5):416–424.

[196] Toma, E. F. and Toma, M. (1992). Tax Collection with Agency Costs: Private Contracting or Government Bureaucrats? *Economica*, 59(233):107–120.

[197] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models.

[198] Tufekci, Z. (2013). "Not This One": Social Movements, the Attention Economy, and Microcelebrity Networked Activism. *American Behavioral Scientist*, 57(7):848–870.

[199] Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 505–514.

[200] Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press.

[201] Twitter, Inc. (2020). Form 10-Q. Technical report, U.S. Securities and Exchange Commission.

[202] Van Zoonen, L. (2005). *Entertaining the Citizen: When Politics and Popular Culture Converge.* Rowman & Littlefield.

[203] Vergani, M., Iqbal, M., Ilbahar, E., and Barton, G. (2020). The three Ps of radicalization: Push, pull and personal. A systematic scoping review of the scientific evidence about radicalization into violent extremism. *Studies in Conflict & Terrorism*, 43(10):854–854.

[204] Vromen, A. (2008). Building virtual spaces: Young people, participation and the Internet. *Australian Political Studies Association*, 43(1):79–97.

[205] Walter, B. F. (2023). *How Civil Wars Start: And How to Stop Them.* Crown.

[206] Walther, S. and McCoy, A. (2021). US extremism on telegram. *Perspectives on Terrorism*, 15(2):100–124.

[207] Waltz, K. N. (1997). Evaluating Theories. *American Political Science Review*, 91(4):913–917.

[208] Webster, J. G. (2014). *The Marketplace of Attention: How Audiences Take Shape in a Digital Age.* Mit Press.

[209] Weingast, B. R., Shepsle, K. A., and Johnsen, C. (1981). The political economy of benefits and costs: A neoclassical approach to distributive politics. *Journal of political Economy*, 89(4):642–664.

[210] Weinstein, J. M. (2005). Resources and the information problem in rebel recruitment. *Journal of Conflict Resolution*, 49(4):598–624.

[211] Wittgenstein, L. (1968). II: Notes for lectures on" Private Experience" and" Sense Data". *The Philosophical Review*, 77(3):275–320.

[212] Wood, R. M. (2014). Opportunities to kill or incentives for restraint? Rebel capabilities, the origins of support, and civilian victimization in civil war. *Conflict Management and Peace Science*, 31(5):461–480.

[213] Wooders, M. H. (1994). Large games and economies with effective small groups. In *Game-Theoretic Methods in General Equilibrium Analysis*, pages 145–206. Springer.

[214] Yan, Q. and Yang, F. (2021). From parasocial to parakin: Co-creating idols on social media. *New Media & Society*, 23(9):2593–2615.

[215] Zhirkov, K. (2014). Nativist but not alienated: A comparative perspective on the radical right vote in Western Europe. *Party Politics*, 20(2):286–296.