Community Competition and Political Extremism

Colin Henry, PhD[1]

[1]Department of Political Science, University of Zurich.

Contributing authors: ;

Abstract

Online extremist movements, although appearing monolithic from the outside, are composed of competing ideologies and strategies. Commitment to and promotion of extremist violence varies widely between the communities that make up the wider movement. What explains this variation? I argue that community-level extremism is driven by competition between online communities for attention and engagement on social media platforms. To support this argument, I construct a theoretical framework for understanding social media platforms as sites of political contestation and distribution of public goods. I gather two novel datasets, using an overlapping snowball chain sampling algorithm and transformer-based classifier to capture community competition and extremist content. I show that inter-movement competition between communities drives the share of extremism expressed in communities, as well as the level of out-group-focused extremism.

Keywords: violence, network analysis, social movements, right-wing extremism, large language model

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 (Davey et al., 2021). 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" (Gartenstein-Ross et al., 2023), "idiosyncratic terrorism" (Norris, 2020), or, in the words of FBI Director Chris Wray, "salad bar extremism" (on Homeland Security and Affairs, 2020).

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 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

[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 (for Economics and Peace, 2022). Therefore, this paper focuses primarily on ideologically-motivated extremism.

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.

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 (Berger, 2018). 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 (Hunt and Benford, 1994; Melucci, 2013; Snow, 2001). 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 (King and Taylor, 2011; Bastug et al., 2020). 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 (Maskaliūnaitė et al., 2015). 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 (Vergani et al., 2020). 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 (Becker, 1968; Dell et al., 2019; Baysan et al., 2019). 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, Zhirkov (2014) or Norris and Inglehart (2019)).

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 (Hiltz, 1985; Rheingold, 1993). 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 (Stryker, 2004), groups, and personal relationships (Stets and Burke, 2014). The salience of a community identity may vary according to situations and interactions (Stryker, 1968; Serpe, 1987). Communities may grow or shrink in response to external events that provoke greater salience. Community identities also vary in centrality (Stryker and Serpe, 1994) or prominence (Gecas, 1982) 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 (Preece and Maloney-Krichmar, 2005). 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.

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 (McCarthy and Zald, 1977a) have mostly focused on social movement organizations as the central component of movements (Tarrow, 1996; Della Porta and Diani, 1999; Burstein and Linton, 2002). In this work, social movements are composed of hierarchical organizations that collaborate and compete for scarce resources (Diani, 1992). 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 (Della Porta, 2013).

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 (Crenshaw, 1985). Or to differentiate themselves through violent tactics and progress towards goals (McCarthy and Zald, 1977b). 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 (Weinstein, 2005). 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 (Kalyvas, 2006; Wood, 2014; Dorff et al., 2023).

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 (Tufekci, 2013). 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, Gitlin (2003) or Benford and Snow (2000)). 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" (Han, 2017), 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.*

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 (Scott, 2007). 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, Amarasingam and Argentino (2020)). 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) (Dep, 2023). 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% (Marc-André, 2023). 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 (Kelly, 2020). 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 (Forberg, 2022). 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 (Argentino, 2022), 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 (Masser and Abrams, 1999). Online this movement expresses itself mostly as antifeminism, a reactionary and hostile response to gendered out-groups entering male spaces (Kelly, 2023). 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 (Kim and Toole, 1999). 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 (Stuart and @keefstuart, 2014). 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 (Hoffman et al., 2020); the so-called "red pill" communities, a classic neo-sexist identity that claims to liberate men from the misandry of modern society (Ging, 2019); 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 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.

4.1 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 (Oba, 2012). 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 (Tufekci, 2014).

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 (Baumgartner, 2019). 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 (Gulati and Abulaish, 2019). 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 (Longyear and Parsons, 1972). '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.

4.1.2 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 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 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.

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 1 Community Dataset, Qanon

See, for example, the community overlap ratio for YouTube in Table 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 2 Community Dataset, Neo-sexism

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, Fortuna and Nunes (2018) or MacAvaney et al. (2019)) 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., Thelwall and Buckley (2013)).

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 (Lombard et al., 2002). 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, Dorff et al. (2023) 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 (Hiaeshutter-Rice and Hawkins, 2022). Some of these linguistic shifts are so-called "algospeak"—changes in language used to evade algorithmic content moderation systems (Steen et al., 2023). 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))) (Ozalp et al., 2020). 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, Newton (2019) or Biddle (2020)). 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 (Loyle and Simoni, 2017). 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 (Schulz et al., 2022). 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 (Joulin et al., 2016) or even larger pre-trained transformer-based language models like BERT (Devlin et al., 2019) 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 (Edwards et al., 2020). 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 (Touvron et al., 2023). 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 (Luccioni and Viviano, 2021).

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 (Cohen, 2013). 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 (Guest et al., 2021). 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 (Fariss, 2014). 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 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 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.

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

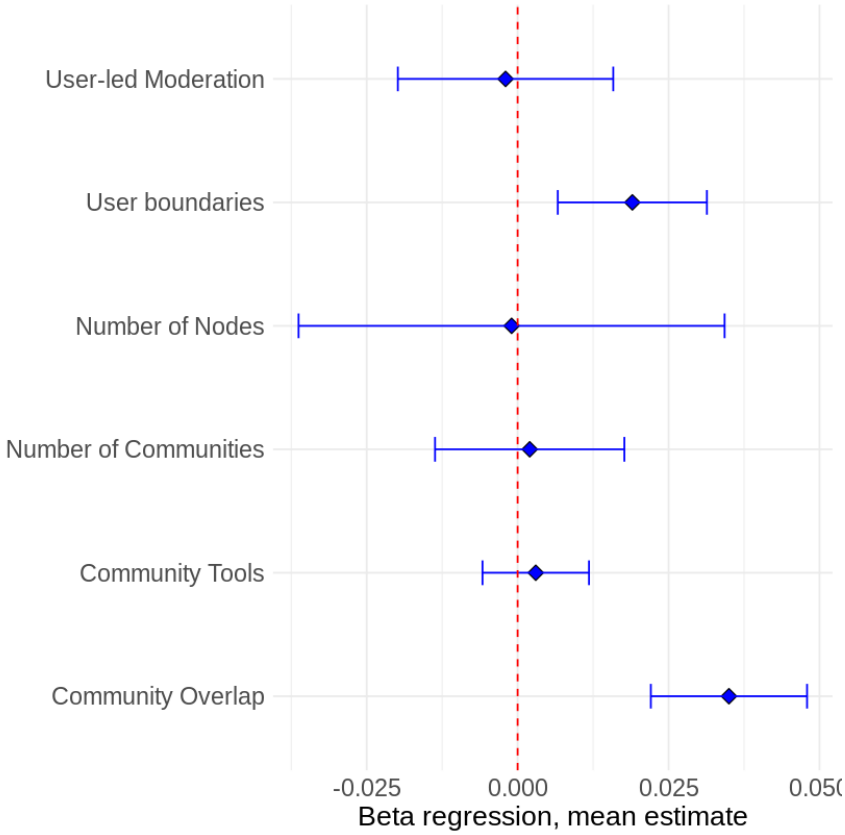

Fig. 1 Model 1, overall share of extremism

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 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 (Walther and McCoy, 2021). 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

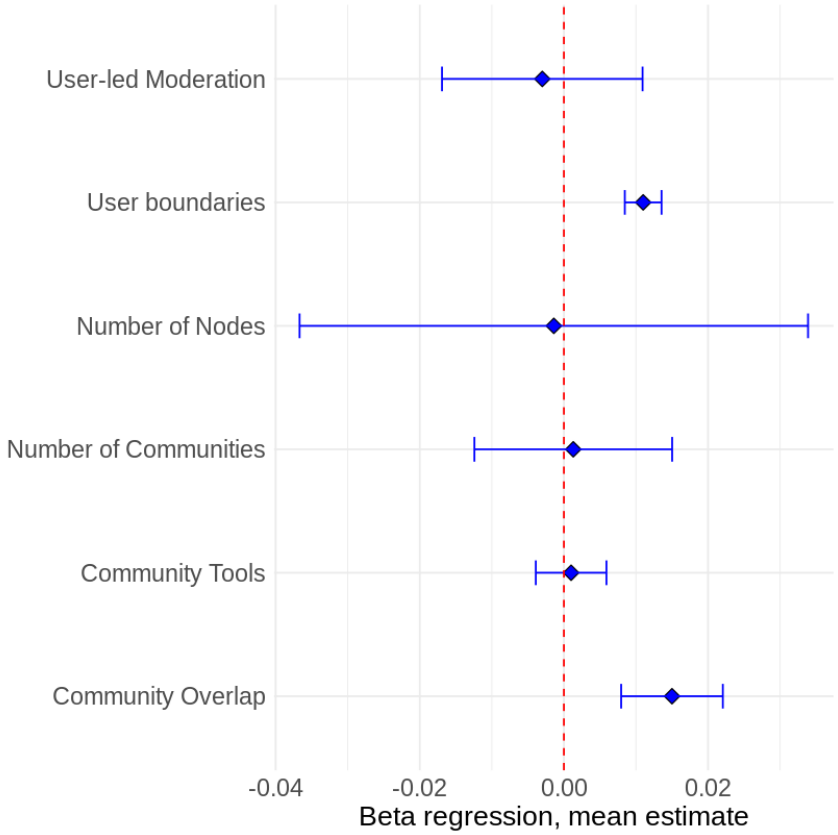

Fig. 2 Model 2, share of out-group-focused extremism

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.

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 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 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

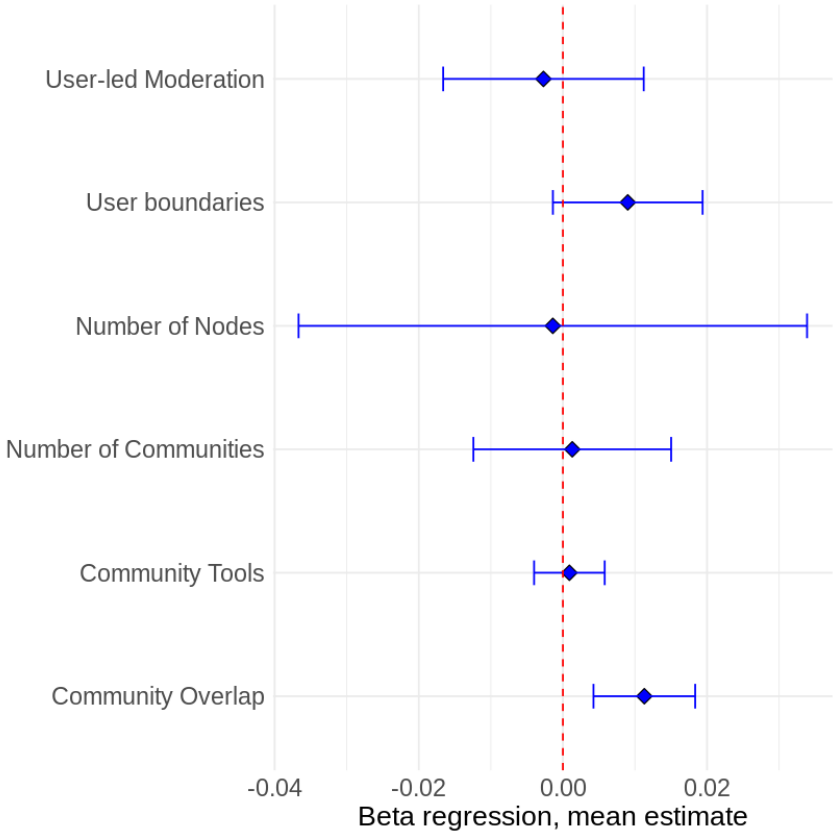

Fig. 3 Model 3, share of in-group-focused extremism

observations have roughly the same among of in-group-focused and out-group-focused extremism overall. 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, Schulze et al. (2022)). 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.

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.

References

2012, October. Obama Networks on BlackPlanet.com. https://web.archive.org/web/20121011172224/http://voices.washingtonpost.com/44/2007/10/obama-networks-on-blackplanetc.html.

2023, September. Departmental Personnel Security FAQs. https://www.energy.gov/ehss/departmental-personnel-security-faqs.

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

Argentino, M.A. 2022. Qvangelicalism: QAnon as a Hyper-Real Religion, *Religious Dimensions of Conspiracy Theories*. Routledge.

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

Baumgartner, J.M. 2019. Pushshift Reddit API. GitHub.

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

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

Benford, R.D. and D.A. Snow. 2000, August. Framing Processes and Social Movements: An Overview and Assessment. *Annual Review of Sociology 26*(1): 611–639. https://doi.org/10.1146/annurev.soc.26.1.611 .

Berger, J.M. 2018. *Extremism*. Mit Press.

Biddle, S. 2020, June. Weeks After PTSD Settlement, Facebook Moderators Ordered to Spend More Time Viewing Online Child Abuse. https://theintercept.com/2020/06/18/facebook-moderator-ptsd-settlement-accenture/.

Burstein, P. and A. Linton. 2002, December. 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. https://doi.org/10.1353/sof.2003.0004 .

Cohen, D.K. 2013, August. Explaining Rape during Civil War: Cross-National Evidence (1980–2009). *American Political Science Review 107*(3): 461–477. https://doi.org/10.1017/S0003055413000221 .

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

Davey, J., M. Comerford, J. Guhl, W. Baldet, and C. Colliver. 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*: 78 .

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

Della Porta, D. 2013. *Clandestine Political Violence*. Cambridge University Press.

Della Porta, D. and M. Diani. 1999. Social movements. *European Studies*: 365 .

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

Diani, M. 1992. The concept of social movement. *The Sociological Review 40*(1): 1–25. https://doi.org/10.1111/j.1467-954X.1992.tb02943.x .

Dorff, C., M. Gallop, and S. Minhas. 2023, April. Network Competition and Civilian Targeting during Civil Conflict. *British Journal of Political Science 53*(2): 441–459. https://doi.org/10.1017/S0007123422000321 .

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

Edwards, A., J. Camacho-Collados, H. De Ribaupierre, and A. Preece 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*, Barcelona, Spain (Online), pp. 5522–5529. International Committee on Computational Linguistics.

Fariss, C.J. 2014, May. Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review 108*(2): 297–318. https://doi.org/10.1017/S0003055414000070 .

for Economics, I. and Peace 2022. Global Terrorism Index 2022: Measuring the Impact of Terrorism. Technical report.

Forberg, P.L. 2022, June. 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. https://doi.org/10.1177/08912416211040560 .

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

Gartenstein-Ross, D., A. Zammit, E. Chace-Donahue, and M. Urban. 2023, March. Composite Violent Extremism: Conceptualizing Attackers Who Increasingly Challenge Traditional Categories of Terrorism. *Studies in Conflict & Terrorism 0*(0): 1–27. https://doi.org/10.1080/1057610X.2023.2194133 .

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

Ging, D. 2019, October. Alphas, Betas, and Incels: Theorizing the Masculinities of the Manosphere. *Men and Masculinities 22*(4): 638–657. https://doi.org/10.1177/1097184X17706401 .

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

Guest, E., B. Vidgen, A. Mittos, N. Sastry, G. Tyson, and H. Margetts 2021, April. 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*, Online, pp. 1336–1350. Association for Computational Linguistics.

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

Han, B.C. 2017. *In the Swarm: Digital Prospects*, Volume 3. MIT press.

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

Hiltz, S.R. 1985. *Online Communities: A Case Study of the Office of the Future*, Volume 2. Intellect Books.

Hoffman, B., J. Ware, and E. Shapiro. 2020, July. Assessing the Threat of Incel Violence. *Studies in Conflict & Terrorism 43*(7): 565–587. https://doi.org/10.1080/1057610X.2020.1751459 .

Hunt, S.A. and R.D. Benford. 1994, January. Identity Talk in the Peace and Justice Movement. *Journal of Contemporary Ethnography 22*(4): 488–517. https://doi.org/10.1177/089124194022004004 .

Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2016, August. Bag of Tricks for Efficient Text Classification.

Kalyvas, S.N. 2006. *The Logic of Violence in Civil War.* Cambridge University Press.

Kelly, A. 2020, September. Opinion — Mothers for QAnon. *The New York Times* .

Kelly, A. 2023. Alpha and nerd masculinities: Antifeminism in the digital sphere. *Patriarchy in Practice: Ethnographies of Everyday Masculinities*: 25 .

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

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

Lombard, M., J. Snyder-Duch, and C.C. Bracken. 2002. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research 28*(4): 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x .

Longyear, J.Q. and T.D. Parsons. 1972, January. The friendship theorem. *Indagationes Mathematicae (Proceedings) 75*(3): 257–262. https://doi.org/10.1016/1385-7258(72)90063-7 .

Loyle, C.E. and A. Simoni. 2017, January. Researching Under Fire: Political Science and Researcher Trauma. *PS: Political Science & Politics 50*(1): 141–145. https://doi.org/10.1017/S1049096516002328 .

Luccioni, A.S. and J.D. Viviano. 2021, May. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. https://arxiv.org/abs/2105.02732v3.

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

Marc-André, A. 2023. *QAnon: A Survey of the Evolution of the Movement from Conspiracy Theory to New Religious Movement*. Ph. D. thesis, Concordia University.

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

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

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

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

Melucci, A. 2013. The process of collective identity, *Social Movements and Culture*, 41–63. Routledge.

Newton, C. 2019, June. Three Facebook moderators break their NDAs to expose a company in crisis. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-tampa.

Norris, J.J. 2020. Idiosyncratic Terrorism: Disaggregating an Undertheorized Concept. *Perspectives on Terrorism 14*(3): 2–18. 26918296 .

Norris, P. and R. Inglehart. 2019. *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press.

on Homeland Security, C. and G. Affairs. 2020, September. *Threats to the Homeland*.

Ozalp, S., M.L. Williams, P. Burnap, H. Liu, and M. Mostafa. 2020, April. Antisemitism on Twitter: Collective Efficacy and the Role of Community Organisations in Challenging Online Hate Speech. *Social Media + Society 6*(2): 2056305120916850. https://doi.org/10.1177/2056305120916850 .

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

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

Schulz, P., A.K. Kreft, H. Touquet, and S. Martin. 2022, April. Self-care for gender-based violence researchers – Beyond bubble baths and chocolate pralines. *Qualitative Research*: 14687941221087868. https://doi.org/10.1177/14687941221087868 .

Schulze, H., J. Hohner, S. Greipl, M. Girgnhuber, I. Desta, and D. Rieger. 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 .

Scott, A.C. 2007, October. *The Nonlinear Universe: Chaos, Emergence, Life*. Springer Science & Business Media.

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

Snow, D. 2001. Collective identity and expressive forms. *University of California, Irvine eScholarship Repository* .

Steen, E., K. Yurechko, and D. Klug. 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 .

Stets, J.E. and P.J. Burke. 2014. Social comparison in identity theory. *Communal functions of social comparison*: 39–59 .

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

Stryker, S. 2004. Integrating emotion into identity theory, *Theory and Research on Human Emotions*, 1–23. Emerald Group Publishing Limited.

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

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

Tarrow, S. 1996, December. Social Movements in Contentious Politics: A Review Article. *American Political Science Review 90*(4): 874–883. https://doi.org/10.2307/2945851 .

Thelwall, M. and K. Buckley. 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 .

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

Tufekci, Z. 2013, July. "Not This One": Social Movements, the Attention Economy, and Microcelebrity Networked Activism. *American Behavioral Scientist 57*(7): 848–870. https://doi.org/10.1177/0002764213479369 .

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, pp. 505–514.

Vergani, M., M. Iqbal, E. Ilbahar, and G. Barton. 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 .

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

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

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 .

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