Ecologies of Hate: Indiscriminate Targeting of Online Hate Speech

Colin Henry, Sean Long, Rick Sear, Nicholas Johnson, Stefan Wuchty, Neil Johnson, Yonatan Lupu

September 24, 2024

Abstract

Extremist groups use hate speech to direct humiliation, contempt, and violence at vulnerable communities, attracting new recruits and radicalizing existing members. While significant progress has been made by policy makers and researchers alike in identifying and understanding the effects of hate speech, few studies track patterns of hate speech and its targets across mainstream and fringe social media platforms. Leveraging a supervised machine learning analysis of 7 types of online hate speech, we examine the discriminate and indiscriminate nature of online hate speech. Using lessons learned from diversity partitioning schemes in community ecology, we develop a new measure of hate speech type assemblage that takes into account both the discriminate nature of targeted hate speech and the overall abundance of hate speech across platforms and extremist groups. We find that offline events like protests or political violence often lead to both short-term and long-term changes in who is targeted by hate speech. We show that mainstream and fringe platforms exhibit different patterns of discriminate hate speech, and that unrelated communities are often caught up in the "crossfire" of indiscriminate hate speech.

1 Introduction

Online hate speech is a worsening problem across social media platforms, posing significant challenges for policymakers, researchers, and platforms. The proliferation of hate speech online has been linked to a range of adverse outcomes, including radicalization of individuals and communities, the incitement of offline political violence, and targeted violence towards vulnerable communities. Despite the efforts to curb its spread, hate speech continues to thrive in both mainstream social media environments like Facebook and Twitter and on fringe platforms such as 4Chan and Gab. This persistence is partly due to the dynamic and adaptable nature of extremist communities, which constantly evolve their strategies to evade detection and suppression. And the algorithmic structures of social media platforms often inadvertently amplify and direct hateful content, enabling it to reach broad audiences or specific targets. The complexity of this issue is compounded by the varying legal and regulatory frameworks across different states, making it difficult to establish a unified approach to combating online hate speech.

Targets of hate speech may suffer a variety of negative consequences. Exposure to online hate speech directed at a community or identity can drive members of those groups off platforms and away from sites of political contestation, social organizing, or economic activity. Users in targeted groups may also experience significant psychological harms. A constant barrage of hate speech can lead to anxiety, depression, and a diminished sense of safety and belonging. Over time, this can erode the social cohesion and resilience of these communities to state and social repression, as individuals may become isolated and less likely to engage in public discourse or build social ties. Furthermore, the normalization of hate speech in online environments can perpetuate harmful stereotypes and biases, further entrenching discrimination and social divides.

This not only affects the immediate victims but also has broader societal implications. The silencing of marginalized voices diminishes political engagement and efficacy, weakening the kind of broad, networked online social and political movements we observed a decade ago. Understanding these impacts is essential for developing comprehensive strategies to counteract the spread of hate speech and support those who are most affected by it.

Extremist communities also use hate speech to construct group identities and police identity boundaries. Indiscriminate hate speech, which targets broad and often unrelated groups, serves to amplify a sense of in-group solidarity among extremists by creating a common enemy. This broad-spectrum hate speech can lead to widespread fear and hostility, affecting diverse and unsuspecting communities. The indiscriminate nature of such speech can create an atmosphere of pervasive hostility, where no group feels entirely safe from attack. This generalized targeting can foster a climate of fear and division, disrupting trust within and between different communities. Additionally, the spillover effects of indiscriminate hate speech can undermine social cohesion and destabilize interactions between online communities, as groups caught in the crossfire may turn on each other.

Understanding the nuances of how hate speech is propagated, its targets, and its impacts is crucial for developing effective interventions and fostering safer online communities. Exploring when hate speech is indiscriminate or targeted towards one particular group has broad implications for both the dynamics of online extremism and the victims of online hate speech.

We study how offline events influence indiscriminate and discriminate hate speech online throughout the platforms and communities that make up online ecological complexes. We draw a distinction between hate speech that focuses on a single vulnerable group (targeted or discriminate) and speech that directs hate a numerous groups at the same time (indiscriminate). We situate the differences between observations containing many types of hate speech and those with only one or two types of hate speech within a framework of violence borrowed from literatures on human rights and violence against civilians. In this framework, violent actors strategically choose to target specific groups with violence or deploy violence indiscriminately, often as a function of their capabilities, resources, or connections to civilians. Although the use indiscriminate or discriminate hate speech by online extremists is an emergent property rather than a strategic choice, it nonetheless may have disparate impacts on wider support for extremist ideologies and vulnerable online communities that reflect the effects of violent repertoires in wartime.

We explore the connection between offline events and the use of indiscriminate hate speech on both moderated and unmoderated platforms. In particular, we ask the following research questions: does hate speech become more or less indiscriminate in reaction to offline events? Is this reaction different across mainstream and 'alt-tech' platforms? Do some platforms harbor more or less indiscriminate hate speech than others in general? And, what types of hate speech typically mix together on more indiscriminate platforms or after events inspiring more indiscriminate hate?

In this paper, we examine how offline events change the use of discriminate and indiscriminate hate speech in online communities. We use a transformer trained on human-coded data to classify a corpus of X posts from Y social media platforms into 7 different types of hate speech. Treating these types of hate speech as "species," we develop a modified version of the Hill Diversity algorithm to produce measures of discriminate hate speech within a variety of samples. Our data approach has significant advantages over similar work on online hate speech. First, the broad cross-section of the classified dataset allows us to make comparisons across multiple platforms, XX online communities, and YY months, while much of the literature on hate speech narrowly focuses on a few platforms and a limited temporal frame. Second, the transformer-based classification algorithm outperforms previous generation machine learning systems, requiring less human-labeled data and producing better accuracy across an enormous dataset.

2 Previous work

The literature on online hate speech has evolved substantially over the last decade, driven by both the increase in hate speech generally and the its social and political implications. The initial focus of research in this area was primarily on detecting and analyzing hate speech using computational methods, with an orientation towards aiding social media platforms in curbing the spread of objectionable content (see, for example, Burnap and Williams (2016); Ribeiro et al. (2018); Zhang et al. (2018), and Qian et al. (2018)). However, while computational detection models have grown in sophistication over time, the scope of analysis has remaind relatively narrow, often limited to binary classifications of hate speech or the categorization of predefined hate speech types.

This narrow focus is limiting, particularly when it comes to understanding the broader implications of online hate speech and its connection to offline events. The majority of studies have concentrated on mainstream platforms like Twitter or Facebook, neglecting the variations that exist on less moderated, fringe platforms where hate speech can be more extreme or differently structured (Fortuna and Nunes, 2018; Jahan and Oussalah, 2023). Jahan and Oussalah (2023), for example, find that fewer than 20% of surveyed hate speech-focused machine learning articles from across the ACM Digital Library and Google Scholar included more than one platform. This gap in the literature highlights the need for more comprehensive frameworks that go beyond binary classifications to capture the full spectrum of hate speech. These frameworks should consider the volume and variety of hate speech across different platforms, which would provide deeper insights into the nature of this kind of expression and its potential to incite offline violence.

This concentration on a narrow set of platforms and a limited scope of hate speech types presents two significant limitations. First, it overlooks the systematic variations in hate speech that likely occur between moderated mainstream platforms and fringe, less moderated ones. Hate speech on fringe platforms may be more extreme, more pervasive, or differently structured, necessitating separate or adapted methodologies for detection and analysis. Second, there is a pressing need for more comprehensive classification frameworks that go beyond binary categorizations or limited hate speech types. A broader taxonomy that combines categories and considers both volume and variety captures hate speech hybridity, providing deeper insights into how this kind of expression evolves and spreads. This expanded understanding is crucial for developing more effective strategies to counteract online hate speech across diverse digital environments.

Some more recent computational work has attempted to expand the categorization and classification of hate speech in interesting and nuanced ways. HateBERT (Caselli et al., 2020), for example, tries to operationalize the multidimensional framework of Poletto et al. (2021) to include both hate speech categories and the intensity or toxicity of posts. However, while HateBERT performs well classifying posts as abusive, offensive, or hateful, the model works on these dimensions separately and does not attempt to generate combined measures for individual posts. HateXplain (Mathew et al., 2021) combines both hatefulness and the type of community targeted, but offers only limited and un-theorized categories drawn from unsupervised clusters in the examined dataset.

Thus, while the body of work on online hate speech has made significant strides, there remains a critical need for research that not only expands the categorization of hate speech but also explores its connection to offline events. Here, we explore the connection between online responsiveness to offline events and the targets of objectionable speech on the social internet.

3 Targeted hate speech

In the political conflict literature, the distinction between discriminate and indiscriminate targeting of violence is critical to understanding the strategies employed by state and non-state actors in armed conflicts. Discriminate targeting refers to violence directed toward specific individuals or groups based on identifiable and separable characteristics, like ethnicity, political affiliation, religion, or military status. This type of targeting is often calculated to pursue precise political or military strategies, such as eliminating perceived threats or punishing specific communities to deter future opposition. Scholars in the field have argued that discriminate violence can be instrumental in controlling populations by creating fear among targeted groups while potentially avoiding backlash from non-targeted communities. This strategy requires significant intelligence and resources, as it relies on the ability to accurately identify and reach the intended targets.

Similarly, discriminate hate speech is directed towards a specific group or community based on group characteristics[1]. Typically targeted hate speech comes from non-state actors online, although state-run accounts can and have engaged in hateful expression against vulnerable communities. The strategic

[1] Although hate speech can also be targeted towards individuals and campaigns of hate against specific people can be constitute of or provoke discriminate hate speech against the greater community, we do not examine this type of targeting here. Scholarship examining this phenomenon can be found, for example, in work on the Gamergate harassment campaign (Walther, 2022; Ferguson and Glasgow, 2021).

goal of discriminate hate speech varies by actor, and is often an emergent property of disaggregated networks of actors. However, for extremist communities, targeted hate speech can serve as a valuable means of coordinating group activity, establishing and enforcing group norms, and giving individuals on the pathway to radicalization a means of integrating extremist identities (Cervone et al., 2021; Schmid et al., 2024). For targeted communities, hate speech directed at individuals or groups can deteriorate intergroup relations and defray perceptions of targeted communities for audiences exposed to hate speech (Bilewicz and Soral, 2020).

In contrast, indiscriminate targeting involves the use of violence in a manner that does not differentiate between combatants and non-combatants or between members of different groups. Such violence often results in widespread harm to civilians and is typically less concerned with the precise identity of the victims. Indiscriminate targeting is frequently associated with tactics like bombing, shelling, or mass executions, where the aim is to inflict maximum damage or chaos rather than to achieve a specific, targeted outcome. Political scientists have debated the conditions under which actors resort to indiscriminate violence, often linking it to situations where resources are scarce, intelligence is poor, or the actor seeks to terrorize or depopulate entire regions to weaken overall resistance. This approach can lead to severe humanitarian consequences and may provoke international condemnation, but it can also be effective in undermining the will or capability of the opponent.

Indiscriminate or untargeted hate speech mirrors the use of indiscriminate violence. Hate speech directed at a wide variety of groups or communities, either in individual posts or as the effects of the aggregated speech of a group, serves to terrorize or drive off platform users. In this way, indiscriminate hate speech is closer to so-called “toxic speech” that generates a wide field of discursive harm for audiences that encounter it (Tirrell, 2017). For extremist communities, it also serves to regenerate and navigate group boundaries, drawing lines between users who accept, reject, or simply tolerate hateful expressions. Users who experience or consume indiscriminate hate speech often face a choice of deserting online spaces or reducing their indignation towards hateful content (Schmid et al., 2024).

4 Data collection

The data used in this project significantly expands on previous data collection efforts from (Lupu et al., 2023). We now include posts from 2021, 2022, and 2023, and cover a broader sway of platforms: Twitter, YouTube, Rumble, and Bitchute. This update aims to provide a more comprehensive and current view of online hate speech dynamics across both mainstream and fringe social media platforms. Thus, we can make comparisons before and after President Donald Trump's tenure in office, a period shown to be correlated with a general rise in hate speech [cite]. Expanding to more so-called 'alt-tech' spaces also allows us to account for the continued Balkanization of online communities from mainstream platforms to networking products that more closely match users' tolerance or demand for offensive speech.

Our data collection focused on publicly available posts without any user interaction. We made no assumptions about the geographic locations of users, although discussions predominantly featured U.S. politics, with frequent mentions of European and other English-speaking regions' issues. Importantly, no personal identifying information was collected, and user names were anonymized. All data collection adhered to the platforms' terms of use and received advanced approval from the Institutional Review Board (IRB).

Exploring hate communities on these new platforms closely mirrored the process of identifying communities described in Lupu et al. (2023). These communities include channels on YouTube, Rumble, or Bitchute, which share similar design architecture; and following or shared hashtag networks on Twitter. The refined sampling procedure detailed in the supplementary information (SI). As before, our team manually searched the updated list of platforms for hate communities, examining their content to identify new hate speech spaces. A community was classified as a hate community if at least two of the twenty most recent posts contained hate speech. For this study, hate speech is defined according to two main criteria:

- Content falling under the United States Code provisions regarding hate crimes or hate speech, as per Department of Justice guidelines.

- Content supporting or promoting fascist ideologies or regime types, such as extreme nationalism and racial identitarianism.

The determination of whether an online community qualifies as a hate community was made manually based on these criteria by subject matter experts on the team. Table 1 below shows the total number of collected posts per platform across the whole dataset.

Platform	**Posts**	**Platform**	**Posts**
4Chan	142,610,212	YouTube	7,433,296
Telegram	3,847,433	Rumble	3,481,579
Gab	1,372,781	Bitchute	6,245,710
Twitter	518,294	Facebook	1,236,016
VK	133,992	Instagram	7,550

Table 1: Total observations per platform, 2019 - 2023

5 Measurement

5.1 Hate speech

Classifying such a large dataset requires the use of automated, computational methods. Our approach improves upon the supervised machine learning classifier used in previous work [cite PLOS]. Extending the trained BERT model [cite Devlin et al], we classify an additional X number of posts. Accuracy from this model matched the previous model in [cite PLOS], ranging from 90% to 97%, depending on the type of hate speech. As before, we validated the machine results using human annotators and found that the results were highly reliable.

Our study categorizes hate speech into seven types: race, gender, religion, gender identity/sexual orientation (GI/SO), immigration, ethnicity/identitarian/nationalism (E/I/N), and anti-Semitism. These categories were chosen based on their prevalence and distinguishability in our manual review of hate communities. Importantly, these categories often co-occur, and in patterns important to understanding discriminate or indiscriminate hate speech. Observations that contain, to choose an extreme example, all seven types of hate speech are especially "indiscriminate"–directing hate speech across many different protected categories. Conversely, of course, observations that contain only one type of hate speech target discriminate hate at that group specifically.

Thus, observations can be classified simply as one of the 7 base level categories, or as one of 120 complex categories covering all possible combinations of hate speech. Figure 1 below shows the percentage distribution of posts occupying these complex categories. Extending our classification scheme uncovers important patterns: the data contains a greater percentage of posts with both racist and E/I/N hate speech than base categories like gender, religion, or immigration, for example. The base category of anti-immigrant hate speech occupies a smaller proportion of hateful posts than less discriminate categories such as

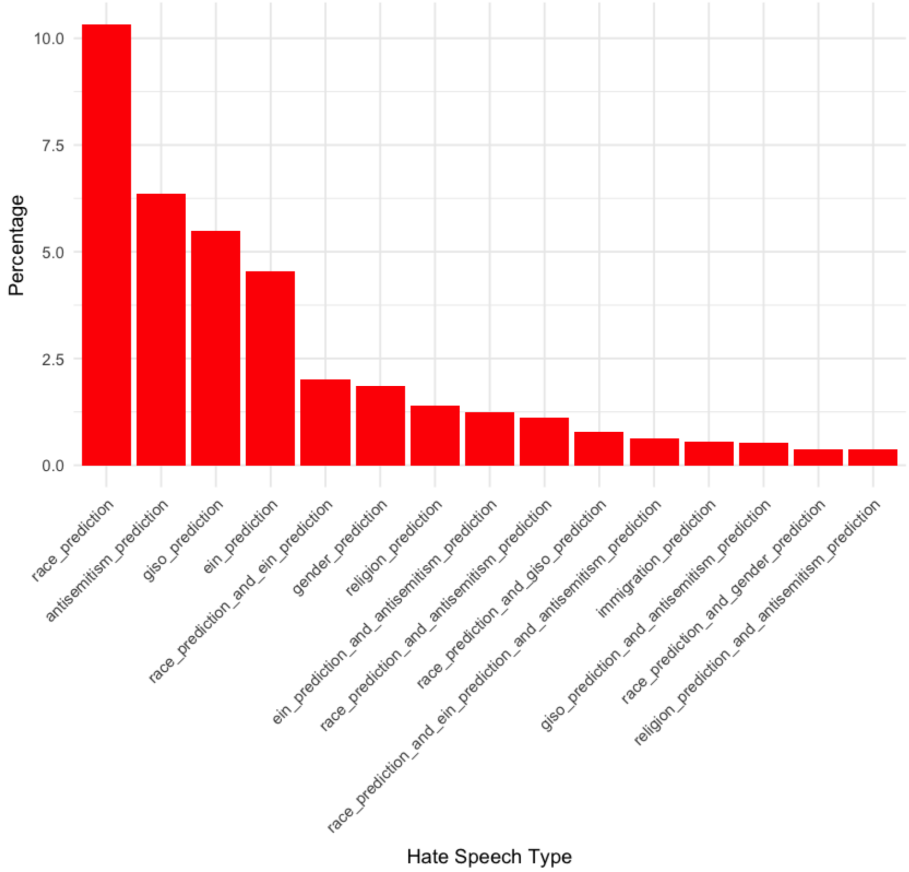


Figure 1: Percentage of Posts Containing Each Type of Hate Speech (Top 15)

racial, E/I/N/, and antisemitic combination posts.

This approach provides us with two advantages. First, compound categories allow us to disaggregate the targets of hate speech. Users or communities with a greater proportion of basic hate speech categories or a higher percentage of one particular basic category can be thought of as particularly discriminatory, targeting one vulnerable population above others. On the other hand, communities with more compound categories are using hate speech in a more indiscriminate manner. The second advantage it gives us is leverage is computing an index of discriminate or indiscriminate targeting. An index that can incorporate both the total volume of all hate speech on the platform as well as the percentages of each basic and compound category will ease interpretation of discriminatory patterns.

5.2 Discriminate and indiscriminate targeting

To compute the level of discriminate and indiscriminate hate speech across a particular sampling frame, we use an algorithmic implementation of the Hill Diversity Index, a measure of diversity that captures the richness and evenness of species within a community. It is part of a family of diversity indices that incorporate different weighting factors to account for the abundance and distribution of species. The Hill Index, also known as the Hill number, is generally defined as:

$$^qD = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}}$$

where:

- ^qD is the Hill number of order q,
- S is the total number of species (or categories),
- p_i is the proportion of individuals (or posts) in the i-th species (or category),
- q is a parameter that determines the sensitivity to species (or category) volume.

For $q = 1$, the formula is undefined due to the division by zero, but in this case, it is typically interpreted as:

$$^{1}D = \exp\left(-\sum_{i=1}^{S} p_i \ln p_i\right)$$

which is the exponential of the Shannon entropy.

A Hill number, denoted as ^{q}D, varies depending on the parameter q, which adjusts the sensitivity of the index to the relative abundances of species. When $q = 0$, the index is equivalent to species richness, counting all species equally regardless of their abundance. When $q = 1$, the index corresponds to the exponential of the Shannon entropy, giving more weight to rare species. As q increases, the index becomes more sensitive to the abundance of common species, with $q = 2$ being equivalent to the inverse of the Simpson index. This flexibility allows us to tailor the Hill Diversity Index to different online ecologies and data, providing a nuanced picture of community diversity.

To be sure, this implementation of Hill numbers is much less complex than those proposed in Chao [2010] or Chio [2014], for example. The measures of diversity, similarity, and differentiation these authors propose as basic tools of ecological analyses seek to explore both functional *and* phylogenetic differences between "species." In other words, this more advanced measure of Hill diversity takes into account the genetic history of unit differences, breaking apart the assumption that all "species" are equally distinct. Units that are on first observation similar may have very different evolutionary histories, and ecologists have theoretical reasons to be interested in these developmental pathways.

Fortunately for work in this field, they also have a wealth of leverage over unit differentiation, incorporating deep phylogenetic depth of observational units. In our dataset, however, we lack the kind of rich information on hate speech content that genetic codes provide ecologists. We can say, for example, that a given post contains antisemitic hate speech, but do not have granular labeling on which words or phrases in the post contribute the most to the antisemitic label. Nor can we observe, without considerable substantive historical or ethnographic work, the evolutionary history of a specific rhetorical phrase used by a given antisemitic post.

Future work that incorporates substantive knowledge and builds off our Hill number measure would likely be of interest to those working on hate speech and extremism, however. For example, we know that hate speech against Jewish, Black, and immigrant communities overlaps in posts about the Great Replacement, a conspiracy theory that argues "global elites" are intentionally depressing the birthrates of white Europeans and attempting to "replace" them with immigrants from the Global South (Thompson et al., 2024). Thus, our Hill measure assumes that posts containing Great Replacement rhetoric are similar to other antisemitic posts, anti-immigrant hate speech, or posts containing mixtures of these two categories. However, subject matter experts on antisemitism and anti-immigrant sentiment might argue that the rhetorical and syncretic trajectories of the Great Replacement and other extremist ideologies reflected in these types of hate speech are rather different. The ability to track these "evolutionary" differences and reassess the distance between seemingly common types of hate speech could have interesting descriptive and theoretical implications for the dynamics of emerging extremist and hate groups online.

5.3 Sampling and interpretation

The advantage that a multidimensional index like Hill numbers gives us is ease of interpretation. Trying to compare platforms, communities, or time periods across multiple dimensions can be difficult. Consider, for example, a sampling frame at the platform-week level called P. If P contained two equally common types of hate speech and *P'* contained four equally common types of hate speech, we could plausibly conclude that *P'* is twice as diverse than *P*. Interpreting this measure of "diversity" in the context of targeted hate speech, we might say that *P* offers more targeted hate speech towards two identity categories, while *P'* contains more indiscriminate hate.

But what happens when types of hate speech are not equally common? This kind of intuitive direct interpretation breaks down. *P*, for example, might have many types of rare hate speech, but only one extremely common type. Using the simple species richness measure, we might conclude that this sampling frame offers indiscriminate targeting of hate despite the overwhelming proportion of actual speech targeting just one group.

Hill numbers are directly comparable, allowing us to overcome this drawback. The Hill number for a particular frame at the platform-week level, for example, describes the "effective number" of hate speech types – also known as the "true diversity" of a sample (Jost, 2007). If the Hill numbers for our above example are $P = 2$ and $P' = 4$, we can conclude that (conditional on the q parameter) the effective number of hate speech types doubles from P to P'. In other words, P contains more targeted hate speech than P'.

However, Hill numbers for the same sampling frame may vary with q, also known as the "order." Referencing the order allows us to go beyond just effective species comparisons and quantify the uncertainty or entropy associated with predicting the type of hate present in the next utterance randomly selected from the corpus of speech. In other words, the greater Hill number between two samples of the same order q indicates a greater uncertainty in the identity of the target in the next randomly selected post from our dataset – a suitable measure of indiscriminate targeting of hate.

In practice, the q parameter determines how much weight should be given to the most abundant species in the sampling frame. A Hill number of order

0 is simply the species richness, a count of the distinct number of categories present in the sample without accounting for abundance. When $q = 1$, the Hill number is equivalent to the exponential of Shannon entropy, giving equal weight to both common and rare types of hate speech. The Hill number when $q = 2$ is the inverse of the Simpson index, and is less sensitive to rare types of hate. As q increases beyond 2, more weight is given to dominant species or hate types.

6 Results

We begin our analysis of this new measure by describing the distribution of Hill numbers across platforms and time periods. In general, we should expect that as q increases, the Hill index either decreases – many types of hate but a few types that dominate, indicating discriminate targeting – or remains similar – a relatively even number of types, indicating indiscriminate hate. In Figure 2 below, we can see that the weekly average Hill index for 4Chan across the dataset from 2019 to 2023 is relatively high for $q = 1$. When equal weight is given to common and rare types, hate speech is indiscriminate, hovering around 3.5 targets for any given draw of a hateful post.

Figure 2 also shows a dramatic shift in hate speech in the summer of 2020, likely in response to the murder of George Floyd and subsequent protests against police violence in the United States. From May to June 2020, we see hate speech of 4Chan become significantly more targeted across all three values of q, consistent with our previous work showing a major increase in the volume of anti-Black hate speech in this time frame (Lupu et al., 2023).

The distribution of Hill numbers across q values for the 4Chan corpus displays the expected pattern – as q increases, the "diversity" score or discriminate targeting of hate speech drops as dominant categories of hate are weighted more. However, in Figure 3 describing hate speech indices for Telegram, we see a dif-

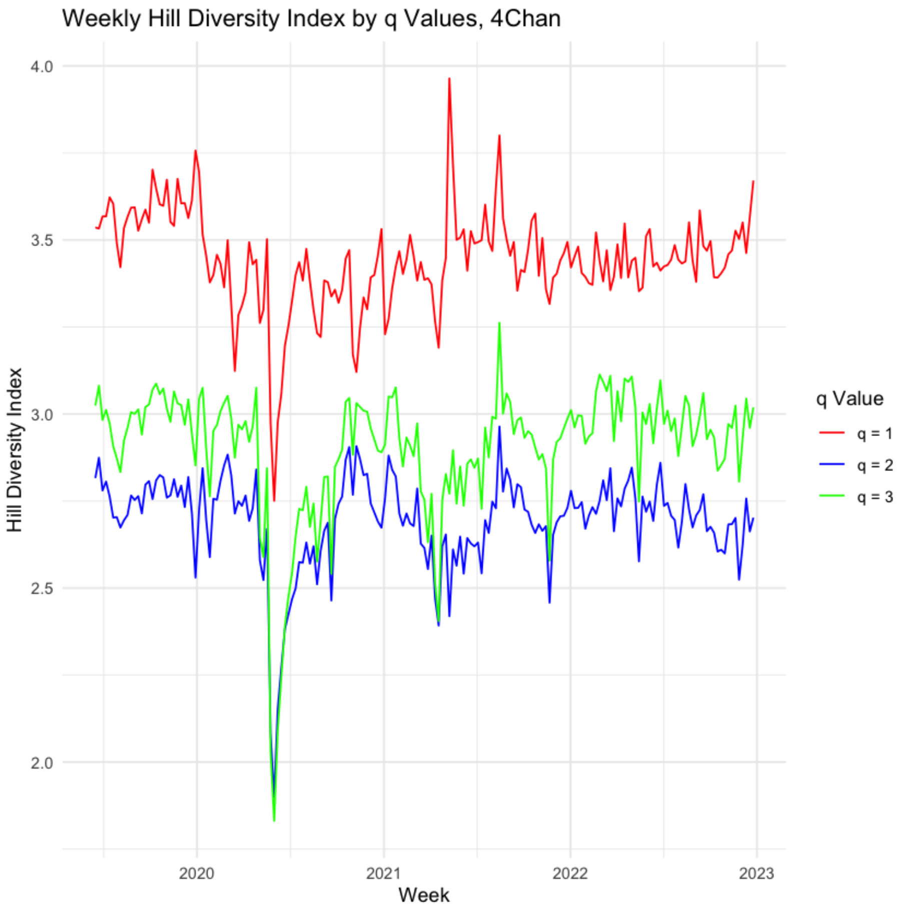


Figure 2: 4chan, Hill Indices

ferent pattern.

In mid-2021, the Hill indices across values of q diverge, showing more indiscriminate hate for order numbers greater than one. This is a puzzling and relatively rare pattern in ecology (Roswell et al., 2021): when more weight is given to the most abundant types of hate, the true diversity of the sample actually increases. There are a few possible interpretations of this outcome.

First, the distribution of abundance in the Telegram corpus may display extreme skewness, such that one type of hate dominates while many other rare

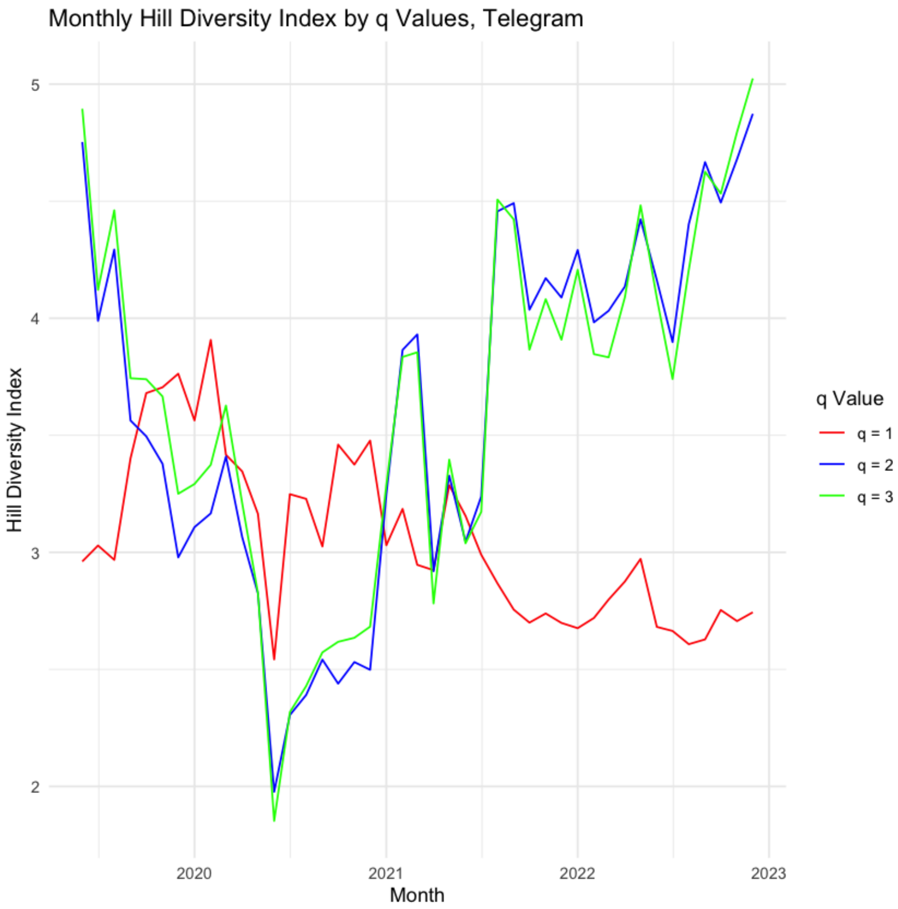


Figure 3: Telegram, Hill Indices

types with small but non-zero proportions collectively contribute significantly to the Hill index. Second, the Telegram corpus may contain clusters of hate speech types with relatively even distributions internal to the group but with large differences between groups. Both of these potential distributions of types within Telegram may have consequences for the discriminate or indiscriminate targeting of hate speech. In the former situation, we see a major dominant form of discriminate targeting with the potential for more indiscriminate speech across a wide variety of groups – what ecological modelers might call "potential niches" (Sillero, 2011). In the latter, we see "realized niches," where clusters of users or clusters of discourse across users indiscriminately but consistently use hate speech towards the same handful of groups. In both cases, we see the potential to use hate speech targeting measures like the Hill index to provoke deeper inter-platform and inter-group research on the relationship between hate speech types and discriminate targeting.

We next analyze the relationship between hate speech targeting and offline events, identifying four event categories across two dimensions: discriminate or indiscriminate targeting; and responsive or unresponsive speech. The former dimension describes a sampling period where the weekly average Hill number is less than or greater than two. We say that a sample is 'discriminate' if the weekly average is less than twp, meaning that any given instance of hate speech from the sample is likely to have less than two targeted groups. Conversely, we say that a sample is 'indiscriminate' if the average weekly Hill number is greater than two, containing speech that is likely to target two or more groups with hate. The latter dimension describes the responsiveness of the sample to the external event. We say that a sample is 'responsive' if the weekly average Hill number shifts by one or more standard deviation (in either direction) in the week after the event takes place.

	Discriminate (${}^1D < 2$)	Indiscriminate (${}^1D > 2$)
Responsive ($\pm 1\sigma$)	George Floyd protests, May 2020 - July 2020	United States Capitol attack, January 2021 - February 2021
Unresponsive	COVID-19 Delta variant, June 2021 - August 2021	2022 FIFA World Cup, November 2022 - December 2022

Table 2: Events classified by responsiveness and type of hate speech.

In Table 2 we show examples of each potential category. The reaction across the dataset to the George Floyd protests in the summer of 2020 is the easiest to describe. Users of hate speech dramatically shifted their focus towards Black

Americans and anti-Black hate speech in response to protest movements that summer. Responsive-discriminate events possess the properties of so-called "focusing events" (Birkland and Lawrence, 2009) with clear salience to vulnerable groups likely to be targeted with hate speech. Compare this, for example, to the unresponsive-discriminate category. The COVID-19 Delta variant surge – and other surges before June 2021 – had clear salience for Asian immigrant communities, who were often the target of hate speech and racially motivated political violence (Gover et al., 2020). However, there were few focusing events associated with this period of time; rather, reporting on the pandemic and infection numbers themselves were a steady drumbeat.

Conversely, the storming of the United States Capitol on January 6th, 2021 drove a significant responsive-indiscriminate change in hate speech. Across the dataset we observe increases in Hill indices in the week after the attack, indicating more indiscriminate hate speech (see Figure 4b).

7 Discussion

References

Pete Burnap and Matthew L Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15, 2016.

Manoel Ribeiro, Pedro Calais, Yuri Santos, Virgílio Almeida, and Wagner Meira Jr. Characterizing and detecting hateful users on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 745–760. Springer, 2018.

Jing Qian, Mai ElSherief, Elizabeth M Belding, and William Yang Wang. Leveraging intra-user and inter-user representation learning for automated hate speech detection. *arXiv preprint arXiv:1804.03124*, 2018.

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

Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546: 126232, 2023.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875, 2021.

Joseph B Walther. Social media and online hate. *Current Opinion in Psychology*, 45:101298, 2022.

Christopher J Ferguson and Brad Glasgow. Who are gamergate? a descriptive study of individuals involved in the gamergate controversy. *Psychology of Popular Media*, 10(2):243, 2021.

Carmen Cervone, Martha Augoustinos, and Anne Maass. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of language and social psychology*, 40(1):80–101, 2021.

Ursula Kristin Schmid, Anna Sophie Kümpel, and Diana Rieger. How social media users perceive different forms of online hate speech: A qualitative multimethod study. *New media & society*, 26(5):2614–2632, 2024.

Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

Lynne Tirrell. Toxic speech: Toward an epidemiology of discursive harm. *Philosophical topics*, 45(2):139–162, 2017.

Yonatan Lupu, Richard Sear, Nicolas Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Beth Goldberg, and Neil F Johnson. Offline events and online hate. *PLoS one*, 18(1):e0278511, 2023.

Andrew Ifedapo Thompson, Maxwell Beveridge, Stefan McCabe, Molly Ahern, Fryda Cortes, Noah Axford, and Jacqueline Martinez Franks. Anti-black political violence and the historical legacy of the great replacement conspiracy. *Perspectives on Politics*, pages 1–18, 2024.

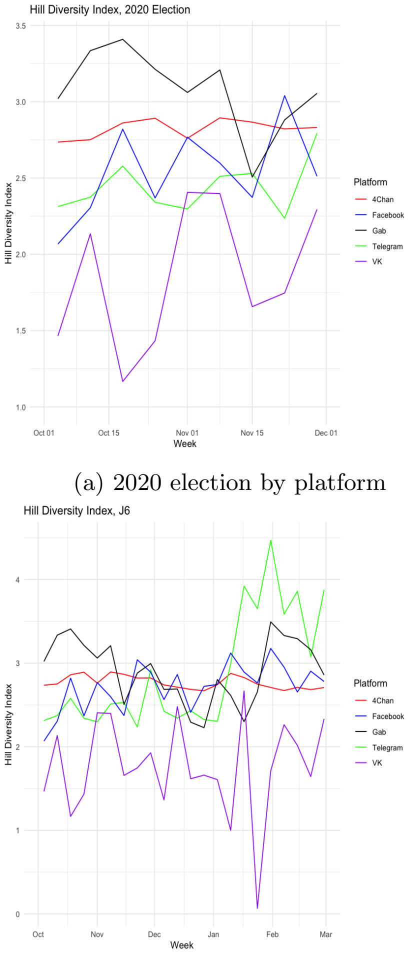
Lou Jost. Partitioning diversity into independent alpha and beta components. *Ecology*, 88(10):2427–2439, 2007.

Michael Roswell, Jonathan Dushoff, and Rachael Winfree. A conceptual guide to measuring species diversity. *Oikos*, 130(3):321–338, 2021.

Neftalí Sillero. What does ecological modelling model? a proposed classification of ecological niche models based on their underlying methods. *Ecological Modelling*, 222(8):1343–1346, 2011.

Thomas A Birkland and Regina G Lawrence. Media framing and policy change after columbine. *American Behavioral Scientist*, 52(10):1405–1425, 2009.

Angela R Gover, Shannon B Harper, and Lynn Langton. Anti-asian hate crime during the covid-19 pandemic: Exploring the reproduction of inequality. *American journal of criminal justice*, 45(4):647–667, 2020.



(a) 2020 election by platform

(b) January 6th attack by platform

Figure 4: Responsiveness to the 2020 Election and the January 6th Attack