Investigating Influential Users' Responses to Permanent Suspension on Social Media

ZIHAN GAO, University of Wisconsin-Madison, Wisconsin, USA
JACOB THEBAULT-SPIEKER, University of Wisconsin-Madison, Wisconsin, USA

Social media platforms use permanent suspension as a measure of last resort to intervene with users who spread harmful or misleading content. However, permanent suspension does not signify the end of a user's online presence, but rather on that specific platform. This issue is particularly salient for influential users with large audiences, as they have the potential to cause substantial shifts in the overall social media information landscape when suspended. Our work employs a mixed methods approach to study the context around, and behavioral patterns after, permanent suspension. We find that migration is a common step after suspension, and characterize a number of behavioral strategies and patterns that occur after influential users are suspended. By focusing on consequences of suspension across more holistically, we have identified numerous opportunities for design and future research to mitigate the potential negative effects of permanent suspensions on the broader social media information landscape.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Permanent suspension, social networks, moderation, social media ecology

ACM Reference Format:
Zihan Gao and Jacob Thebault-Spieker. 2024. Investigating Influential Users' Responses to Permanent Suspension on Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 79 (April 2024), 41 pages. https://doi.org/10.1145/3637356

1 INTRODUCTION

Social media applications serve as a "platform" for user engagement, expression, and social connection [46]. However, problematic conduct on social media, such as hate speech, misinformation, and harassment, has proven to be a significant issue [3, 82, 130]. Given the popularity of social media platforms and the various stakeholder goals associated with them, policing and enforcing what content and behavior is allowed has become a necessary part of large-scale social media systems. To protect users and groups from adversaries and to remove offensive, disturbing, or illegal content and content creators, platforms are left needing to moderate content [8, 46, 133]. Content moderation, broadly, refers to a series of sociotechnical procedures used by platforms to establish boundaries between what is acceptable and what is not (e.g., community guidelines and terms of service), detection approaches to identify unacceptable content (e.g., collective flagging and detection algorithms), and enforcement measures on that content or its creators (e.g., suspension and shadow banning) [49, 85, 106]. How platforms respond to violations of their rules and regulations varies widely, from deleting individual posts to permanently suspending a user's account [60, 85, 101]. Permanent suspension is typically the most severe form of enforcement that

Authors' addresses: Zihan Gao, zihan.gao@wisc.edu, University of Wisconsin-Madison, Madison, Wisconsin, USA; Jacob Thebault-Spieker, jacob.thebaultspieker@wisc.edu, University of Wisconsin-Madison, Madison, Wisconsin, USA.

2573-0142/2024/4-ART79
https://doi.org/10.1145/3637356

platforms impose on users, resulting in a user losing access to all account data, including profiles, posts, and social networks [20, 61, 70]. Through one lens, permanent suspension is the response of a platform to users (influential or not) who unduly shape the information landscape on the platform.

Our research focuses on influential social media users, characterized by their extensive follower base and significant content generation. Remarkably, a significant fraction of permanently suspended users belong to this influential bracket [60, 101]. Unlike typical users, these individuals possess an amplified potential to sway their followers, which can result in unbalanced negative repercussions, such as the spread of misinformation or inappropriate content. In specific cases, permanently suspending influential users seeks to hinder the swift and widespread spread of harmful content [132, 136]. The situation of Alex Jones, a prominent user across various social media platforms and the Infowars creator, serves as a prime example. Notorious for spreading false information and endorsing conspiracy theories, Jones and his platform significantly influenced their substantial YouTube followership, thereby exacerbating misinformation-related harm. In light of these events, YouTube suspended Jones's account permanently as part of its comprehensive approach towards addressing misinformation and hate speech [25]. This highlights the role of permanent suspension as a tool employed by platforms to handle users, influential or otherwise, who adversely affect the information landscape.

Importantly, however, permanent suspension does not mark the end of a user's presence on social media, only the end on one platform. Permanent suspension, unlike content removal that lets users remain active, introduces unique challenges as it may drive users, particularly influential ones, to explore alternate online venues. Prior research has investigated the implications of such moves on other platforms and the communities where banned users might relocate. For example, Horta Ribeiro et al. [55] illustrated that communities formed post-ban may face enhanced toxicity and radicalization. Similarly, Ali et al. [5] found that users expelled from one platform often display increased activity and toxicity on alternative platforms. These studies underscore the complex ramifications of permanent suspension.

Examining user behavior, whether influential or not, across diverse social media platforms aligns with research on social media ecology. This research mainly examines how individuals shift their content and engagement across platforms based on audience, platform features, and their perception of these spaces [33]. This field suggests that users continuously modify their social media ecology, altering behaviors and platform preferences according to their objectives and priorities. If users perceive their current social media environment as less rewarding or convenient, they may modify behaviors on their initial platform or explore new platforms [89, 140]. Permanent suspension could destabilize a user's social media ecology, prompting them to reevaluate their setup [140]. However, how permanent suspension influences users' social media ecosystems and the tactics suspended users employ to maintain their online presence remains unclear.

The consequences of suspending influential users from a platform are both (a) not fully understood and (b) potentially harmful due to influential users' large audiences and unique weight on the information landscape. Therefore, their absence can lead to unexpected outcomes. This becomes particularly significant in the era of data-driven technologies that heavily lean on user-generated content from these platforms for various applications [17, 40, 97, 98, 129]. Currently, many social media platforms are grappling with the ethical implications of their vast troves of user-generated content. These platforms are now reconsidering their content sharing policies, which has been evidenced by actions like restricting API access for various reasons, including commercial [57]. The ripple effect of these actions can influence many sectors, including tech companies that harness social media content. While it is understood that suspending influential users can be an approach to curtail the spread of toxicity or misinformation, the broader implications on the digital content

ecosystem remain uncertain. For instance, some tech tools have had to employ human moderators to filter out harmful output, a portion of which may have roots in social media content [94, 128]. The overarching concern here is that while platforms might succeed in creating a temporary information void by suspending users, the long-term effects on the content landscape, especially with tools that use this data, remain a subject of inquiry.

Given the potential for outsize consequences across the social media landscape, our work here seeks to develop a rich understanding of *what happens when influential users are permanently suspended*, as well as *what consequences this has* for the broader social media information landscape. Specifically, we pose the following three primary research questions:

RQ1: How do influential users react to permanent suspension, and what changes emerge in their social media ecology?

RQ2: How significantly does permanent suspension trigger influential users migration to alternative social media platforms and what discernible patterns characterize such migrations?

RQ3: How can insights from the social media ecology dynamics of suspended influential users inform the development of more effective moderation strategies to curb the proliferation of harmful content?

To tackle these research questions, we employ a mixed-methods approach to deeply investigate influential users' responses following permanent suspension. We first conduct a qualitative thematic analysis to understand the contextual reasons for, and influential users' reactions to, permanent suspension, drawing inspiration from the public health methods used by Lin et al. [73] which rely on news articles as a way of understanding relatively rare events. We further augment our thematic analysis through a "natural experiment" of permanently suspended influential users, examining how their behavior changes across platforms after suspension, in comparison to similar non-suspended users. By combining these two analytical approaches, our work makes five primary contributions:

- First, we demonstrate that how platforms communicate the reasoning behind suspension is a significant factor influencing suspended influential users' subsequent behavior. Our results indicate that when the rationale for suspension is unclear, there is a risk that suspended users and the public may develop false narratives about the reasons. We found that it is more likely that suspended users will attempt to return to the same platform under these circumstances. Based on these findings, we offer several research and design recommendations for effectively communicating suspension reasons to mitigate these risks.
- Second, our research reveals that migration to a different social media platform is a common response among influential users who have faced permanent suspension. This finding aligns with previous studies on social media migration and the dynamics of social media ecology [5, 60, 140]. It suggests that the separation between platforms can serve as both a hurdle for individuals seeking to rebuild their social media presence and a barrier against the proliferation of harmful content. While mainstream social media platforms are often the destination of choice for these influential users, we observe an increasing trend towards alternative platforms over time.
- Third, we show that raising the cost of migration may be an effective way to prevent the spread of problematic content. Influential users migrate between social media platforms with the intention of preserving their social media ecology — reaching the most relevant audiences and maintaining their self-presentation. However, that process of migration, including migrating data, rebuilding social networks, and potentially rebranding oneself, can be costly and may deter users from moving to other platforms.
- We identify four archetypal migration modes, and characterize their contexts and uses, which reflect the objectives and traits of influential user migration. Developing an understanding

of migration modes can help minimize the risks associated with influential users influence on the information landscape, and can lay the groundwork for future research to identify influential users who have migrated to different platforms.
- Finally, we develop a series of design recommendations, like enhancing the clarity of suspension notifications and minimizing the negative consequences of cross-platform migration. Notably, we propose the exploration of cross-platform content moderation strategies to effectively address the challenges posed by influential users who have been suspended and subsequently migrate to other platforms.

2 RELATED WORK

Our work here is contextualized within, and builds upon, three bodies of prior work. Namely: research on content moderation processes, permanent suspension as a technique more specifically, and research exploring and characterizing users' social media ecologies. By examining the intersection of these three fields, our work here builds on and extends these bodies of prior work and develops a deeper understanding of the consequences of permanent suspension.

2.1 The process of moderation

Social media content moderation involves a series of procedures to ensure that content and behavior within a community or platform adhere to community guidelines. Various aspects of how companies enact content moderation include a combination of policies, community rules, detection methodologies, and enforcement. Platform policies (often termed "community guidelines") are shaped by companies' values, users' values, and laws that platforms are held to. Detecting infractions of these policies occurs in a variety of ways as well, including through community volunteer moderators, paid employees or contractors serving as moderators, and algorithmic recognition systems, among others. Once a user or piece of content has been deemed in violation of the policies, enforcement happens through a number of methods, including content removal, temporary account suspension, and permanent account suspension.

2.1.1 Community guidelines. The community guidelines clearly outline the types of content that are not permitted on the platform. The terms and criteria used by platforms for content moderation are often precise, if somewhat ambiguous, providing operational definitions for prohibited content categories and examples of acceptable and unacceptable content [85, 113, 125]. This broadness of criteria and examples is intended to help platforms navigate the tension between creating global standards for a diverse user base while also adhering to the conventions and values of the various local markets in which they operate. However, this can lead to a lack of precision in the definitions provided, which platforms argue is necessary to effectively address potential rule violations by users [7]. Community guidelines for mainstream social media platforms are often similar, despite platform differences, as they typically provide an overarching vision for the type of discourse the platform aims to promote. This may be due to the fact that these platforms rely on user-generated content and play a significant role in shaping public discourse, thus requiring the implementation of rules that protect the user experience and uphold the integrity of public discourse. These platforms often consult each other when dealing with similar forms of harmful content and behavior, drawing upon a long history of speech management for guidance on when to intervene [46]. The community guidelines are not intended as a guide for moderating content, but rather as a definition of the community's desired state. Moderation policies serve as a reflection of the platform's ideals, as well as the friction that arises between platform values, user experience, and public discourse [46].

The written community guidelines, while providing a clear framework for prohibited content, do not fully encompass the complexities of content moderation in practice [58]. The distinction

between acceptable and inappropriate content can be challenging to make, and categorical terms such as "sexually explicit" or "vulgar or filthy" only serve to complicate the matter further. The subjectivity of determining offensive content, coupled with the limitations of written explanations, only risks creating more ambiguous lines that must be constantly evaluated and monitored [46]. This is further compounded by the complexity of the enforcement process, which is often marred by errors, exaggerated responses, and biases [46, 106]. The gray areas, or boundaries, of content moderation present the greatest challenges.

2.1.2 Detection approaches. The task of moderating content on social media platforms is a complex and challenging one, characterized by the large scale and diversity of users, the vast amount of content, and the rapid pace of its dissemination. These characteristics set social media apart from traditional media, and require different approaches to moderation. Social media platforms must navigate a variety of cultural, national, and ideological groups, all with different motivations and goals. In light of this complexity, Grimmelmann [49] gave a taxonomy of the many methods for performing content moderation, including centralized and decentralized approaches, manual and automatic methods, ex post and ex ante moderation, and transparent and secret methods. He also mentioned community features that can influence moderating strategies, such as infrastructure capability, user community size, ownership distribution, and participant identity (not anonymity).

In addition to the above taxonomy, most researchers have divided content moderation implemented by platforms into three categories: editorial review, automatic detection and community flagging [46, 106]. One common approach to moderation is platform editorial review, which is similar to traditional media's handling of offensive content. Some platforms also employ automatic detection software to flag potentially harmful content for human review. Recent research has focused on developing algorithms to automatically detect and remove harmful content at scale [12, 18, 50, 72, 88]. Another approach is community-based moderation, such as flagging systems that allow users to report content that they believe violates rules or conventions [28, 77, 114, 135]. Some platforms also rely on volunteer-based solutions such as bot-based collective blocklists to address harassment and help regulate user experiences [34, 43].

The literature on detection approaches has primarily focused on reactive methods, such as removing problematic conduct after it occurs. However, an emerging area of research has begun to explore proactive approaches to prevent such conduct from occurring in the first place. Seering et al. [107] found that utilizing interface components designed based on psychological principles can be effective in promoting thoughtful and engaged engagement among users. Additionally, research has been conducted on the potential use of chatbots within communities to strengthen community cohesion and establish guidelines [108]. The development of technology to proactively encourage positive community behavior is a promising field for further research and intervention [64].

2.1.3 Sanction enforcement. Social media platforms can delete or hide harmful content or accounts on their platform. Removal as a method of dealing with problematic content or accounts on social media platforms has been extensively studied by academics. Singhal et al. [113] characterized it as "hard moderation," in contrast to "soft moderation." While this approach has some benefits, such as promoting public safety and avoiding association with inappropriate content or behavior, it also has several drawbacks [47, 123]. Hard moderation can be a heavy-handed approach that removes content for all users, which may contradict principles of open participation and freedom of speech [60, 123]. It may also be perceived as censorship and may not be legally mandated for private companies [10, 54]. Furthermore, suspended users may still be able to access other platforms or the internet at large [5].

Soft moderation, also known as concealing, does not involve the deletion of any material. Instead, it aims to bring potential concerns about the content to the attention of other users through

methods such as adding warning labels, limiting the spread of questionable content (i.e., shadow banning), or restricting the ability of users to interact with the content (e.g., by disallowing replies or re-sharing). A significant amount of research has been conducted on the topic of soft moderation through various means, such as interviews, surveys, and crowd-sourcing studies. Some studies have found that warning labels can lead users to seek out additional information for verification, and decrease the intent to share information [15, 42, 63, 81, 83, 93, 105, 109]. However, more recent research has also uncovered the potential negative effects of warning labels, such as the implied truth effect, which suggests that users may view posts containing misinformation but without warning labels as more credible [92]. Another form of soft moderation is shadow banning [85], which involves keeping the information available, but limiting its reach. This method can help decrease the presence of problematic content. However, it can also generate feelings of prejudice and ambiguity among users [9]. Despite its potential benefits, soft moderation poses certain political risks as it can complicate the management of public debate by enabling the dissemination of a wide range of information to large audiences in ways that are difficult to identify or critique. As a result, users may be under the impression of being in the same discourse, while in reality they might be moving across slightly different but overlapped discourse worlds within the same platform [9, 13, 46].

2.2 Permanent Suspension

In addition to removing particular content, social media platforms, under severe circumstances, may permanently suspend users' accounts. While many of these suspensions pertain to clear violations such as child pornography, there have also been high-profile cases of permanent suspension in recent years. A notable example being the permanent suspension of former US President Donald Trump's (@realDonaldTrump) Twitter account in January 2021 due to concerns about the potential for further incitement of violence [56]. Despite legal protections under Section 230 of the United States Communications Decency Act [23, 46], discussion surrounding social media permanent suspension revolves around the issue of freedom of speech, the rights of individuals to access and use social media platforms, and the rights of private information providers to shape and occasionally restrict material continue [22, 46, 77]. Complicating matters further, social media platforms deal with legal changes when operating across cultures and nations, and their position as internet intermediaries is being questioned as they gain ownership of content and organize it to keep users on the platform [24, 46, 69].

Research in the field of Human-Computer Interaction (HCI) on the topic of permanent suspension has mainly focused on the consequences of this form of content moderation. Studies have found that permanent suspension can have negative effects on the suspended user, such as social isolation and economic consequences if the user depends on the platform for communication and networking. Myers West [85] revealed that many users view social media platforms as a means of staying connected with their support networks. For these users, losing access not only means the loss of a platform for their speech, but also the loss of an essential channel of communication with the outside world and an increased likelihood of social isolation [101]. Additionally, Myers West [85] highlighted the economic cost of suspension, as social media platforms serve as a primary mechanism for obtaining referrals for users' businesses, leading users to seek new accounts on the same platform [60, 101]. Other research has further found that users will try to keep their accounts on social media platforms in order to maintain access to modes of receiving donations, processing payments, and selling merchandise [101].

The behavioral and cognitive reactions of users who have been permanently suspended from a social media platform can be intricate and multifaceted. Social media users often use multiple

platforms and are not confined to a single platform, which allows for the possibility of migration to other alternative platforms following content moderation [87].

2.3 The Impact of Permanent Suspension

The act of permanent suspension is deeply embedded in the complexities of online ecosystems. Numerous studies have delved into the ramifications of these suspensions, considering their impact not just on individual platforms, but across the wider digital landscape. These studies take into account both individual and communal viewpoints. From the individual perspective, Jhaver et al. [60] showed that suspending influential users disrupts discussions about them, decreasing their digital footprint and reducing associated anti-social ideas. Ali et al. [5] found that a large fraction of users migrate to platforms like Gab after suspensions, often exhibiting heightened toxicity. This notion of "digital exodus", as articulated by Edwards and Boellstorff [36], points to platforms unintentionally driving users towards less moderated environments. On the community front, while studies like Chandrasekharan et al. [20] and Saleem and Ruths [104] attest to the effectiveness of suspensions on the source platform, the broader implications hint at the migration of behaviors elsewhere. DeCook [31] and Horta Ribeiro et al. [55] emphasized how communities might reform on alternative platforms, often with intensified ideologies. Furthermore, Chang and Danescu-Niculescu-Mizil [21] spotlighted the significance of perceived fairness in suspensions, influencing user reactions post-suspension. Indeed, questions of fairness in permanent suspension echo broader questions of fairness in content moderation as well (e.g. [122]). Contrasting this platform-centric view, Kou [70] approached from the user's lens, arguing against stereotyping suspended users and advocating for a more restorative moderation strategy. Holistically, permanent suspension intersects with issues of digital equity, freedom of speech, and platform responsibility [127]. Understanding these intricate dynamics is pivotal for framing balanced and effective moderation paradigms.

2.4 Social Media Ecologies

Taking the user-centered view on content moderation, as Kou [70] do, accentuates an important perspective with its own body of prior work. Namely: users are the "recipients" of content moderation decisions. As such, content moderation decisions play a role in a user's understanding of their social media world, and inform how users navigate the multitude of platforms and social media systems that are available to them. Prior research has studied social media ecologies broadly. While this work has not focused on content moderation specifically, this body of research informs our work here as well.

2.4.1 Characteristics and dynamics. The impacts of the constantly evolving social media information landscape on individuals are complex [11]. The features, norms, and audience of a particular platform can all influence a user's experience on that platform [140]. However, the availability of multiple platforms allows users to select and arrange them to meet their diverse needs, such as communication [74, 140] and self-presentation [32, 33]. Research has shown that people use different platforms to access different audiences and social networks, switching between them to communicate with various social groups [74]. The affordances of a platform — the potential actions enabled by the combination of a user's intentions and the technology's capabilities — can also shape a user's self-presentation on different platforms [32]. Rather than viewing multiple social media platforms as distinct environments, individuals often see them as part of a larger, interconnected ecosystem, which they can customize to suit their specific needs and preferences, like accessing imagined audiences and accomplishing self-presentation goals [11, 14, 121]. Zhao et al. [140] identified tensions in how users manage their platform choices, specifically the conflicting desires for separation and permeability, and stability and change. On the one hand, users may try to maintain

boundaries between different platforms through tactics such as disguising account names, creating fake accounts, and withholding information about their social media profiles. On the other hand, users may also seek to establish connections between platforms and create stability in their overall social media ecosystem by making explicit links between platforms. These conflicting desires can create challenges for users when new platforms emerge, as they must determine whether to adopt them and how to integrate them into their existing social media landscape.

Taken holistically, previous research indicates that it can be challenging for users to maintain or alter their social media ecologies, which has implications for their behavior in response to content moderation or suspension decisions [11]. For example, Zhao et al. [140] found that if users feel that their existing social media ecosystem is insufficient, they will seek to adopt new platforms. A permanent suspension, for instance, might significantly disrupt a user's social media ecosystem, prompting them to consider options such as trying to regain access to their account, deciding whether to create a new account and on which platform, and finding ways to prevent future suspensions while still maintaining an audience. In light of these issues, we aim to examine the effects of permanent suspension on user reactions and behavior, with a focus on understanding how users rebuild their social media ecologies.

2.4.2 Mainstream and alternative platforms. An integral aspect of how users reconstruct their social media ecosystems is their ability to understand and choose social media platforms. Newer and smaller platforms, such as Gab and Gettr, have become increasingly popular alternatives to mainstream social media platforms [101]. These platforms often tout their defense of free speech and present themselves as more accommodating to the views of extreme users [6, 48, 65, 139]. Some research has suggested the idea that these platforms provide a more implicit concept within these ideas of free speech — these platforms would allow users to keep their content available and avoid the threat of deletion or account suspension [101, 140]. This is particularly relevant for users who have recently experienced permanent suspension, particularly for influential users who are seeking new channels to broadcast their views [86, 101]. For example, Telegram's end-to-end encryption makes it an appealing option for some types of users [110]. Telegram is primarily a messaging app, but it also allows users to create unlimited groups and channels with an unlimited number of subscribers [101]. In other words, popular alternative platforms are perceived as both 'protected spaces' and 'publicity spaces', combining both the need for privacy and the need for amplifying one's voice [86].

Rogers [101] compiled a list of permanently suspended social media celebrities and mapped them to accounts on alternative platforms, finding that popular alternative platforms include BitChute (as an alternative to YouTube), Minds (as an alternative to Facebook), Gab (as an alternative to Twitter), and Telegram. Additionally, they found that other websites, such as personal websites or news subscription services, are also frequent destinations for content from suspended celebrities. Despite the increasing popularity of alternative platforms, YouTube and Twitter continue to be important sources of extreme content. Zannettou et al. [139] similarly conducted a study focused on Gab, a platform known for its defense of "free speech" that is considered a sanctuary for white supremacists [91]. They found that Gab is primarily used for the discussion and dissemination of news and events, and that it attracts hate groups, conspiracy theorists, and other internet trolls. They also found that Gab has a high proportion of hate speech, positioning it as a bridge between mainstream social networks like Twitter and fringe web communities like political discussions on 4chan [139].

Moreover, these alternative communities receive a lot of attention and see an influx of new users when influential people migrate to them after being suspended from mainstream platforms [101, 139]. For example, Alex Jones (founder of conservative site InfoWars) migrated to Real.Video

(a video streaming platform alternative to YouTube) after his account was permanently suspended on YouTube, resulting in a surge of new users [138]. However, even with these bursts of growth, mainstream social media platforms continue to have substantially larger audiences and drive more traffic to extreme content than alternative platforms [101].

2.5 Our Work Here

Taken holistically, prior work suggests that the consequences of suspension can be clearly positive for the platform. However, the effects of suspension on individual users' behavior and choices within their social media ecology, as well as the consequences of a suspension across the broader social media information landscape, complicate the potential upsides of permanent suspension. In other words, while suspension may be good for a given platform, further study is needed to understand how permanent suspension changes individual behaviors and what that means for other platforms. Our work here seeks to address this important gap in the literature.

3 METHODS

Given the heterogeneity of features, groups of users, and content moderation strategies across platforms, we focus our study here on one platform. In an early pilot survey, we found that within the past five years, Twitter received the most attention in public conversations about permanent suspension. According to search results from Nexis Uni, 36.3% of news reports related to Twitter, followed by 21% for Facebook. As such, to address our research questions that focus on the outcomes of suspension in a number of different ways, we scope our study to focus on influential users who were permanently suspended from Twitter. Therefore, we start by providing a brief introduction to Twitter's specific policy on permanent suspensions.

3.1 Twitter's Permanent Suspension Policy

According to Twitter's rules, the platform may enforce various actions based on the type, frequency, and severity of violations it detects, such as tweet deletion, profile modification requests, labeling, temporary suspension, and permanent suspension [125]. There are multiple types of violations that may result in permanent suspension, including information authenticity, account authenticity, sexual content, abusive behavior, hateful conduct, violence-related activity, involvement with violent organizations, suicide and self-harm, illegal activity, and copyright and trademark infringement, etc. [125]. These infractions are described in further detail in Appendix A [125].

As noted above, certain types of infractions can result in permanent suspension, such as promoting the illicit activities of a terrorist organization, posting someone's intimate photos without their consent, and using platform manipulation to artificially inflate certain information [125]. For other types of infractions, the platform will consider the frequency and severity of the violation to determine whether to permanently suspend the account, such as repeatedly promoting suicide, repeatedly violating others' copyright, and repeatedly manipulating elections (a form of civil integrity violation) [125]. Once a decision has been made, Twitter typically notifies the account that will be permanently suspended and provides the user with a nonpublic explanation. Users can file an appeal with Twitter after being permanently suspended if they believe that their suspension was unfair.

3.2 Our Mixed Methodological Approach

Investigating the trajectories of individuals who have been permanently suspended across various social media platforms and the subsequent changes in their social media ecologies presents a significant challenge. When a user is initially suspended from Twitter, all information pertaining to the account, including profiles, posts, and followers, is kept hidden from the public for moderation

purposes, making it impossible to access using standard social media data collection methods like the Twitter API. Furthermore, once a user leaves their original account on the original platform, the process and method by which they respond to a permanent ban are opaque and difficult to trace — after all, this information may be spread out across many different accounts on many different platforms, making it difficult to gather and analyze. Exacerbating the difficulty of this data gathering process even further, permanent suspensions are relatively infrequent occurrences, making it difficult to compile a comprehensive library of cases.

In light of these challenges, this study employs a mixed-method approach that combines qualitative and quantitative methods to explore the complex process of influential users' responses to permanent suspensions. We take a richer, more exploratory thematic analysis approach to understand how influential users react to permanent suspension, and we augment this with a cross-platform "natural experiment" focused on how influential users' behavior changes after being permanently suspended. We bridge both qualitative and quantitative methods, and by bringing these complementary approaches together we mutually validate the results of each approach, strengthening our confidence in our findings.

3.3 Constructing our Corpus and Dataset

In order to conduct our mixed-methods approach, we needed to develop our datasets to serve both qualitative and quantitative goals. This meant (1) establishing a set of influential users who had been suspended from Twitter, (2) developing a corpus of news articles to help us understand this somewhat rare phenomenon (following Lin et al. [73]), and (3) building a dataset of social media user behavior data to evaluate our research questions quantitatively as well.

3.3.1 Identifying Suspended Users. To start constructing our full set of suspended accounts, we aggregated our initial seed list based on the list of suspended users collected by Hanania [51] and Jhaver et al. [60]. We then developed a set of relevant search terms about permanent suspension on Twitter. With these search terms in hand, we followed the content analysis approach described by Stryker et al. [119], and drew random sub-samples from the set of search terms to iteratively develop our search strategy. Ultimately, our dataset was narrowed based on specific criteria aimed at ensuring that it was both recent and composed of accounts that were influential prior to suspension. Specifically, the accounts selected had to meet the following criteria:

(1) The suspended user had at least 10,000 followers before suspension.
(2) The suspended user was suspended between 2017 and 2022.
(3) At least five relevant news reports exist about the suspension

All three criteria were designed to ensure that the included accounts were influential. Of course, many social media accounts can reach 10,000 followers and get suspended, but this may not be a precise indication of influence. For instance, accounts might buy followers to boost their audience size, or use other techniques to game "influence" that would cause a confound for our results. Therefore, we included our last criterion to ensure that the accounts were indeed influential, as the presence of a significant number of related news reports likely indicates public interest and attention around the suspension.

This process resulted in a list of 155 influential accounts that were suspended for a variety of reasons. Because we were interested in how accounts portrayed themselves prior to being suspended, we also used the Internet Archive's Wayback Machine and Social Blade to collect information about accounts on Twitter prior to suspension. Out of the 155 accounts examined, 12 of them were found to have either no data or were deemed inaccessible. This inaccessibility arose from certain screenshots not successfully capturing the content, displaying only a message

indicating that the site could not be accessed. As a result, these accounts were excluded from the analysis, leading to a final dataset comprising 143 suspended influential users.

3.3.2 Building our News Reports Corpus. With this list of accounts in hand, we then turned to building our overall corpus for analysis. We used the news search engine, Nexis Uni, to identify relevant articles. We again followed the method for identifying search terms described by Stryker et al. [119]. To identify all related news reports about a given suspension, we focused on four important terms in the search query. The first three of these terms were: users' username (handle), display name, and "Twitter". The fourth term was intentionally more open-ended, in order to encompass the variety of concepts related to permanent suspension [119]. For example, some news reports use synonyms to describe the phenomenon (such as banning from the platform, deplatforming, etc.). Thus, we generated a set of synonyms and hypernyms to capture the broader concept, such as ban and deplatform. Then, using these terms, we constructed our search phrases by iteratively assigning different search terms. We restricted our search to news reports that were published from 2017 to 2022, and articles in English. After getting the search results for each suspended user, we first sorted the results by relevance and removed duplicate results. For the first ten search results, we manually browsed the snippet summary on the search result page to understand the general content of each news article, and then manually filtered out the irrelevant articles. Finally, we constructed a corpus of 1,144 news articles, with an average of 8 articles per suspended user.

In constructing this corpus, we were also curious about some generalized information about the suspended user's profile, prior to suspension. To gather this data and include it in our thematic analysis, we also recorded the account information (username and handle) of the permanently suspended users reported in the news and the new platform used by the users after they were suspended. Because it is impossible to gather this data from Twitter after an account has been suspended, we collected archival data from the Internet Archive's Wayback Machine [1] and Social Blade[2]. We recorded some basic numeric information about users' profiles before they were permanently banned, including the number of posts, number of followers, number of people the account was following, and the profile text on their Twitter account page.

3.3.3 Building our User Behavior Dataset. Our News Article Corpus enables our qualitative thematic analysis approach (described below). Since permanent suspensions may lead to cross-platform behaviors like users registering accounts on new platforms or transferring content to their existing accounts on other platforms, we needed a dataset that enabled studying the impact of permanent suspensions on user behavior through comparisons across platforms. One intuitive way to do this is to compare user activity levels on the original and new platforms. Therefore, to conduct our "natural experiment" and answer our more quantitative research questions, we also needed to construct a dataset that includes metrics of user behavior of suspended users, and identify similar users who were not suspended as well, for the purposes of comparison (Figure 1).

To identify these similar users, we started with the 1000 users with the largest number of followers on each platform, as a set of potential users who would be similar to our influential suspended users. To identify the top 1000 users on Twitter, Facebook, and Instagram, we utilized Social Blade to identify the top 1000 users on each platform. Social Blade provides comprehensive data and rankings based on follower counts and engagement metrics for users on these platforms. For Gab, we employed a large-scale dataset consisting of posts and user profiles [39] to identify potential similar users. Fair and Wesslen [39] built this dataset through web scraping conducted between

[1]https://web.archive.org
[2]https://socialblade.com

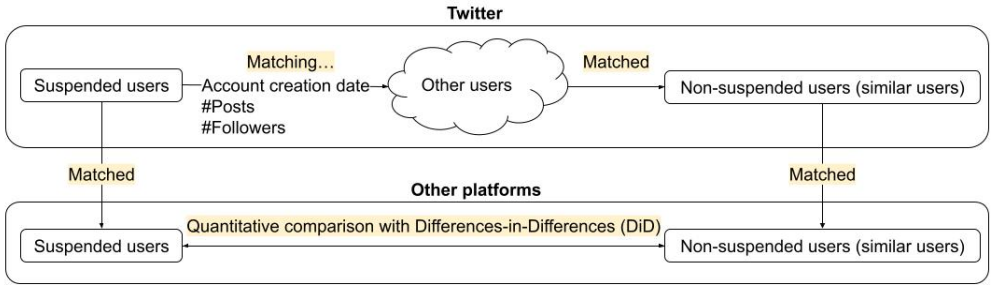

Fig. 1. We match suspended influential users with non-suspended users on Twitter. Then find these users' accounts on the new platforms, and use Differences-in-Differences (DiD) method to compare user activities on new platforms.

August 2016 and December 2018. For Gettr, we accessed a public Gettr API client, *GoGettr*, to extract data and platform dynamics from Gettr, enabling us to identify influential users on the platform [90].

Following previous studies on user matching [20, 21], we employ a similar approach to determine the similarity between potential similar users and our set of suspended users — a necessary component of our "natural experiment". To compute this similarity, we used the Mahalanobis distance metric, which takes into account factors like account creation date, number of posts, and number of followers. Our goal in doing this matching is to identify users who have similar characteristics to our suspended influential users and can serve as a comparison point. To identify the most similar non-suspended user for each suspended influential user in our data, we selected the pair with the largest similarity score from a computed set of pairwise similarity scores. For the suspended users, we did indeed rely on their bios and media reports to locate their migrated handles. For the control group, considering their sizable follower base, we initially searched for their presence on the migrated platform using their known handles and profile information. It is worth noting that while the suspended users predominantly migrated to right-wing/conservative platforms, not all control users (even with large followings) were present on these platforms. If the most similar non-suspended user was not present on the right-wing platform, we then proceeded to the next most similar user in our list. We continued this approach, iterating until we located a user with an account on the platform in question. After matching the suspended influential user with a similar non-suspended user, we again collected account data for both our suspended influential users and their similar non-suspended users through the Internet Archive's Wayback Machine and Social Blade, for six months before, and after, the date when the relevant suspension occurred. In short, we developed a dataset that captures six months of user data before suspension, and six months of user data after suspension on other platforms, for both our suspended user and their most similar non-suspended influential user. This enables our cross-platform "natural experiment" comparison between suspended and non-suspended users, which we describe in more detail below (Figure 2). A key aspect of this dataset is that it includes account information for both our suspended user and their paired similar user *across multiple platforms*; after all, differences in user activity levels across platforms can be affected by any number of platform-specific factors, including platform size, features, norms, etc. Our dataset accounts for this possibility by ensuring that we can measure both the suspended user, and the same paired, non-suspended, user across multiple different social media platforms.

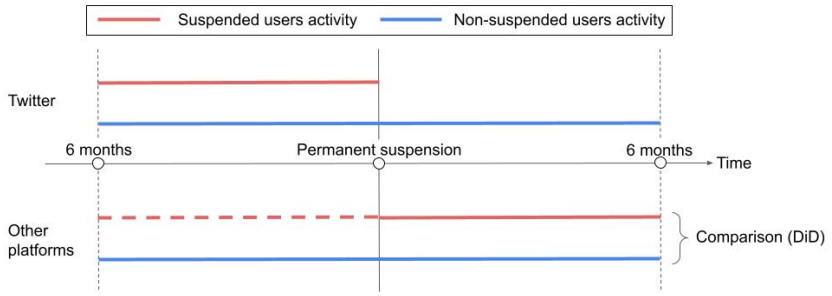

Fig. 2. We employ the DiD method to analyze the activity of two groups of users who migrated to other platforms after being suspended. The DiD method allows us to compare and assess the impact of the suspension on influential user activity. Within the graph, the dashed line corresponds to the activity levels of suspended users who had existing accounts on those platforms before facing permanent suspension.

3.4 Our Data Analysis Methods

3.4.1 Qualitative Thematic Analysis with Our News Article Corpus. Our research questions center on understanding the reactions of influential users to permanent suspensions, which first requires identifying suspended users' Twitter usernames or handles and then using web archive websites to access the suspended users' account data. As described above, this data collection process is impossible with common API or scraping-based methods. In other words, as described by Lin et al. [73], we are studying "relatively rare phenomena of media interest for which no authoritative dataset is available and for which simulations are not currently tractable", we follow Lin et al. [73]'s approach of relying on a corpus of news articles describing permanent suspensions. Specifically, we built a corpus of news articles about influential Twitter users who were permanently suspended over five years (2017-2022), and we rely on thematic analysis to analyze this data.

To understand the impact of permanent suspensions on influential users' social media ecology, two coders coded the news corpus according to a thematic analysis approach. The two coders performed pre-coding (inter-rater reliability is 0.719) and formal coding (inter-rater reliability is 0.874) together according to the six phases of thematic analysis defined by Braun and Clarke [16]. We discuss each of these phases in turn, below. In addition to the coding, we gathered user information, like account creation date, number of followers, number of posts, etc., from their social media accounts (including Twitter and other platforms they utilized after being permanently suspended) and produced a social media ecology table based on it. This provided us with a relatively full picture of the influential user's migration and social media ecologies change.

Data familiarity. In the pre-coding phase, in order to familiarize ourselves with the data, the lead author of this paper read fifteen suspended users' news reports in the corpus, and extracted several initial potential codes. We then iteratively aligned these potential options with research questions, and finally specified the initial codebook, which we used to extract relevant portions of text from the full corpus. During the formal coding phase, the two coders integrated their understanding of the codebook by comparing and discussing differences in the pre-coding results. Then the two coders read through the whole corpus and modified the codebook as needed to guide the subsequent coding process.

Generating initial codes. After extracting these salient segments from the full corpus, the first author then performed an iterative, open-coding process on those segments. In cases where these segments of articles indicated that the owners of these accounts would be migrating to other social media platforms, we used a snowball sampling strategy to make targeted queries about that

migration. For example, if news articles about a suspended user in our data report that the user's plan is to migrate to Gab, we use a similar query to our example above (substituting "Gab" for "Twitter") to retrieve additional news sources about this user and Gab. Throughout this coding process, we worked to identify themes in the text of our news corpus. When a new theme arose, we iteratively revisited the previous articles we had seen with the new theme as a possible code as well.

Identifying themes. After reaching a stable set of codes, and completing the coding of our corpus, we followed an affinity diagramming process, grouping codes together into different potential themes. This affinity diagramming process culminated in an initial thematic diagram [16], drawing relationships among different groups of codes and themes.

Reviewing themes. We then examined the codes within each potential theme to ensure that themes accurately reflected the clear contours of the underlying data. We did this iteratively throughout the entire set of initial themes, ensuring (1) that each theme was specific and cohesive, and (2) that the thematic map accurately reflected our entire dataset.

Defining and naming themes. Once our themes were finalized, we worked to characterize their scope and content in a couple of sentences and assigned a name to each theme.

3.4.2 Quantitative Comparison with Differences-in-Differences (DiD). In addition to our thematic analysis approach, which offers qualitative insights into how individuals respond to permanent suspensions, we bolster our methodology by adopting a "natural experiment" framework to investigate the broader implications of permanent suspensions on influential user behavior within the social media ecosystem. We operationalize this "natural experiment" through a quasi-experimental econometric technique known as Differences-in-Differences (DiD). This technique has previously been used for assessing the causal impact of permanent suspensions on key account behaviors — such as the number of posts and number of followers [20] — and enables us to perform a quantitative analysis of behavioral changes attributable to permanent suspension.

More specifically, using differences-in-differences, we are able to treat permanent suspension as an "experimental intervention," and compare users in the "experimental treatment" group with users in the "experimental control" group. This comparison group relies on pairing each suspended Twitter user from our dataset with a similar social media user who had not experienced suspension, and the paired users serve as our "experimental control". The DiD approach focuses on the average change in our behavioral variables over time for users who have been permanently suspended, and contrasts this with the average change over time for a comparable set of non-suspended users. Crucially, some users had established accounts on alternative platforms prior to being suspended, thereby providing us with direct access to their pre-suspension data. However, for those who created accounts on other platforms post-suspension, we created a "synthetic control" by combining control units in a weighted manner, such that it closely mimics the characteristics of the treatment group prior to the suspension, following best practice [1].

To provide a more intuitive sense, the DiD approach permits us to deduce that a permanent suspension significantly alters user behavior if we observe a statistically significant and meaningful variation in the behavior variables we are examining, relative to non-suspended users. In formal terms, we evaluated our research questions by constructing two Ordinary Least Squares regression (OLS) models to analyze the user behavior data and estimate the relevant parameters. In these models, the dependent variables were the number of followers and the number of posts for the accounts of users who were suspended. We constructed our model as follows:

$$Y = \beta_0 + \beta_1 PS_i + \beta_2 Time_i + \beta_3 PS_i \times Time_i + e_i$$

In this model, PS_i represents permanent suspension, where if $i = 1$, the group of users is the group of permanently suspended users (treated users) and if $i = 0$, the group of users is the group of paired users (control users). $Time_i$ is the period before and after permanent suspension, where $i = 0$ represents the pre-period and $i = 1$ represents the post-period. The parameters of the model are as follows: β_0 represents the baseline level of activity for the control group, which is the level of activity for the paired users before suspension. β_1 represents the increment in activity level for the treated group compared to the control group. β_2 represents the increment in activity level for the post-period compared to the pre-period, regardless of the group. β_3 represents the incremental impact of going from the pre-suspension period to the post-suspension period and from the control group to the treated group, which is the differences-in-differences estimator we focus portions of our analysis on here.

In other words, for the purposes of our analysis, this regression includes a few control variables (PS_i and $Time_i$) that control for the independent effects of being in the control group versus the treatment group or being before or after the point of suspension. The *intersection* of these variables represents the "experimental condition", comparing suspended users to non-suspended users both before and after suspension.

3.5 Methodological Reliability

To enhance the reliability and robustness of our methodology, we implemented several measures to address potential biases in press coverage and strengthen the validity of our findings. Acknowledging the inherent biases that can arise in press coverage, particularly when influential figures and political users are involved, we approached the data with caution and worked to ensure that our own interpretations of the data and the findings were not skewed based on biased reporting. Further, we cross-referenced information from multiple sources, including social media posts, public statements, and user testimonies whenever available. This approach allowed us to verify and validate the reasons for suspension, motivations to migrate, and other relevant factors.

Recognizing the potential for self-serving narratives from banned influential users seeking to present their suspensions strategically, we critically evaluated the information obtained from press coverage. Rather than relying solely on one press account, we treated them as one piece of evidence among others. By seeking corroborating evidence and alternative perspectives, we aimed to ensure a comprehensive understanding of the motivations and migration tactics of suspended users. This careful evaluation helped us mitigate the influence of misleading or self-serving narratives. Our "natural experiment" serves as another reference point to establish confidence in our results, allowing us to observe and analyze the behavior and influence changes of suspended influential users, providing important empirical evidence to complement the limitations of relying solely on press coverage data.

3.6 Methodological Limitations

As with all research, our study has certain limitations due to methodological choices that may affect the generalizability of our results. First, our sample was not randomly selected. We relied on a previously generated list and expanded it based on reports in the public media. While this was necessary for our research questions, it did limit the size and representativeness of our sample. Second, our reliance on the news media corpus for thematic analysis limits the breadth of information available in our data. We tried to supplement this with manual data collection from influential users' social media accounts, but our public media corpus is primarily composed of written news. Third, we limited the timeframe of our analysis to six months before the suspension because we were relying on archival approaches like the Wayback Machine to gather this data. A

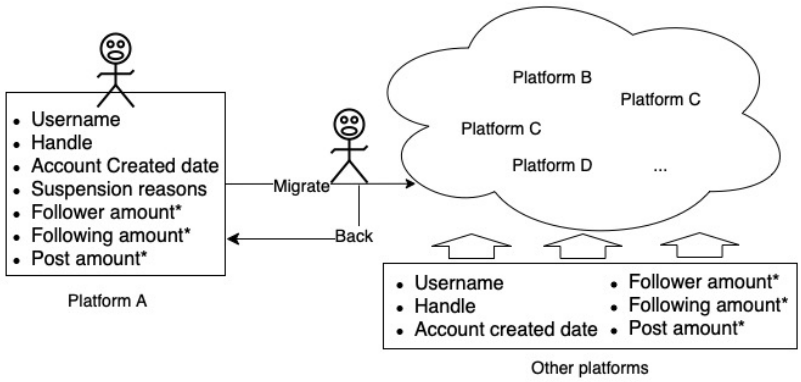

Fig. 3. This figure shows influential users' account information on the original platform and platforms where suspended users have migrated to. Variables marked with an asterisk (*) indicate that we have collected data for a period of six months before and after the permanent suspension, if the data is available.

more longitudinal approach to capturing suspension data might allow for a wider window of data, and we think this is an interesting direction for future work.

4 RESULTS

Our thematic analysis yielded six distinct themes, which notably align with findings from our DiD analysis. Consequently, we have organized our Results section based on these themes and have integrated the DiD findings where they are relevant and applicable.

To facilitate comprehension and based on the outcomes of our thematic analysis and the data gathered from the Internet Archive website, we have created a visual representation, Figure 3, to illustrate our high-level results and the pertinent account information employed in our analysis. We delve into this in greater detail in Section 4.2, providing a comprehensive discussion.

4.1 Permanent suspension reasons

The first theme that emerged in our analysis is the reasons behind permanent suspension. Generally, when a suspension occurs, Twitter notifies the user about the suspension and provides an explanation for the policy or policies they have violated and the content they posted that was in violation [124]. However, the reason for the suspension is only visible to the suspended user, while the public can only see that a particular user has been suspended (e.g., Figure 4). Therefore, in our data, the suspension reasons were all derived from related news reports. In the majority of cases in our dataset, the suspended users themselves (or individuals close to them) provided the reason for suspension, either through statements to journalists or posts on other social media platforms. In some instances, Twitter itself provided an official reason for suspension. For example, Twitter's spokesperson informed BuzzFeed News about the suspension of Gavin McInnes (@Gavin_McInnes):

> "We can confirm that these accounts have been suspended from Twitter and Periscope for violating our policy prohibiting violent extremist groups," a company spokesperson said in a statement to BuzzFeed News [103].

According to our analysis, the majority of permanent suspension cases were accompanied by a reported reason for the suspension (90%). As shown in Table 1, the top three reasons for permanent suspension were hateful conduct (13.29%), misleading or deceptive identities (12.59%), and platform manipulation and spam (11.19%). These three categories accounted for the majority of suspensions

Fig. 4. An example of the permanent suspended user's profile page

among all suspended influential users, reflecting a significant consequence for the broader social media information landscape if left unaddressed. The continued spread of hostile, misleading, or manipulative content poses risks to both the culture of other social media platforms and the perspectives reflected in AI systems that rely on social media content for training data.

However, for some users, news reports did not reveal their rationale for suspension. We found that 14 users (9.79%) in our dataset were suspended for unknown reasons. For example, the former leader of the English Defence League Tommy Robinson (@TRobinsonNewEra) was suspended from Twitter in 2018 for "hateful conduct", but Twitter did not provide an explanation regarding the specific content that led to the violation [99]. In such instances, news articles (like those in our corpus) can only speculate about the cause of a particular suspension based on the context, making it difficult to confidently identify the reason behind the suspension.

Our analysis revealed that in many cases, the reasons for permanent suspension were not limited to a single infraction. In fact, several cases in our dataset showed that influential users were suspended for multiple violations and reasons. For example, in the case of Gemma O'Doherty (@gemmaod1), she was suspended due to both hateful conduct and abusive behavior [35]. Similarly, David Duke (@DrDavidDuke) had previously violated the rule about "violent organization" and was subsequently suspended for "hateful conduct" [37, 102]. Furthermore, we discovered that certain individuals, despite receiving a temporary suspension for breaking a particular rule, would create new accounts in order to continue their actions. This can result in a permanent suspension for "ban evasion." These observations underscore the complexity of the decision-making process behind suspensions and the need for a comprehensive evaluation of all possible violations and reasons to ensure fair and appropriate outcomes.

4.2 Suspended users' responses

The second theme we found in our data focuses on influential users' responses to suspension. It is important to note that the responses of suspended users to their permanent suspension are varied and complex. By tracing users' footprints across various social media platforms after permanent suspension (Figure 3), we have observed three kinds of reactions among suspended influential users: working to stay on the original platform, disappearing altogether, or migrating to a different platform. The first reaction, working to stay on the original platform, typically involves the suspended user attempting to regain access to the platform by appealing their suspension or creating new accounts. The second reaction, disappearing altogether, involves the suspended user

Table 1. Top-10 reasons for permanent suspension of influential users

Reasons	Descriptions	#	%
Hateful conduct	Abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized	19	13.29%
Misleading and deceptive identities	Engaging in impersonation or using a misleading or deceptive fake identity	18	12.59%
Platform manipulation and spam	Severely and artificially amplifying or suppressing information or engaging in behavior that manipulates or disrupts people's experience	16	11.19%
Abusive behaviors	Behaviors that harasses or intimidates, or is otherwise intended to shame or degrade others	12	8.39%
COVID-19 misleading information	Misrepresenting accounts' affiliation as a medical practitioner, and spreading the COVID-19 false information or misinformation	11	7.69%
Violent organizations	Affiliating with or promoting the illicit activities of a terrorist organization or violent extremist group	9	6.29%
Civic integrity	Sharing content about manipulating or interfering in elections or other civic processes	8	5.59%
Coordinated harmful activity	Using specific, detectable techniques of platform manipulation to engage in the artificial inflation or propagation of a message or narrative	6	4.20%
Ban evasion	Circumventing a Twitter enforcement action (such as a permanent suspension) by creating accounts or repurposing existing accounts to replace or mimic a suspended account	6	4.20%
Perpetrators of violent attacks	Individual perpetrators of terrorist, violent extremist, or mass violent attacks	6	4.20%

Note: The detailed descriptions of each reason are in the Appendix A.

leaving the social media ecosystem altogether and discontinuing their online presence. The third reaction, migrating to a different platform, involves the suspended user moving their presence to another social media platform in order to maintain their online presence and audience. Working to stay can be seen as one kind of migration — back to the original platform. Here, we distinguish between these two reactions, aiming to separate the effect of permanent suspension on the original platform and on other platforms.

4.2.1 Working to stay. We found that users who work to stay on the original platform may take various actions, such as creating new accounts or taking legal action against the platform. An example of this is Craig R. Brittain (@SenatorBrittain) who attempted to sue Twitter for suspending his account, but the suit was dismissed [30, 100]. However, he later created a new account, Craig R. Brittain for US Senate, Arizona R-2022 (@BrittainAZ), which is potentially in violation of Twitter's "ban evasion" rule, despite his claims that it was a staff account and that he had moved to Facebook.

Moreover, our DiD results (Table 2) shed further light on this phenomenon: users with second accounts on Twitter, prior to suspension, saw a substantial increase in the number of followers per month on their second accounts, compared to their similar non-suspended counterparts (β_3 =

Table 2. Results of OLS regression with Account Types

Platforms	Account Creation Time	Dependent Variables	Coefficient: β_3	Std	t	P >|t|	R^2
Twitter	Pre-Suspension	Number of followers	430.743	207.613	2.075	0.038*	0.040
		Number of posts	5.792	1.151	5.030	0.000*	0.101
	Post-Suspension	Number of followers	423.490	165.457	2.560	0.011*	0.021
		Number of posts	10.759	1.487	7.234	0.000*	0.125
Facebook	Pre-Suspension	Number of followers	136.049	68.898	1.975	0.049*	0.071
		Number of posts	7.013	0.856	8.189	0.000*	0.135
	Post-Suspension	Number of followers	59.538	165.176	0.360	0.719	0.033
		Number of posts	1.990	1.212	1.642	0.102	0.154
Instagram	Pre-Suspension	Number of followers	288.788	116.373	2.482	0.013*	0.014
		Number of posts	1.968	0.422	4.658	0.000*	0.091
	Post-Suspension	Number of followers	513.709	316.016	1.626	0.105	0.021
		Number of posts	3.586	1.295	2.770	0.006*	0.040
Gab	Pre-Suspension	Number of followers	293.465	117.109	2.506	0.013*	0.145
		Number of posts	4.517	2.164	2.087	0.037*	0.060
	Post-Suspension	Number of followers	404.011	188.406	2.144	0.033*	0.025
		Number of posts	5.233	2.297	2.278	0.023*	0.055
Gettr	Pre-Suspension	Number of followers	452.649	221.299	2.045	0.042*	0.048
		Number of posts	12.329	2.283	5.400	0.000*	0.245
	Post-Suspension	Number of followers	472.231	221.746	2.130	0.034*	0.037
		Number of posts	3.634	1.227	2.961	0.003*	0.151

Note: The coefficient "β_3" is our DiD estimator in the regression model, signifying the additional effect of suspension on influential users compared to non-suspended ones. A p-value less than 0.05, denoted by an asterisk (*), confirms this effect is statistically significant at the 5% level.

430.743, $p = 0.038$). Permanent suspension drives on-platform growth for accounts associated with the suspended user, suggesting that maintaining a second account on the platform may be a way to remain resilient to suspension and maintain audiences. Moreover, users who created new accounts on Twitter after suspension also exhibited a significant increase in the number of followers (Table 2, $\beta_3 = 423.490$, $p = 0.011$). Interestingly, we found that both categories of users tend to post more frequently from their new accounts on the platform after suspension, compared to control users, providing evidence that resiliance may be possible, but requires active effort to reestablish one's presence. This is somewhat expected, given that the audience associated with their old account has been lost.

4.2.2 Disappearance. In our thematic analysis of our news corpus, some suspended influential users disappear after permanent suspension. The reasons for their disappearance are variable — either passive or active in nature. In some cases, accounts maintained false identities — like DCLeaks (@dcleaks_) and Guccifer 2.0 (@GUCCIFER_2), who were identified by the Justice Department as fronts for Russia's Main Intelligence Directorate (GRU) agents — and passively disappeared once exposed. Another example of this style of passive disappearance is accounts claiming to be associated with major news organizations: BBC Afghanistan(@BBCAfghanNews), MSNBC Afghanistan (@MSNBCAfghan) and CNN Afghanistan (@CNNAfghan). These accounts were permanently suspended for propagating a false story about a fictitious CNN journalist named "Bernie Gores", who they claimed was killed by the Taliban during the Fall of Kabul. None of these accounts were verified as being affiliated with the news organizations in question, but posts from

these accounts about this story still received dozens of retweets prior to being suspended. A tweet by "CNN Afghanistan" even received more than 900 retweets and 1600 likes before being suspended [29]. Because our analysis relies heavily on news articles and user account information found on the Internet Archive website, we cannot be sure that the people behind these accounts left social media platforms completely. It may be that some of these users are intentionally less publicly visible, but are still participating in less visible spaces online.

Other forms of disappearance from social media are more proactive, as some users choose to voluntarily relinquish their individual social media accounts. One example is Paul Golding (@GoldingBF), a prominent leader of an organization called British First, who has previously received suspensions. In Golding's case, he gave up his individual account and focused on operating the organizational account, a phenomenon that is discussed in more detail in Section 4.4 [44]. In other cases, users actively left social media, in part because they did not rely on it to reach their audiences. For example, Raúl Castro (@RaulCastroR) was suspended from Twitter, and chose not to seek out new accounts on other social media platforms. However, being a prominent figure outside of social media, Castro is likely still able to reach his audience through other means [76].

4.2.3 Migration as an alternative. Migration to other platforms was the third option chosen by influential users in our dataset, with 113 users (79.02%) transitioning to alternative social media platforms. This migration, coupled with the potential of their audience following suit, may exacerbate the spread of harmful content across the social media landscape. We define user migration by three indicators: (1) users announcing their migration publicly, (2) users creating new accounts on different platforms post-Twitter suspension, and (3) noticeable shifts in users' platform activities post-suspension, such as increased posting. The first two indicators are discernible through our news articles corpus and account data. The third, subtler indicator is inferred from behavioral changes without direct migration confirmation. In such cases, a Mann-Whitney test evidences statistically significant differences in users' pre-and-post suspension posting volumes.

Using our defined metrics, we identified individuals in our dataset who responded to their suspension by migrating to different platforms. Our data show that migrating users seek out both mainstream platforms like Facebook, alternative platforms like Gab, and even subscription-based services such as Substack. For instance, the account of True Indology (@tiinexile), an Indian right-wing historian, was suspended from Twitter, leading them to start a subscription-based email newsletter, which gained 123,359 supporters [53]. Systematic suspension across multiple mainstream platforms can push users towards more alternative platforms. A case in point is Laura Loomer (@LauraLoomer), who, after spreading hateful and anti-Muslim rhetoric, was suspended from a swath of social media, ride-sharing, and money transfer platforms such as Twitter, Instagram, Facebook, PayPal, GoFundMe, Venmo, Uber, and Lyst [116]. Another case is Owen Benjamin(@OwenBenjamin), banned from Twitter, Facebook, Instagram, and YouTube for policy violations. Benjamin created accounts on multiple alternative platforms like Odysee, Bitchute, Rumble, among others. He kept his audience updated on his streaming and social media activities on these alternative platforms through his personal website [111].

Our DiD analysis sheds further light on the effects of permanent suspension as well. These results generally show that on most platforms, users' number of followers and posts significantly change after permanent suspension compared to control users (Table 2). On Facebook, the effects of suspension varied. While there was a statistically significant increase in followers and posts pre-suspension, post-suspension saw a smaller and statistically insignificant increase in followers ($\beta_3 = 59.538$, $p = 0.719$) and a significant but smaller increase in posts ($\beta_3 = 1.990$, $p = 0.102$) compared to pre-suspension. For Instagram, there was a substantial increase in followers and posts both pre-suspension and post-suspension. However, the increase in followers post-suspension was not

statistically significant (β_3 = 513.709, p = 0.105), whereas the increase in posts was significant (β_3 = 3.586, p = 0.006). The results from the Gab platform indicated a significant increase in followers and posts both pre-suspension and post-suspension, with a higher increase post-suspension in both followers (β_3 = 404.011, p = 0.033) and posts (β_3 = 5.233, p = 0.023). On Gettr, while there was a statistically significant increase in followers and posts pre-suspension, the post-suspension increase was also significant, but slightly lower for followers and much lower for posts. Broadly, permanent suspension generally leads to increases in posting behavior and audience growth across all examined platforms, but to varying extents. This change was consistent for users with both pre-existing and newly created accounts on mainstream and alternative social media platforms.

The reactions of suspended influential users to their suspensions raise significant complications for current moderation strategies and show how a permanent suspension from one platform can shape dynamics across the broader social media landscape. For instance, suspension can prompt users to adopt a coded or cryptic language to avoid detection and stay on the original platform. While this may seem like an individual survival tactic, it has wider implications. It can exacerbate the challenge of detecting harmful or misleading information across social media platforms, potentially affecting downstream technologies that use social media for training data, perpetuating subtle, coded forms of harmful content [50]. Furthermore, the mass migration of users from one platform to another following suspension can significantly shift the linguistic norms and overall culture of the recipient platform. For example, a significant influx of users following one suspended influential user can risk altering the platform's culture. Overall, while suspension might be immediately effective for a particular platform, our study shows that suspension has broader, systemic implications, likely leading to negative alterations in the social media information landscape.

4.3 Motivation of migration

Motivations for migration among permanently suspended influential users also emerged as a key theme in our analysis. We analyzed news articles to understand the reasons why suspended influential users migrated and the barriers they faced. Our analysis revealed that users' decisions to migrate to other social media platforms were primarily influenced by two factors: (1) the users' goals for their social media use, and (2) the cost of migration. Specifically, we find that individuals may be inclined to transition to a different platform if it more effectively satisfies their social media goals. Conversely, the cost of migration, including the time and resources necessary to transfer personal data and connections to a new platform, acts as a deterrent.

4.3.1 Goals for social media usage. We found three distinct dimensions of suspended influential users' goals for their social media use: the desire to reach imagined audiences, the need for communication, and the availability of features and policies on the platform. Specifically, regarding the first dimension, we found that individuals in our data frequently seemed motivated to migrate because of a desire to continue reaching their imagined audience. In certain instances, this may include a desire to maintain social influence or retain supporters. For example, as previously mentioned in 4.2.3, Laura Loomer (@LauraLoomer) was banned from Twitter and other platforms and claimed that these bans from "Big Tech" prevented her from reaching her "millions" of followers, damaged her livelihood, and hindered her political aspirations [116]. On the other hand, in some cases, users may desire to tell their side of the story regarding their account suspension, such as David Vance (@DVATW) who was suspended in 2020 for violating Twitter's rules on hateful content. After his suspension, he migrated to Parler and alleged that he was targeted by a left-wing mob on Twitter [96].

From a platform policy standpoint, it has been observed that certain users, upon facing suspension, actively seek out platforms that offer more lenient policies regarding permissible topics or

content. For example, Robert W. Malone (@RWMaloneMD) was suspended in late 2021 for sharing misleading information about COVID-19. In an interview, he argued that *"what the media does not understand is that you can't suppress information. It'll find a way to be free"*, and subsequently migrated to Gettr:

> "We've had all this suppression, and yet we persist... People don't tolerate this type of censorship and propaganda that is being pushed on us [80, 115]."

In rare instances, suspended users may resort to migrating to other platforms in order to prevent impersonation by fake accounts, thereby mitigating potential confusion and related issues for the public. For example, Pikachu(@TruelndoIogy) took a handle designed to look like a user named "True Indology", whose real handle was @tiinexile. This fake account posted death threats against the real account holder and remained active even after the real account was suspended [53]. These incidents highlight the significant influence migration objectives can have on users prior to their actual transition to alternative platforms.

4.3.2 Costs of migration. In the context of user migration to new platforms, the perceived costs associated with the transition play a crucial role in shaping users' decisions. Our analysis has identified three primary types of perceived costs that users consider. The first type of cost is the cost of learning. When contemplating a migration, users must assess various factors, including the likelihood of being permanently suspended, strategies for maintaining influence, and the process of rebuilding social networks. For instance, the case of Marjorie Taylor Greene (@mtgreenee) highlights the issue of suspension, as she was banned from Twitter for repeatedly violating the platform's COVID-19 misinformation policy [4]. As a result, she announced her intention to migrate to the conservative platform Gettr, where she would be among like-minded individuals [38], and the "free speech" platform Gab, where there is a lower likelihood of suspension [95].

The second type of cost associated with user migration pertains to the changing affordances of the platforms themselves. Different platforms may possess distinct attributes or affordances that can act as barriers for users seeking to migrate, influencing their choice of alternative platforms. Factors such as audience size, interface design, and functionality play a significant role in shaping users' decisions. For example, if a user places a high value on audience size, they will likely remain on mainstream social media platforms. The case of the account Politics for All (@PoliticsForAll) illustrates this point, as it was suspended from Twitter for violating the platform's policies on manipulation and spam. The account, which provided a "one-stop rapid aggregation service of mainstream outlets with limited context" had amassed over 450,000 followers and received more than 80 million views per month prior to its suspension [27, 131]. Given that this account's success was closely tied to the huge audience base of mainstream platforms, the account shifted its focus to posting news to Instagram, another mainstream platform with design attributes — one-directional following mechanisms and broadcast-style posting — that are similar to Twitter's [131].

The final type of cost associated with migration is the effort required to rebuild one's audience. Even if users are familiar with the new platforms and the platforms themselves are suitable, they must still rebuild their entire social network and content from scratch. Gettr, for example, promoted the ability to import content from Twitter. However, Twitter later disabled this feature, making it more difficult for users to establish themselves on Gettr [2, 79]. In our analysis, we found that these three elements play a crucial role in influencing suspended users' decisions on where to migrate and in the overall process of operating their accounts. Overall, it seems that the costs of migration have a significant impact on the success of the migration. Users with sufficient motivation to migrate may choose to do so, but we identified a number of factors that may shape users' migration choices and strategies.

4.4 Factors influencing migration

4.4.1 Preparation: proactive V.S. reactive. We found that many users in our corpus included links to other social media accounts on their Twitter profiles, allowing their audience to find them on other platforms. These links could include references to personal websites or social media landing pages such as Linktree, which in turn connect to other platforms. The two most common types of links found were to users' personal websites (60.12%) and Facebook (29.45%). However, only 6.19% of users who migrated to other platforms had zero outlinks in their profiles. By contrast, 79.17% of users who did not migrate to other platforms did not include any outlinks in their profiles. These findings suggest that users who incorporate links to other social media accounts in their Twitter profiles may already be active on multiple platforms, indicating a broader online presence beyond just Twitter. Such links could also signify users' proactive steps to anticipate or prepare for potential suspensions, aiming to ensure continuity in audience connectivity. Conversely, users who do not include outlinks in their profiles may demonstrate a stronger commitment to Twitter as their primary platform and show a lower propensity for migration to other social media platforms. This highlights the potential influence of profile outlinks on users' online behavior and platform preferences. It is worth noting that by using outlinks such as Linktree, users can provide their followers with a comprehensive list of all the social media platforms they are using. This can be an effective way for users to successfully migrate across platforms and regain their audience.

One example is the official Twitter account belonging to the Supreme Leader of Iran, Ali Khamenei (@KhameneiSite) which was suspended for releasing an animation targeting former US President Donald Trump in revenge for the assassination of Qasem Soleimani in 2020, a content decision that Khamenei likely knew would be controversial. The profile of this account includes links to other social media accounts and a personal website [62, 71]. In contrast, user accounts that do not anticipate account suspension may find it challenging to guide their audiences to their new accounts, particularly when responding reactively to a suspension. For example, Li-Meng Yan (@LiMengYAN119), a Chinese virologist, was suspended for posting misleading COVID-19 information. This account was shut down within two days of its creation, leaving nearly no time for it to get prepared [84]. We could not find any messages this user left to guide her followers to other social media platforms.

4.4.2 Destination: mainstream V.S. alternative. Of the 143 suspended influential users included in our data, we found the average number of migration destinations for users was 2.79. In many cases, we observed that users tend to migrate to both mainstream and alternative social media platforms. Approximately 20% of users solely migrated to mainstream platforms, while nearly 10% solely migrated to alternative platforms. Users who exclusively migrated to alternative platforms had a tendency to have been suspended from multiple mainstream platforms in addition to Twitter. Moreover, we noticed that around 20% of suspended users either left social media altogether or created new identities that we did not identify as the same user (see more details in Table 3). We also analyzed the proportion of each specific destination among all alternative platforms, and found that three mainstream social media platforms: Facebook (39.2%), Instagram (38.5%), and YouTube (36.4%) were the most popular migration destinations. A significant number of users also chose to migrate back to Twitter (30.8%). Gab (23.8%), Telegram (23.1%), and Gettr (14.7%) followed the aforementioned four platforms in popularity.

We discovered intriguing trends for users who exclusively migrated to alternative platforms following their suspension. First, a subset of these users had been suspended from multiple mainstream platforms. This accumulation of suspensions seems to have influenced their decision to explore alternative platforms as a new avenue for maintaining an online presence. Second, we found cases where users announced their migration to less restrictive alternative platforms as a

Table 3. Permanently Suspended Influential Users' different migration destinations

Destinations	#	%
Mainstream social media platforms	29	20.28%
Alternative social media platforms	17	11.89%
Both mainstream and alternative social media platforms	71	49.65%
None	26	18.18%

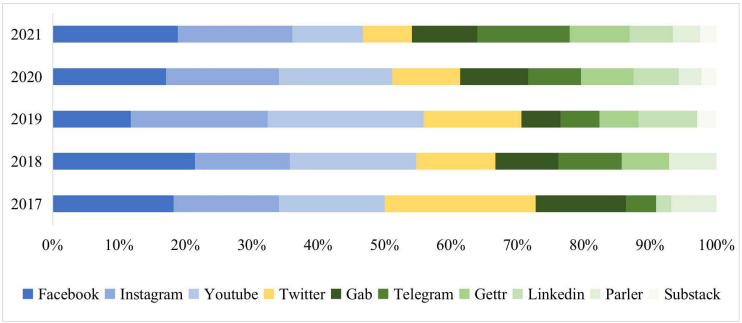

Fig. 5. Percentage of user's migration destinations for each alternative platform per year

form of protest against platform bans or restrictions. These users actively sought out platforms that aligned with their ideologies and positioned themselves as advocates for free expression. Third, users migrating solely to alternative platforms may have been motivated by a desire to find features and functionalities similar to those offered by mainstream platforms. They sought alternatives that could mirror their previous online experiences or provide similar social interactions. These findings highlight the complex motivations that drive users to exclusively migrate to alternative platforms. Factors such as multiple suspensions, ideological alignment, and platform features significantly influenced their choices. Understanding these diverse considerations provides valuable insights into user behavior and platform migration dynamics. Figure 5 further illustrates this point, showing a growing trend of users choosing a higher number of alternative platforms as migration destinations over time.

4.4.3 Reason: clear V.S. unclear. Our analysis revealed that influential users' understanding of the reason for their suspension significantly impacts their subsequent actions. Based on our thematic analysis, when influential users are unaware of why they were suspended, they create a new account on the original platform at higher rates. Our data show that when the reasons for suspension are clear, 31.78% of users return to Twitter, whereas in cases where the reasons for suspension are unclear, 64.29% of users return to Twitter. Additionally, if a user disputes the reason for their suspension, we found that they are more likely to attempt to return to Twitter. This can be observed in the case of Daniel Sieradski (@selfagency), who was suspended in 2017 for multiple violations, but believed the ban could have also resulted from a campaign of harassment by a right-wing user [67]. This case highlights that even when platforms provide some rationale for content moderation but don't fully disclose all reasons, it can create room for rumors, confusion, or even intentional misinterpretation of the suspension motives. It is worth noting that circumventing a ban on Twitter violates the platform's policies. If users disagree with or do not understand the reasons for a non-permanent suspension, they may try to evade it, leading to a possible permanent suspension.

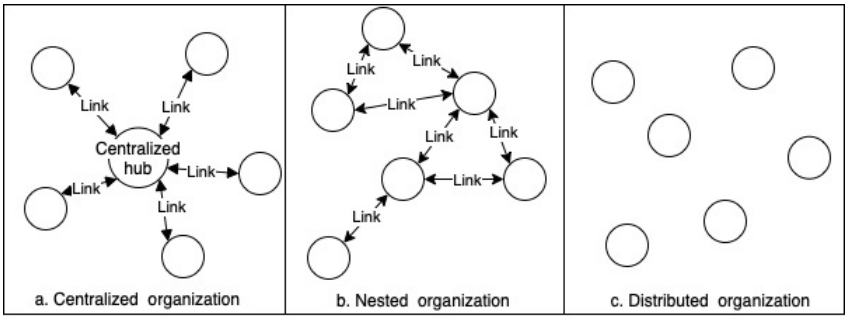

Fig. 6. Three kinds of platform organization. The circle represents a social media platform, and the line between platforms indicates that the user has connected two sites through a link.

Communicating the reasons for suspensions clearly may facilitate users and the public accepting the suspension, and potentially prevent future suspensions.

4.4.4 Platform organization: centralized, nested and distributed. There is a variety of methods by which the influential users in our news corpus organize the links between their various social media accounts. Based on the migration table we created (Figure 3), we analyzed the distribution and direction of outlinks on different platforms and identified three distinct organizational structures: centralized, nested, and distributed. In the centralized arrangement, all social media references are linked to a single location, such as a personal website or Linktree (Figure 6.a). Nearly all of the suspended influential users in our data utilize their personal websites as the central hubs for their social media accounts, allowing their followers to potentially follow them to their new migration locations. For instance, Ed Krassenstein (@EdKrassen) includes links to all of his social media accounts on his homepage. In the nested arrangement, users' social media accounts on different platforms link to one another, producing a "nesting" or chain structure (Figure 6.b). 5.6% of users organize their platforms in this pattern. Due to the fact that references to other platforms vary based on the account profile in question, audiences will be sent from one platform to another as opposed to a central destination. The final structure is a distributed one, in which users neither centralize a list of other social media profiles nor nest linkages between platforms (Figure 6.c). 4.2% of users arrange their platforms using this structure. In this strategy, accounts belonging to the same user are not linked together, even though they are run by the same person.

4.4.5 Time: multiple V.S. once. Based on the account creation dates at users' migration destinations, it appears that suspended influential users usually establish new accounts sequentially, rather than in bulk. This behavior is typical amongst users who are banned from mainstream social media platforms and subsequently make the move to alternative ones. The time and effort needed to build a following often discourage users from migrating to multiple alternative platforms all at once. However, there are exceptions where a small fraction of users do decide to migrate to several alternative platforms simultaneously. An illustrative case is Anjem Choudary (@anjemchoudary_), who was suspended from Twitter due to policy violations related to violent organizations and individuals [134]. Despite numerous unsuccessful attempts to reestablish a presence on Twitter, he countered further suspensions from various popular sites by concurrently setting up new accounts on Telegram, Snapchat, Pinterest, and Odysee.

4.4.6 Accounts types: individual V.S. organizational. Our data indicate that some influential users suspended from Twitter try to return by creating new accounts or leveraging associated accounts,

such as organizational or official profiles. Among the fifty users who tried to migrate back to Twitter, only four of them came back with their original accounts (8%), eleven of them were suspended again (22%), and twenty-one of them returned with their organizational or official accounts (42%). This suggests that official or organizational accounts may have a lower risk of being suspended. For example, Marjorie Taylor Greene, a Congresswoman representing Georgia's 14th congressional district, had her personal account suspended for multiple violations of Twitter's COVID-19 misinformation policy [4]. However, her official congressional account was not suspended. Additionally, we have observed a trend in which users who have founded an organization tend to use organizational accounts, rather than individual accounts, after being suspended in order to remain active on Twitter. This can be seen in the examples of Anthony Cumia and Meghan Murphy, both of whom have been suspended for violating the platform's rules. Anthony Cumia (@AnthonyCumia) was suspended in 2017 for abusive behavior. He then created several accounts to get around the initial suspension such as @CompoundBoss and @BoeingMax8, which were suspended as well [112]. However, his professional show accounts, @CompoundAmerica and @TheCumiaShow, both remain active, and are primarily used to promote shows. Meghan Murphy (@MeghanEMurphy) and Feminist Current are another example. Meghan Murphy was suspended for violating rules about hateful conduct in 2018 [75]. The organizational account sqlulz (@FeministCurrent) was created in 2021 and remains active on Twitter as of the time of this writing. However, compared to the 24,004 followers of her individual @MeghanEMurphy account, immediately before suspension, the organizational account only has 103 followers as of this writing.

4.4.7 Occupation. Finally, we found that influential users' professions can play a role in their cross-platform migrations. Politicians whose Twitter accounts have been suspended often use their personal websites as hubs for their social media presence (83.33%), listing all of their social media accounts on their websites. Additionally, if they are suspended from mainstream platforms as well as crowdfunding channels such as PayPal and GoFundMe, these politicians often turn to their personal websites as a means of receiving donations. Similarly, as previously mentioned, individuals who lead public organizations tend to migrate to their organizational accounts in order to continue promoting their cause. Furthermore, there are various professional platforms that can be used by those who are proficient in a particular field. For example, writers or journalists may use subscription platforms such as Substack to continue publishing their work.

4.5 Migration modes

The fifth theme in our results is the ways in which influential users migrate. Drawing on the analysis provided in Section 4.4, we noticed some persistent migration patterns or "migration modes" — characteristic behavioral patterns that suspended influential users exhibit throughout their migration journey. These modes offer a nuanced understanding of how these individual factors are intertwined, culminating in unique strategies for user migration. A visualization of these migration modes, represented as an interplay of various factors, can be found in Figure 7.

One of the notable migration modes is the "Strategic Mode," in which users combine proactive preparation, multiple migration instances, and nesting of social media accounts. Unlike simply identifying proactive preparation as a factor, this may indicate an intentional plan, by creating a network of interlinked accounts across platforms. For instance, Shiva Ayyadurai's case illustrates this mode, where he constructed a well-knit presence across various platforms before his suspension in 2017, potentially indicating a premeditated and methodical approach [41].

We also identified a "Responsive Mode" which is characterized by reactive measures and centralized migration timing. This contrasts with the "Strategic Mode," as it typically involves users who were caught unawares by the suspension. They scramble to maintain their online presence by

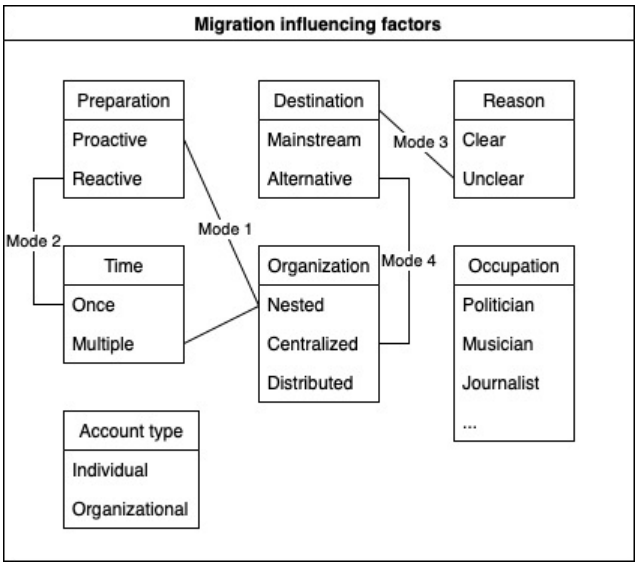

Fig. 7. The figure shows the conceptual model of different factors that can influence users' migration mode. The lines connecting different levels of factors indicate migration modes. Details about the four migration modes can be seen in Section 4.5.

immediately moving to alternative platforms. An illustration is Li-Meng Yan (@LiMengYAN119), who adapted rapidly to maintain her reach [84]. A third mode we observed is the "Diversified Mode". This mode is significant among influential users who receive vague or unclear reasons for their suspension. They seem to cast a wide net by branching out to both mainstream and alternative platforms, possibly to mitigate the risk of further suspensions.

Finally, the "Centralized-Alternative Mode" is noteworthy among users who mainly migrate to alternative platforms but centralize their presence by listing all social media accounts in one place. This may be intended as a pragmatic approach, where users concentrate on alternative platforms but still aim to reach wide audiences and provide easy access to their various online personas. Michael Flynn (@GenFlynn) exemplifies this mode, by having a consolidated presence through his personal website and Linktree page, after being suspended in 2020 [26].

4.6 Announcing migration

In the final phase of the migration process, we noted a crucial step: publicizing the new social media destination to followers. This step is not just pivotal for audience retention but also functions as a bridge, connecting influential users' new accounts with their original and potential audience. In this section, we dive deeper into the specific communication tactics leveraged by influential users to announce their migration.

The most prevalent communication strategy employed by influential users is the use of personal websites as a stable anchor, as described above. These sites frequently serve as a central repository for their social media links, effectively directing their audience towards other accounts held by the user. As previously mentioned, these users sometimes use their existing social media platforms or specialized services, such as Linktree, to create a consolidated hub for all their social media links.

An exemplar of this strategy is Jared Taylor (@jartaylor) and American Renaissance, which have amalgamated their social media links into American Renaissance's Bitchute channel[3].

However, it is important to note that these curated lists of social media links provided by users do not always encompass all of their accounts. Intriguingly, the platforms typically omitted from these lists are alternative ones. For instance, Naomi Wolf (@naomirwolf) excluded her Gab, Telegram, and LinkedIn accounts from her list, while Jim Hoft (@gatewaypundit) did not include his Telegram and Gab accounts. This selective omission is an interesting pattern that we think warrants future work.

Another common strategy we observed is the use of "traditional" media channels to announce migration. In these instances, the news of the migration is disseminated via media coverage. This was evident with Paul Golding (@GoldingBF), whose shift to Gab garnered widespread attention from various media outlets. Additionally, we identified a more implicit migration announcement strategy: maintaining username consistency across platforms. This strategy, while seemingly simple, effectively aids follower retention post-migration by making influential users easily discoverable across different platforms. However, its feasibility is dependent on username availability and can pose complications for pseudonymous users. Lastly, a proactive strategy that some influential users adopt is the "Pre-emptive Redirection" — embedding links to alternative social media accounts in their profiles before facing any suspension. This strategy differs from simply listing accounts. It is a strategic measure taken in anticipation of potential suspensions, ensuring a seamless transition for the audience, who are pre-informed of alternate points of contact.

5 DISCUSSION

Our work here explores the causes and consequences of permanent suspension on Twitter, with a particular eye to the process of migrating to alternative platforms and the consequences this has for the broader social media information landscape. We work to contextualize our results above, step back and discuss a number of key takeaways and associated opportunities for design and future work.

5.1 The Value of Clarity in Suspension Notifications

Previous research has underscored the advantages of transparency in content moderation, such as enhancing comprehension and acceptance of the moderation process. Transparency also serves an educational purpose by helping users grasp the platform's rules and expectations, thus motivating them to comply. For instance, Jhaver et al. [59] discovered that participants exhibited negative attitudes towards their content removals, and the absence of notifications exacerbated their dissatisfaction. This finding suggests that community rules and explanations for removals contribute to perceiving content moderation as fair.

Building upon recent calls by Jhaver et al. [59], Gill et al. [45], Suzor et al. [120], and others, we emphasize the significance of providing explanations for suspension decisions. Our research indicates that, particularly in the context of influential users, the general public may have an interest in suspension explanations too. Our findings demonstrate that even when a platform offers some information about the reasons behind content moderation, withholding certain details from the public can facilitate the spread of rumors or misrepresenting the reason for personal ends. In our data, influential users portrayed this lack of information as leaving them uninformed or creating suspicion of the platform's intentions. If this sense of suspicion is pervasive, either because of eroded trust in the platform or because influential people work to sow suspicion in their audiences, one consequence is an environment that is conducive to the proliferation of rumors, which may

[3]https://www.bitchute.com/channel/5Q4sa6rObtGx/

exacerbate issues of mis- and disinformation [117]. One path to potentially mitigating these risks is increasing the clarity, accuracy, and precision of suspension decisions, both for the account holder and potentially for the public. This may serve to help reduce confusion, uncertainty, and the dissemination of rumors. Others have suggested that failure to explain moderation reasons can lead to user confusion and uncertainty, cultivating an environment ripe for rumor propagation [59]. In the absence of clear and accurate information regarding moderated content or behaviors, users may resort to speculation or unverified sources of information, resulting in misinformation and false narratives. These consequences undermine trust in the platform and its moderation process.

Our results point to a key reason for valuing this kind of transparency — it may deter suspended influential users from trying to return to the same platform again. When users and the public receive clear and specific explanations for suspensions, we show that in the majority cases these influential users migrate to a different platform. This may indicate that being clear about suspension reasons helps users understand the infraction and accept the suspension as justified. Consequently, this reduces the likelihood that influential users attempt to circumvent suspensions by creating new accounts or returning under different identities. Several factors may influence the effectiveness of transparency in moderation, including the level of detail in the explanation, timing, presentation, and degree of disclosure. Future research could delve into these factors' impact on user migration and other outcomes related to trust and acceptance of moderation. Such insights would be valuable to platform designers seeking to optimize the clarity and transparency of their moderation processes.

In our study, we observed that rather than being suspended for a single specific reason, many influential users were found to have either repeatedly violated a single policy or committed several less severe rule violations before a final, more severe infraction led to their suspension. Some platform policies specify that users will only be suspended if they repeatedly violate them, serving as a buffer mechanism, which can serve a dual purpose. On one hand, it prevents the over-penalization of users who unintentionally violate a policy, allowing for a fair and measured response. On the other hand, it provides an opportunity to educate and warn users about the platform's guidelines. However, our analysis of the news corpus revealed numerous instances in which users were suspended for multiple violations of the same policy, suggesting that the buffer may not be effective for the purposes of education and mitigation of continuing infractions. This indicates the need for continuous evaluation and refinement of buffer mechanisms to effectively educate users and deter repeated violations.

There may be alternative designs to a similar buffer mechanism that more effectively strike a balance between accommodating users who unintentionally violate policies and effectively addressing deliberate, antisocial behavior. Platforms might consider personalizing how they communicate policy violations and which techniques they use to help educate people who unintentionally violate policies. However, not all users who violate multiple policies do so inadvertently. A more tiered system of consequences for policy violations, or mandating further education about platform expectations before allowing a user to rejoin, might serve to curtail intentional violations. Our study suggests a tension between affording leniency to well-intentioned users and effectively preventing harmful behavior, which we see as a key design consideration. Successfully distinguishing between accidental infractions made in good faith and deliberate, antisocial conduct presents a significant challenge [46]. However, if platforms can reliably make this distinction, it could pave the way for more personalized and effective content moderation responses.

5.2 The Significance of Prioritizing User Migration

When platforms' permanently suspend users, frequently the action a platform takes stops after permanent suspension. In one way, this is intuitive, the company running the platform only controls their platform — their primary goals are to enforce their own policies, create a safe environment

within their own ecosystem, and focus on engaging and retaining users on their own sites. Of course, what a user does after being suspended is a difficult thing for a platform to intervene on, and users might rightfully have surveillance and privacy concerns if platforms were seeking to extend their influence beyond the point of suspension, and platforms may not even legally have the authority to pursue such goals.

Our work here complicates this issue, however. After all, a permanent suspension is not the end of an influential user's presence online. When platforms ban users for violating community guidelines or engaging in harmful behavior, the successful migration of these banned users to alternative platforms may serve to undermine the perception of the banning platform's effectiveness in maintaining a safe and healthy environment. Moreover, if suspended users migrate to rival platforms, it can bring about shifts in user demographics or content trends. Through one lens, this is the goal — suspending people who are spreading hateful and misleading content from your platform reduces hateful and misleading content on your platform. It may also reduce the number of people who participate in the platform primarily because of the suspended account.

Our results show that in many cases, influential users who have been permanently suspended work hard to retain their public visibility and audience, either by working to return to the original platform, or by increasing their posting rates and working to regrow their audience elsewhere. Because permanent suspension from one platform is not the end of a user's presence online, permanent suspension does not end the spread of hateful or misleading content, it only ends the spread of hateful or misleading content on one platform. Our results suggest that permanent suspension *shifts* the spread of this kind of content and the growth of these audiences elsewhere. In other words, this content is not *removed from* the social media information landscape, merely *moved around* the social media information landscape.

Moreover, while shifting harmful or misleading content to alternative, less popular platforms may help address how this content affects audiences, it does not necessarily address how else this content may be used. We see it as likely that there may always be audiences for such content, and this continued interest may not be something that can be solved by a technology platform. However, AI systems are being trained on social media content today [129], and social media platforms are choosing to charge very high rates for access to their content through APIs, essentially blocking AI companies from leveraging the content on their sites [57, 118]. Again, this is somewhat intuitive, companies may not be able to easily control what happens outside the purview of the platform they run.

However, we see cause for concern stemming from social media companies that focus solely on their own platforms, rather than the social media information landscape more broadly. That is, shifting harmful or toxic influential users elsewhere in the social media information landscape — through permanent suspension — and subsequently closing off the content on social media platforms from use as training data, creates a potentially dire situation. The consequences of permanently suspending influential users ripple across the information landscape, rather than removing them from it. That same information landscape is used as training data for AI companies, and shifting harmful and misleading information off of large, increasingly closed platforms, and into more readily accessible places for gathering training data risks unduly weighting this harmful and misleading content in the data being used for AI systems like LLMs [40].

Importantly, we do not advocate that platforms merely allow people who spread harmful or misleading content to remain on their platforms. Moreover, we do not suggest that this means that AI companies ought to have unfettered access to social media platforms. Rather, we see it as critical that social media platforms shift their focus away from exclusively their own platforms, towards a more global view of the broader social media information landscape. What might it mean

to consider the costs of permanent suspension more broadly? Could platforms learn to recognize migration and federate suspension decisions?

5.3 Migration is Common, Despite Potential Barriers

In this study, we informed our analysis in part through the lens of social media ecologies, because permanent suspension does not remove people from the wider social media information landscape, only from one platform. Our findings complement previous research, which has highlighted that influential users tend to migrate to other platforms after being banned. [5, 20, 101]. Moreover, our work helps to fill a gap in prior work by demonstrating that migration is a common phenomenon that can have significant implications for platform usage and user engagement. Our findings suggest that social media platforms should take into account the impact of permanent suspensions on other platforms, and the ways in which these suspensions may affect the overall internet ecosystem. While previous research has evaluated the effectiveness of permanent suspension on the original platform [19, 60], and some studies have attempted to understand user behavior changes on an alternative platform after being permanently suspended [5]. Our findings of common migration patterns suggest that future evaluations of the effectiveness of permanent suspension should take a holistic view of users' social media ecology, consider users' migration maps, and include users' migration routes and behavioral changes as evaluation metrics, in order to fully comprehend the full range of consequences of this moderation strategy.

Moreover, our results indicate that most influential users choose to migrate in response to suspension, suggesting that, despite the potential loss of audience or other detrimental consequences, most influential users deem the benefits of migration to outweigh the downsides [101]. In recent years, the emergence of alternative platforms, in some cases specifically designed as a replacement for certain users, may help to minimize the costs of migration by facilitating the transition process and reducing the learning curve for users. In fact, some alternative platforms have large user bases of their own, which can also help minimize the risk of audience loss, depending on the degree of overlap in audiences between platforms [139]. Our findings also reveal that a significant number of influential users link to different social media platforms from their profiles, seemingly in preparation for potential suspension, which may further mitigate the cost of audience loss. Given these conditions, it seems that users commonly see migration as a worthwhile endeavor after being permanently suspended, enabling them to continue engaging in effective self-presentation, reaching and communicating with their imagined audiences, all while potentially benefiting from more relaxed platform norms in some cases.

Consequently, our work emphasizes the importance of understanding and addressing the contextual factors that contribute to user migration after permanent suspension. We view our results as beginning to characterize the contextual behavioral patterns of cross-platform migrations, in order to inform the design of alternative barriers to easy cross-platform migration. We see continuing to understand the contexts of migration, the mechanisms that influential users leverage to prepare for migration, and the influential sociotechnical approaches that ease or create barriers to migration as important directions for continued research.

5.4 Mitigating Cross-Platform Migration Harms

Our results suggest that an important technique for migrating to other platforms seems to be leaving a train of nested links to other platforms across multiple social media accounts. This proactive approach seems intended to help audiences follow influential users to new platforms, in case the account is suspended. In addition, we also find that migration rates to alternative platforms are gradually increasing. Indeed, we find that influential users make proactive preparations create new accounts on other platforms one by one, actively nesting links to other platforms in their social

media profiles, seemingly with the goal of having a ready-made audience on a new social media account shortly after being suspended. These findings, taken together, suggest a design opportunity as well. It may be possible to detect, and intervene on, this behavior. For instance, it may be fruitful to simply consider disabling links in user profiles for users who have a track record of violating platform policies. This would serve as a way to limit the ability to prepare for a suspension by preemptively pointing audiences elsewhere. More broadly, computationally recognizing these modes of migration would enable platforms to predict users' migration trajectories before and after suspension, which might enable a more cohesive cross-platform approach to suspension, and may help reduce the harms of suspended users migrating toxicity to other platforms. Future research could also investigate the possibilities of collaboration among different social media platforms. This could include sharing information of problematic users and content or establishing similar content policies to prevent the transfer of issues from one platform to another. We also see other potential low-hanging fruit interventions that increase barriers to migration while a person is still on a platform. For instance, limiting the ability to export data from one platform to another may serve as a migration hindrance. Of course, it also limits the flexibility of what non-migrating, non-policy-violating users could do with their data if applied too broadly, and applying such a policy universally may even be illegal in some jurisdictions.

5.5 Eradicating Harmful Content: Moving Beyond Permanent Suspension

Our findings illuminate the intricate nature of user migration strategies and announcement methods, particularly among influential users. Importantly, we observe that permanent suspension, as a content moderation measure, can inadvertently facilitate the dissemination of harmful content, thereby undermining the effectiveness of content moderation. Our study uncovers several migration strategies employed by influential users, as well as their methods of announcing such migrations. Future research should further analyze the consequences of these migrations on the communities that influential users abandon, the new communities they integrate into, and the possibility of exacerbating harmful content in different environments. This trajectory of research is likely fruitful for the design of strategies to curb the adverse effects of such migrations. One interesting facet of this trajectory is also focusing on the psychological and societal ramifications of these migrations, especially concerning influential users, which may help develop additional insight into how these migrations sculpt online communities, discourse, and the broader social media information landscape.

Additionally, it is imperative to consider content moderation within a broader context. Previous research has predominantly categorized permanent suspension as a punitive measure and acknowledged the inherent limitations of such punitive content moderation approaches. As an alternative, some scholars have advocated for educational strategies to encourage improved user behavior and adherence to community guidelines [59, 85]. Moreover, researchers have recommended that platforms and community managers establish transparent guidelines, provide explanations for content removal, and utilize automated tools [45, 61, 68, 78, 108, 120]. Notably, Kou [70] accentuated the negative effects of permanent suspensions within gaming communities, instead advocating for a more restorative approach that accounts for context, relationships, and the specific circumstances of individual cases as a superior strategy in mitigating toxicity. Others have proposed the adoption of restorative justice principles or subsidiarity as a means of achieving a balanced approach to context and content management in online communities [52, 137]. We see ideas like restorative justice and related principles as a compelling alternative vision for the social media information landscape. Engaging users in a discourse that emphasizes the ramifications of their content, fostering reflection, and facilitating resolutions or reparations may present a more fruitful approach compared to immediate punitive measures, and may even diminish users inclinations to migrate off of the

platform in question. Furthermore, it is essential to recognize cultural and regional heterogeneity in content moderation and user behavior. The application of a universal content moderation strategy across diverse cultures and regions is inherently fraught with challenges, many of which arise from a lack of sensitivity to local and cultural contexts [46, 113, 126]. Future research on culturally- and regionally-contextual content moderation is likely fruitful, and points to another setting in social computing systems where local knowledge and context are likely critical for success [66].

5.6 Ethical Considerations of our Work

Our work here relies on second-hand news reports and archival collections of social media content on the web that has been removed by the platforms themselves. Using this type of data for research warrants a high level of sensitivity and ethical interrogation — after all, this kind of content is usually removed for violating community guidelines, and accessing such content through the web archives may raise questions about privacy and ethical research practices. Here, we reflect on and unpack the ethical considerations in our work.

In our work, we intentionally focused on the behavioral migration patterns of the influential users post-suspension, rather than delving deeply into the content itself, that led to their suspension. Moreover, we operationalized the news articles in our corpus as reports on suspension events and collected basic information about suspended accounts through the Wayback Machine. By focusing on these event-driven and behavioral metrics, we sought to avoid the dissemination of sensitive, removed content, and to maintain a non-judgmental perspective to ensure that our work does not contribute to the further spread of the problematic content.

Broadly, as researchers, we pursue research trajectories and topics that serve a greater public interest, and we see studying our study here as reflective of those goals. By studying the migration patterns of influential users after permanent suspensions, we hope to inform the design of more holistic and effective content moderation strategies to benefit both specific online communities, individual platforms, and the broader social media information landscape.

6 CONCLUSION

In this paper, we explore the consequences of permanently suspending influential users on Twitter and evaluate how suspension affects users' behavior within their own social media ecologies and across the social media information landscape. We apply thematic analysis to explore the context around and consequences of influential users' permanent suspension. Because permanent suspension signifies the end of a user's online presence on that platform, migration is a common reaction, and migration to alternative platforms is growing over time. We also explore variables that can influence users' migration modes and observe several kinds of migration modes. This observation suggests an opportunity to proactively identify users' migration trajectories, thus helping platforms make permanent suspension decisions and reduce the harms of migration.

ACKNOWLEDGMENTS

Partial support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison with funding from the Wisconsin Alumni Research Foundation.

REFERENCES

[1] Alberto Abadie and Javier Gardeazabal. 2003. The economic costs of conflict: A case study of the Basque Country. *American economic review* 93, 1 (2003), 113–132.

[2] Brain Adam. 2021. *Setback for Gettr:Can no longer import Twitter Content.* Retrieved July 12, 2022 from https://voonze.com/setback-for-gettr-can-no-longer-import-twitter-content/

[3] Kayode Sakariyah Adewole, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. 2017. Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications* 79 (2017), 41–67.

[4] Davey Alba. 2022. *Twitter Permanently Suspends Marjorie Taylor Greene's Account.* Retrieved July 12, 2022 from https://www.nytimes.com/2022/01/02/technology/marjorie-taylor-greene-twitter.html

[5] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*. 187–195.

[6] Virginia Allen. 2021. *New Social Media Platform Gettr Says No to Cancel Culture, Yes to Free Speech.* Retrieved July 12, 2022 from https://www.dailysignal.com/2021/07/09/new-social-media-platform-gettr-says-no-to-cancel-culture-yes-to-free-speech/

[7] Marvin Ammori. 2014. THE" NEW"" NEW YORK TIMES": FREE SPEECH LAWYERING IN THE AGE OF GOOGLE AND TWITTER. *Harvard Law Review* 127, 8 (2014), 2259–2295.

[8] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. 2019. Measuring the Facebook advertising ecosystem. In *NDSS 2019-Proceedings of the Network and Distributed System Security Symposium*. 1–15.

[9] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* (2021), 1–18.

[10] Enrique Armijo. 2021. Reasonableness as Censorship: Section 230 Reform, Content Moderation, and the First Amendment. *Fla. L. Rev.* 73 (2021), 1199.

[11] Joseph B Bayer, Penny Triu, and Nicole B Ellison. 2020. Social media elements, ecologies, and effects. *Annual review of psychology* (2020).

[12] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics*. Springer, 405–415.

[13] Danielle Blunt, Ariel Wolf, Emily Coombes, and Shanelle Mullin. 2020. Posting into the void: Studying the impact of shadowbanning on sex workers and activists. *Retrieved September* 6 (2020), 2021.

[14] Pablo J Boczkowski, Mora Matassi, and Eugenia Mitchelstein. 2018. How young users deal with multiple platforms: The role of meaning-making in social media repertoires. *Journal of computer-mediated communication* 23, 5 (2018), 245–259.

[15] Leticia Bode and Emily K Vraga. 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication* 65, 4 (2015), 619–638.

[16] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[18] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet* 7, 2 (2015), 223–242.

[19] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–26.

[20] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[21] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of blocked community members: Redemption, recidivism and departure. In *The world wide web conference*. 184–195.

[22] Danielle Keats Citron. 2014. *Hate crimes in cyberspace*. Harvard University Press.

[23] Danielle Keats Citron and Benjamin Wittes. 2017. The internet will not break: Denying bad samaritans sec. 230 immunity. *Fordham L. Rev.* 86 (2017), 401.

[24] Danielle Keats Citron and Benjamin Wittes. 2018. The Problem Isn't Just Backpage: Revising Section 230 Immunity. (2018).

[25] Jane Coaston. 2018. *YouTube, Facebook, and Apple's ban on Alex Jones, explained.* Retrieved August 6, 2018 from https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-youtube-conspiracy-theories

[26] Ben Collins and Brandy Zadrozny. 2021. *Twitter bans Michael Flynn, Sidney Powell in QAnon account purge.* Retrieved July 12, 2022 from https://www.nbcnews.com/tech/tech-news/twitter-bans-michael-flynn-sidney-powell-qanon-account-purge-n1253550

[27] Marie L Conte. 2022. *The Inside Story of Politics for All, Twitter's Infamous UK News Account.* Retrieved July 12, 2022 from https://www.vice.com/en/article/jgmjex/what-happened-to-politics-for-all
[28] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.
[29] Colleen Cronin. 2021. *Fake news accounts claim Taliban executed CNN journalist.* Retrieved July 12, 2022 from https://www.dailydot.com/debug/fake-news-accounts-tweet-death-fake-journalist-afghanistan/
[30] Tim Cushing. 2019. *Court Says Section 230 Shields Twitter From Revenge Porn Bro's Stupid Lawsuit.* Retrieved July 12, 2022 from https://www.techdirt.com/2019/06/13/court-says-section-230-shields-twitter-revenge-porn-bros-stupid-lawsuit/
[31] Julia R DeCook. 2022. r/WatchRedditDie and the politics of Reddit's bans and quarantines. *Internet Histories* 6, 1-2 (2022), 206–222.
[32] Michael A DeVito, Jeremy Birnholtz, and Jeffery T Hancock. 2017. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 740–754.
[33] Michael A DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. 'Too Gay for Facebook' Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.
[34] Judith S Donath. 2002. Identity and deception in the virtual community. In *Communities in cyberspace*. Routledge, 37–68.
[35] Rónán Duffy. 2020. *Twitter permanently suspends Gemma O'Doherty's account over 'repeated violations'.* Retrieved July 12, 2022 from https://www.thejournal.ie/gemma-odoherty-twitter-5165384-Jul2020/
[36] Emory James Edwards and Tom Boellstorff. 2021. Migration, non-use, and the 'Tumblrpocalypse': Towards a unified theory of digital exodus. *Media, Culture & Society* 43, 3 (2021), 582–592.
[37] Oliver Effron. 2020. *David Duke has been banned from Twitter.* Retrieved July 12, 2022 from https://www.cnn.com/2020/07/31/tech/david-duke-twitter-ban/index.html
[38] Greg Evans. 2021. *This Twitter account documents all the madness you're missing on Parler.* Retrieved July 12, 2022 from https://www.indy100.com/science-tech/parler-twitter-account-trump-right-wing-b1784844
[39] Gabriel Fair and Ryan Wesslen. 2019. Shouting into the void: A database of the alternative social media platform gab. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 608–610.
[40] Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).
[41] Jenna Fisher. 2021. *Twitter Suspends Senate Candidate Shiva Ayyadurai's Account.* Retrieved July 12, 2022 from https://patch.com/massachusetts/cambridge/twitter-suspends-senate-candidate-shiva-ayyadurais-account
[42] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on Facebook and Twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
[43] R Stuart Geiger. 2016. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society* 19, 6 (2016), 787–803.
[44] Samuel Gibbs and Martin Belam. 2017. *Twitter suspends Britain First leaders as it enforces new anti-abuse rules.* Retrieved July 12, 2022 from https://www.theguardian.com/technology/2017/dec/18/twitter-enforcing-anti-abuse-rules-clean-up-act-abuse-hate-symbols-sexual-advances-violent-groups
[45] Lex Gill, Dennis Redeker, and Urs Gasser. 2015. Towards digital constitutionalism? Mapping attempts to craft an internet bill of rights. *Mapping Attempts to Craft an Internet Bill of Rights (November 9, 2015). Berkman Center Research Publication* 2015-15 (2015).
[46] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.
[47] João Gonçalves, Ina Weber, Gina M Masullo, Marisa Torres da Silva, and Joep Hofhuis. 2021. Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion. *new media & society* (2021), 14614448211032310.
[48] Jazmin Goodwin. 2021. *Gab: Everything you need to know about the fast-growing, controversial social network.* Retrieved July 12, 2022 from https://www.cnn.com/2021/01/17/tech/what-is-gab-explainer/index.html
[49] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[50] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N Asokan. 2018. All you need is" love" evading hate speech detection. In *Proceedings of the 11th ACM workshop on artificial intelligence and security*. 2–12.
[51] Richard Hanania. 2019. *It Isn't Your Imagination: Twitter Treats Conservatives More Harshly Than Liberals.* Retrieved June 24, 2022 from https://quillette.com/2019/02/12/it-isnt-your-imagination-twitter-treats-conservatives-more-harshly-than-liberals/l

[52] Amy A Hasinoff and Nathan Schneider. 2022. From Scalability to Subsidiarity in Addressing Online Harm. *Social Media+ Society* 8, 3 (2022), 20563051221126041.
[53] HinduPost. 2020. *Twitter targeting TrueIndology for sharing inconvenient truths?* Retrieved July 14, 2022 from https://hindupost.in/media/twitter-hounds-true-indology-for-sharing-inconvenient-truths/
[54] Matthew P Hooker. 2019. Censorship, Free Speech & Facebook: Applying the First Amendment to Social Media Platforms via the Public Function Exception. *Wash. JL Tech. & Arts* 15 (2019), 36.
[55] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.
[56] Twitter Inc. 2021. *Permanent suspension of @realDonaldTrump.* Retrieved January 8, 2021 from https://blog.twitter.com/en_us/topics/company/2020/suspension
[57] Mike Isaac. 2023. *Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems.* Retrieved July 10, 2023 from https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html
[58] Shagun Jhaver. 2020. *Identifying opportunities to improve content moderation.* Ph. D. Dissertation. Georgia Institute of Technology.
[59] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. " Did you suspect the post would be removed?" Understanding user reactions to content removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–33.
[60] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.
[61] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
[62] Dustin Jones. 2021. *Twitter Bans Account Linked To Iran's Supreme Leader.* Retrieved July 12, 2022 from https://www.npr.org/2021/01/22/959736537/twitter-bans-account-linked-to-irans-supreme-leader
[63] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. 2021. Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21).* 1163–1180.
[64] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.
[65] Ayushman Kaul. 2020. *Telegram — a free speech Russian platform is a haven for far-Right terror groups.* Retrieved July 12, 2022 from https://theprint.in/opinion/telegram-a-free-speech-russian-platform-is-a-haven-for-far-right-terror-groups/407357/
[66] David A Kaye. 2019. *Speech police: The global struggle to govern the Internet.* Columbia Global Reports.
[67] Sam Kestenbaum. 2017. *'Antifa's Most Prominent Jew' Booted From Twitter.* Retrieved July 12, 2022 from https://forward.com/fast-forward/374276/antifas-most-prominent-jew-booted-from-twitter/
[68] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* 1 (2012), 4–2.
[69] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.
[70] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–21.
[71] Zaira Lakhpatwala. 2021. *Why won't Twitter ban Khamenei when it permanently suspended Trump?* Retrieved July 12, 2022 from https://www.arabnews.com/node/1829296/media
[72] Cliff Lampe, Paul Zube, Jusil Lee, Chul Hyun Park, and Erik Johnston. 2014. Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. *Government Information Quarterly* 31, 2 (2014), 317–326.
[73] Allen Yilun Lin, Kate Kuehl, Johannes Schöning, and Brent Hecht. 2017. Understanding" Death by GPS" A Systematic Study of Catastrophic Incidents Associated with Personal Navigation Technologies. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* 1154–1166.
[74] Siân E Lindley, Catherine C Marshall, Richard Banks, Abigail Sellen, and Tim Regan. 2013. Rethinking the web as a personal archive. In *Proceedings of the 22nd international conference on World Wide Web.* 749–760.
[75] Julia Manchester. 2018. *Self-described feminist banned from Twitter says platform is setting 'dangerous' precedent.* Retrieved July 12, 2022 from https://thehill.com/hilltv/rising/420033-self-described-feminist-banned-from-twitter-says-platform-is-setting-a/
[76] Sarah Marsh. 2019. *Twitter blocks accounts of Raul Castro and Cuban state-run media.* Retrieved July 12, 2022 from https://www.reuters.com/article/us-cuba-twitter/twitter-blocks-accounts-of-raul-castro-and-cuban-state-run-

media-idUSKCN1VX2AH
[77] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.
[78] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.
[79] Emma Mayer. 2021. *Gettr CEO Says Twitter Blocking Tweet Imports, Suggests That Makes It a Publisher.* Retrieved July 12, 2022 from https://www.newsweek.com/gettr-ceo-says-twitter-blocking-tweet-imports-suggests-that-makes-it-publisher-1608577
[80] Bill McCarthy. 2020. *Who is Robert Malone? Joe Rogan's guest was a vaccine scientist, became an anti-vaccine darling.* Retrieved July 12, 2022 from https://www.politifact.com/article/2022/jan/06/who-robert-malone-joe-rogans-guest-was-vaccine-sci/
[81] Paul Mena. 2020. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & internet* 12, 2 (2020), 165–183.
[82] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*. 85–94.
[83] Patricia L Moravec, Antino Kim, and Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* 31, 3 (2020), 987–1006.
[84] Jason Murdock. 2020. *Twitter Suspends Account of Chinese Virologist Who Claimed Coronavirus Was Made in a Lab.* Retrieved July 12, 2022 from https://www.newsweek.com/twitter-suspends-dr-li-meng-yan-wuhan-lab-coronavirus-covid19-1532193
[85] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[86] Peter Nagy and Gina Neff. 2015. Imagined affordance: Reconstructing a keyword for communication theory. *Social Media+ Society* 1, 2 (2015), 2056305115603385.
[87] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*.
[88] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.
[89] Fayika Farhat Nova, Michael Ann DeVito, Pratyasha Saha, Kazi Shohanur Rashid, Shashwata Roy Turzo, Sadia Afrin, and Shion Guha. 2021. " Facebook Promotes More Harassment" Social Media Ecosystem, Skill and Marginalized Hijra Identity in Bangladesh. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–35.
[90] Stanford Internet Observatory. 2001. *gogettr.* Retrieved May 15, 2005 from https://github.com/stanfordio/gogettr/
[91] Abby Ohlheiser and Ian Shapira. 2018. *Gab, the white supremacist sanctuary linked to the Pittsburgh suspect, goes offline (for now).* Retrieved July 12, 2022 from https://www.washingtonpost.com/technology/2018/10/28/how-gab-became-white-supremacist-sanctuary-before-it-was-linked-pittsburgh-suspect/
[92] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.
[93] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
[94] Billy Perrigo. 2023. *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic.* Retrieved July 10, 2023 from https://time.com/6247678/openai-chatgpt-kenya-workers/
[95] Ewan Plamer. 2022. *Marjorie Taylor Greene Gets Gab, Gettr Boost After Twitter Ban—Way Down Overall.* Retrieved July 12, 2022 from https://www.newsweek.com/marjorie-taylor-greene-twitter-ban-gab-gettr-boost-1667988
[96] Andrew Quinn. 2020. *Twitter ban former TUV election candidate David Vance over tweet sent to Manchester United striker Marcus Rashford.* Retrieved July 14, 2022 from https://www.newsletter.co.uk/news/politics/twitter-ban-former-tuv-election-candidate-david-vance-over-tweet-sent-to-manchester-united-striker-marcus-rashford-2967257
[97] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
[98] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[99] Kevin Rawlinson. 2018. *Tommy Robinson permanently banned by Twitter.* Retrieved July 12, 2022 from https://www.theguardian.com/technology/2018/mar/28/tommy-robinson-permanently-banned-twitter-violating-rules-hateful-conduct
[100] Adi Robertson. 2018. *TFormer revenge porn mogul Craig Brittain sues Twitter in absurdist censorship complaint.* Retrieved July 12, 2022 from https://www.theverge.com/2018/6/7/17437608/craig-brittain-revenge-porn-senate-

candidate-twitter-censorship-antitrust-lawsuit
[101] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* 35, 3 (2020), 213–229.
[102] Aja Romano. 2017. *Twitter suspended, then restored, the account of former KKK grand wizard David Duke.* Retrieved July 12, 2022 from https://www.vox.com/culture/2017/3/7/14831800/david-duke-twitter-ban-restored
[103] Mac Ryan and Montgomery Blake. 2018. *Twitter Suspended Proud Boys' And Founder Gavin McInnes' Accounts Ahead Of The 'Unite The Right' Rally.* Retrieved June 28, 2022 from https://www.buzzfeednews.com/article/ryanmac/twitter-suspends-proud-boys-and-founder-gavin-mcinnes
[104] Haji Mohammad Saleem and Derek Ruths. 2018. The aftermath of disbanding an online hateful community. *arXiv preprint arXiv:1804.07354* (2018).
[105] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6.
[106] Joseph Seering. 2020. Reconsidering Community Self-Moderation: the Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proc. ACM Hum.-Comput. Interact* 3 (2020).
[107] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong'Cherie' Chen, Likang Sun, and Geoff Kaufman. 2019. Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[108] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It takes a village: integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
[109] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation. In *Proceedings of the 10th ACM Conference on Web Science*. 265–274.
[110] Ahmad Shehabat, Teodor Mitew, and Yahia Alzoubi. 2017. Encrypted jihad: Investigating the role of Telegram App in lone wolf attacks in the West. *Journal of strategic security* 10, 3 (2017), 27–53.
[111] Spencer Silva. 2021. *Racist troll Owen Benjamin is evading bans from major social media platforms to continue spreading hate and bigotry.* Retrieved July 12, 2022 from https://www.mediamatters.org/facebook/racist-troll-owen-benjamin-evading-bans-major-social-media-platforms-continue-spreading
[112] Seth Simons. 2021. *Twitter Permanently Suspends Anthony Cumia.* Retrieved July 12, 2022 from https://www.pastemagazine.com/comedy/anthony-cumia/twitter-permanently-suspends-anthony-cumia/
[113] Mohit Singhal, Chen Ling, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2022. SoK: content moderation in social media, from guidelines to enforcement, and research to practice. *arXiv preprint arXiv:2206.14855* (2022).
[114] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media & Society* 21, 9 (2019), 1910–1927.
[115] Angelica Stabile. 2022. *Dr. Robert Malone on Joe Rogan interview censorship, Twitter ban: 'You can't suppress information'.* Retrieved July 12, 2022 from https://www.foxnews.com/media/dr-robert-malone-joe-rogan-covid-ingraham-twitter-ban-youtube-censorship
[116] Christine Stapleton. 2019. *This far-right provocateur is banned from social media, but she's still running for Congress.* Retrieved July 12, 2022 from https://www.palmbeachpost.com/story/news/local/2019/10/08/this-far-right-provocateur-is-banned-from-social-media-but-shes-still-running-for-congress/2545693007/
[117] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
[118] Chris Stokel-Walker. 2023. *Twitters $42,000-per-Month API Prices Out Nearly Everyone.* Retrieved July 15, 2023 from https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/
[119] Jo Ellen Stryker, Ricardo J Wray, Robert C Hornik, and Itzik Yanovitzky. 2006. Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism & Mass Communication Quarterly* 83, 2 (2006), 413–430.
[120] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.
[121] Edson C Tandoc Jr, Chen Lou, and Velyn Lee Hui Min. 2019. Platform-swinging in a poly-social-media context: How and why users navigate multiple social media platforms. *Journal of Computer-Mediated Communication* 24, 1 (2019), 21–35.
[122] Jacob Thebault-Spieker, Sukrit Venkatagiri, Naomi Mine, and Kurt Luther. 2023. Diverse Perspectives Can Mitigate Political Bias in Crowdsourced Content Moderation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1280–1291.

[123] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. 243–258.
[124] Twitter. 2022. *Notices on Twitter and what they mean.* Retrieved June 24, 2022 from https://help.twitter.com/en/rules-and-policies/notices-on-twitter
[125] Twitter. 2022. *Rules and policies.* Retrieved June 24, 2022 from https://help.twitter.com/en/rules-and-policies#general
[126] Kristen Vaccaro, Ziang Xiao, Kevin Hamilton, and Karrie Karahalios. 2021. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–28.
[127] Reed Van Schenck. 2023. Deplatforming "the people": media populism, racial capitalism, and the regulation of online reactionary networks. *Media, Culture & Society* (2023), 01634437231169909.
[128] Daniel Victor. 2016. *Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.* Retrieved July 15, 2023 from https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html
[129] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. *arXiv preprint arXiv:2211.04325* (2022).
[130] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1231–1245.
[131] Jim Waterson. 2022. *TechScape: Inside the rise and fall of Politics For All.* Retrieved July 14, 2022 from https://www.theguardian.com/technology/2022/jan/12/techscape-politics-for-all-twitter
[132] Brian E Weeks, Alberto Ardèvol-Abreu, and Homero Gil de Zúñiga. 2017. Online influence? Social media use, opinion leadership, and political persuasion. *International journal of public opinion research* 29, 2 (2017), 214–239.
[133] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L Mazurek, and Blase Ur. 2020. What Twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own Twitter data. In *29th USENIX Security Symposium (USENIX Security 20)*. 145–162.
[134] Mark White. 2021. *Anjem Choudary: Islamist hate preacher banned from Twitter.* Retrieved July 12, 2022 from https://news.sky.com/story/anjem-choudary-islamist-hate-preacher-banned-from-twitter-12367411
[135] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
[136] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*. 705–714.
[137] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2022. Addressing harm in online gaming communities–the opportunities and challenges for a restorative justice approach. *arXiv preprint arXiv:2211.01524* (2022).
[138] Brandy Zadrozny. 2018. *Right-wing platforms provide refuge to digital outcasts — and Alex Jones.* Retrieved July 14, 2022 from https://www.nbcnews.com/tech/tech-news/right-wing-platforms-provide-refuge-digital-outcasts-alex-jones-n899161
[139] Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. In *Companion Proceedings of the The Web Conference 2018*. 1007–1014.
[140] Xuan Zhao, Cliff Lampe, and Nicole B Ellison. 2016. The social media ecology: User perceptions, strategies and challenges. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 89–100.

A RULES OF TWITTER PERMANENT SUSPENSION

Table 4. Rules of permanent suspension[125]

Level-1	Level-2	Behavior that can lead to permanent suspension
Safety	Violent threats	Statements of an intent to kill or inflict serious physical harm on a specific person or group of people
	Glorification of violence	Glorifying, celebrating, praising or condoning violent crimes, violent events where people were targeted because of their membership in a protected group, or the perpetrators of such acts

Continued on next page

Table 4 – continued from previous page

Level-1	Level-2	Behaviors that can lead to permanent suspension
	Violent organizations	Affiliating with or promoting the illicit activities of a terrorist organization or violent extremist group
	Child sexual exploitation	Any content that depicts or promotes child sexual exploitation
	Abusive behaviors[1]	Behaviors that harasses or intimidates, or is otherwise intended to shame or degrade others
	Hateful conduct[1]	Abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized (including women, people of color, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities)
	Perpetrators of violent attacks	Individual perpetrators of terrorist, violent extremist, or mass violent attacks
	Suicide and self-harm	Repeatedly promoting or encouraging suicide or self-harm, or dedicated to promoting or encouraging self-harm or suicide
	Sensitive media	Repeatedly posting live photos and profiles images about graphic violence, adult content, and hateful imagery, or dedicated to posting graphic violence, adult content, hateful imagery, violent sexual conduct and gratuitous gore.
	Illegal or certain regulated goods or services	Repeatedly selling, buying, or facilitating transactions in illegal goods or services, as well as certain types of regulated goods or services or dedicated to the sale of illegal or regulated goods and/or services
Privacy	Private information and media	Repeatedly sharing private information (such as home address, identity documents etc.)
	Non-consensual nudity	Posting or sharing intimate photos or videos of someone that were produced or distributed without their consent
Authenticity	Platform manipulation and spam	Severely and artificially amplifying or suppressing information or engaging in behavior that manipulates or disrupts people's experience on Twitter (such as using any of the tactics described on this page to undermine the integrity of elections, buying/selling accounts, creating accounts to replace or mimic a suspended account)
	Civic Integrity	Sharing content about manipulating or interfering in elections or other civic processes, and the number of strikes caused by violating this policy is over 5

Continued on next page

Table 4 – continued from previous page

Level-1	Level-2	Behaviors that can lead to permanent suspension
	Misleading and Deceptive Identities[1]	Engaging in impersonation or using a misleading or deceptive fake identity
	Synthetic and manipulated media[1]	Sharing harmful misleading narratives that violate the synthetic and manipulated media policy
	Copyright and trademark	Repeatedly violating others' intellectual property rights, including copyright and trademark
	Parody, commentary, and fan account[1]	Depicting another person, group, or organization in account profile to discuss, satirize, or share information about that entity, and make insufficient edits to the profile after the first warning.
	Coordinated harmful activity	Using specific, detectable techniques of platform manipulation to engage in the artificial inflation or propagation of a message or narrative on Twitter
	Distribution of hacked materials	Account directly operated by hackers, hacking groups, or people acting for or on behalf of such hackers, and engaging in the direct distribution of hacked materials
	COVID-19 misleading information	False affiliation: If the account is determined to misrepresent their affiliation, or share content that falsely represents its affiliation as a medical practitioner, public health official or agency, research institution, or that falsely suggests expertise on COVID-19 issues. Repeated Violations: If the account repeatedly violates the COVID-19 misinformation policy over a 30-day time period, or if the account has been set up for the expressed purpose of Tweeting false or misleading information about COVID-19.
	Ban evasion	Circumventing a Twitter enforcement action (such as a permanent suspension) by creating accounts or repurposing existing accounts to replace or mimic a suspended account

[1] Violators' penalty will be determined by a number of factors including, but not limited to, the severity of the violation and the violators' previous records of rule violations.

Received 15 January 2023; revised 18 July 2023; accepted 18 September 2023