No Easy Way Out: The Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment

Anh V. Vu[†], Alice Hutchings[†], Ross Anderson[†§]
[†]University of Cambridge, [§]University of Edinburgh

***Abstract*—Legislators and policymakers worldwide are debating options for suppressing illegal, harmful and undesirable material online. Drawing on several quantitative data sources, we show that deplatforming an active community to suppress online hate and harassment, even with a substantial collective effort involving several tech firms, can be hard. Our case study is the disruption of the largest and longest-running harassment forum KIWI FARMS in late 2022, which is probably the most extensive industry effort to date. We collected complete snapshots of this site and its primary competitor LOLCOW FARM, encompassing over 14.7M posts during their lifespan over the past decade. These data are supplemented with a full scrape of the Telegram channel used to disseminate new updates when the forum was down, tweets made by the online community leading the takedown, and with search interest and web traffic to the forum spanning two months before and four months after the event. Despite the active participation of a number of tech companies over several consecutive months, this campaign failed to shut down the forum and remove its objectionable content. While briefly raising public awareness, it led to rapid platform displacement and traffic fragmentation. Part of the activity decamped to Telegram, while traffic shifted from the primary domain to previously abandoned alternatives. The forum experienced intermittent outages for several weeks, after which the community leading the campaign lost interest, traffic was directed back to the main domain, users quickly returned, and the forum was back online and became even more connected. The forum members themselves stopped discussing the incident shortly thereafter. The net effect was that forum activity, active users, threads, posts and traffic were all cut by about half. The disruption largely affected casual users (of whom roughly 87% left), while half the core members remained engaged. It also drew many newcomers, who exhibited increasing levels of toxicity during the first few weeks of participation. Deplatforming a community without a court order raises philosophical issues about censorship versus free speech; ethical and legal issues about the role of industry in online content moderation; and practical issues on the efficacy of private-sector versus government action. Deplatforming a dispersed community using a series of court orders against individual service providers appears unlikely to be very effective if the censor cannot incapacitate the key maintainers, whether by arresting them, enjoining them or otherwise deterring them.**

***Index Terms*—deplatforming, hate, harassment, online forums, website takedown, content moderation; censorship; KIWI FARMS.**

I. INTRODUCTION

Online content is now prevalent, widely accessible, and influential in shaping public discourse. Yet while online places facilitate free speech, they do the same for hate speech [1], and the line between the two is often contested. Some cases of stalking, bullying, and doxxing such as Gamergate have had real-world consequences, including violent crime as well as political mobilisation [2]. Content moderation has become a critical function of tech companies, but also a political tussle space, since abusive accounts may affect online communities in significantly different ways [3]. Online social platforms employ various mechanisms to detect, moderate, and suppress objectionable content [4], including "hard" and "soft" techniques [5]. These range from reporting users of illegal content to the police, through deplatforming users who break terms of service [6], to moderating legal but obnoxious content [7], which may involve actions such as flagging it with user warnings, downranking it in recommendation algorithms, or preventing its being monetized through ads [8], [9], [10].

Deplatforming may mean blocking individual users, but sometimes the target is not a single bad actor, but a whole community, such as one involved in crime [11]. It can be undertaken by industry, as when Cloudflare terminated service for the Daily Stormer after the Unite the Right rally in Virginia in 2017 [12] and for 8chan in August 2019 [13]; or by law enforcement, as with the FBI taking down DDoS-for-hire services in 2018 [14], [15] and 2022 [16], and seizing Raid Forums in 2022 [17]. Industry disruption has often been short-lived; both 8chan and Daily Stormer re-emerged shortly after being disrupted. Police intervention is often slow and less effective, and its impact may also be temporary [11]. After the FBI shut down Silk Road in 2013 [18], the online drug market fragmented among multiple smaller marketplaces [19]. The seizure of Raid Forums led to the emergence of its successor Breach Forums. Furthermore, the takedowns against DDoS-for-hire services cut the attack volume significantly, yet the market recovered rapidly [14], [15].

KIWI FARMS is the largest and longest-running online harassment forum [20], [21]. It is often associated with real-life trolling and doxxing campaigns against feminists, gay rights campaigners and minorities such as disabled, transgender, and autistic individuals; some have killed themselves after being harassed [22]. Despite being unpleasant and widely controversial, the forum has been online for a decade and had been shielded by Cloudflare's DDoS protection for years. This came to an end following serious harassment by forum members of a Canadian trans activist, culminating in a swatting incident in August 2022.[1] This resulted in a community-led campaign on Twitter to pressure Cloudflare and other tech firms to drop the forum [23], [24], [25]. This escalated quickly, generating significant social media attention and mainstream headlines. A series of tech firms then attempted to take the forum down; they included DDoS protection services, infrastructure providers, and even some Tier-1 networks [26], [27], [28],

[1] This is when a harasser falsely reports a violent crime in progress at the victim's home, resulting in the arrival of a special-weapons-and-tactics (SWAT) team to storm the premises, placing the victim and family at risk.

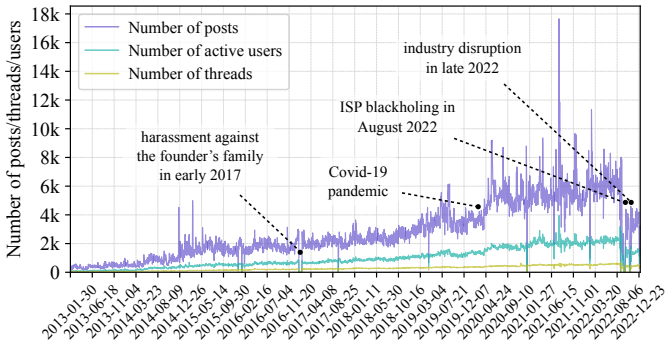

Figure 1: Activity levels and major incidents affecting KIWI FARMS during its one-decade lifetime from 2013 to late 2022.

[29], [30]. This extraordinary series of events lasted for a few months and was the most sustained effort to date to suppress an active online hate community. It is notable that tech firms gave in to public pressure in this case, while they have in the past resisted substantial pressure from governments.

Existing studies have investigated the efficacy of deplatforming social-media users [31], [32], [33], [34], [35], [36], [37], yet there has been limited research into the effectiveness of industry disruptions against hate communities – both quantitatively and qualitatively. This paper investigates how well the industry dealt with a hate site. Our goals were to evaluate the efficacy of the effort; to understand the impacts and challenges of deplatforming as a means to suppress online hate and harassment; and to examine the role of industry in censorship and content regulation.

We outline the disruption landscape in §II, then describe our methods and datasets in §III. Sections §IV and §V assess the impacts on the forum itself and the relevant stakeholders. We discuss the role of industry in tackling online harassment, censorship and content regulation, as well as legal, ethical, and policy implications of the incident in §VI. Our data collection and analyses were approved by our institutional Ethics Review Board. Our data are available to academics on request.

II. DEPLATFORMING TO SUPPRESS ONLINE HATE AND HARASSMENT

There is a complex ecosystem of online abuse, which has been evolving for decades [38]. There can be a large grey area between criminal behaviour and socially acceptable behaviour online, just as in real life. And just as a pub landlord will throw out rowdy customers so platforms have acceptable use policies backed by content moderation [39], to enhance their users' experience and protect advertising revenue [40].

A. Deplatforming and its Efficacy

Deplatforming refers to blocking, excluding or restricting individuals or groups from using online services, on the grounds that their activities are unlawful, or that they do not comply with the platform's acceptable use policy [6]. Various extremists and criminals have been exploiting online platforms for over thirty years, resulting in a complex ecosystem in which some harms are prohibited by the criminal law (such as terrorist radicalisation and child sex abuse material) while many others are blocked by platforms seeking to provide welcoming spaces for their users and advertisers. For a history and summary of current US legislative tussles and their possible side-effects, see Fishman [41]. The idea is that if a platform is used to disseminate abusive speech, removing the speech or indeed the speakers could restrict its spread, make it harder for hate groups to recruit, organise and coordinate, and ultimately protect individuals from mental and physical harm. Deplatforming can be done in various ways, ranging from limiting users' access and restricting their activity for a time period, to suspending an account, or even stopping an entire group of users from using one or more services. For example, groups banned from major platforms can displace to other channels, whether smaller websites or messenger services [6].

Different countries draw the line between free speech and hate speech differently. For example, the USA allows the display of Nazi symbols while France and Germany do not [42]. Private firms offering ad-supported social networks generally operate much more restrictive rules, as their advertisers do not want their ads appearing alongside content that prospective customers are likely to find offensive. People wishing to generate and share such material therefore tend to congregate on smaller forums. Some argue that taking down such forums infringes on free speech and may lead to censorship of legitimate voices and dissenting opinions, especially if it is perceived as politically motivated. Others maintain that deplatforming is necessary to protect vulnerable communities from harm. Debates rage in multiple legislatures; as one example, the UK Online Safety Bill will enable the (politically-appointed) head of Ofcom, the UK broadcast regulator, to obtain court orders to shut down online places that are considered harmful [43]. This lead us to ask: how effective might such an order be?

Most studies assessing the impact of deplatforming have worked with data on social networks. Deplatforming users may reduce activity and toxicity levels of relevant actors on Twitter [31] and Reddit [32], [33], limit the spread of conspiratorial disinformation on Facebook [34], and minimise disinformation and extreme speech on YouTube [35]. But deplatforming has often made hate groups and individuals even more extreme, toxic and radicalised. They may view the disruption of their platform as an attack on their shared beliefs and values, and move to even more toxic places to continue spreading their message. There are many examples: the Reddit ban of r/incels in November 2017 led to the emergence of two standalone forums, incels.is and incels.net, which then grew rapidly; users banned from Twitter and Reddit exhibit higher levels of toxicity when migrating to Gab [36]; users migrated to their own standalone websites after getting banned from r/The_Donald expressed higher levels of toxicity and radicalisation, even though their posting activity on the new platform decreased [44], [45]; the 'Great Deplatforming' directed users to other less regulated, more extreme platforms [46]; the activity of many right-wing users moved to Telegram increased multi-fold after being banned on major social media [37];

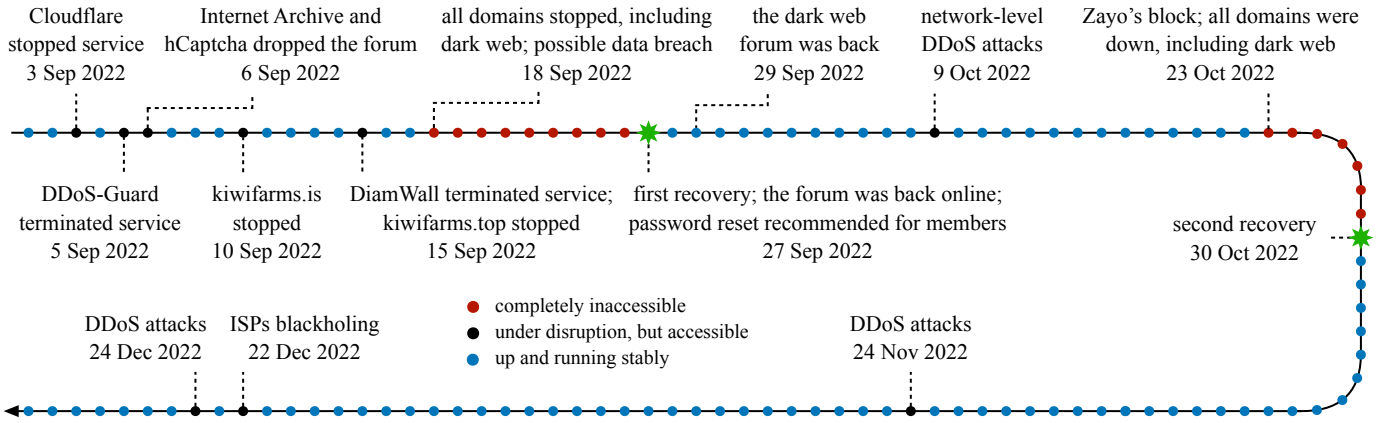

Figure 2: Major incidents disrupting KIWI FARMS from September to December 2022. Green stars indicate the forum recovery.

users banned from Twitter are more active on Gettr [47]; and communities migrated to Voat from Reddit can be more resilient [48]. Blocking can also be ineffective for technical and implementation reasons: removing Facebook content after a delay appears to have been ineffective and had limited impact due to the short cycle of users' engagement there [49].

Standalone communities, such as websites and forums, may be more resilient as the admin has control of all the content, facilitating easy backups and restores. Previous work has documented the impacts of law enforcement interventions on online cybercrime marketplaces and services [14], [15], [19], yet how effective the industry can be in dealing with such extreme, radicalised communities remains unstudied.

B. Kiwi Farms and the Disruptions

KIWI FARMS had been growing steadily over a decade (see Figure 1) and had been under Cloudflare's DDoS protection for some years.[2] An increase of roughly 50% in forum activity happened during the COVID-19 lockdown starting in March 2020, presumably as people were spending more time online. Prior interventions have resulted in the forum getting banned from Google Adsense, and from Mastercard, Visa and PayPal in 2016; from hundreds of VPS providers between 2014–2019 [50]; and from selling merchandise on the print-on-demand marketplace Redbubble in 2016. XenForo, a close-source forum platform, revoked its license in late 2021 [51]. DreamHost stopped its domain registration in July 2021 after a software developer killed himself after being harassed by the site's users. This did not disrupt the forum as it was given 14 days to seek another registrar [52]. While these interventions may have had negative effects on its profit and loss account, they did not impact its activity overall. The only significant disruption in the forum's history was between 22 January and 9 February 2017 (19 days), when the forum's owner suspended it himself due to his family being harassed [53].[3]

[2] Cloudflare's service tries to detect suspicious patterns and drop malicious ones, only letting legitimate requests through.

[3] Minor suspensions found in our data are on 2 February 2013, 24 January 2016, 29 September 2017, and 11 January 2021, without clear reasons.

The disruption studied in this work was started by the online community in 2022. A malicious alarm was sent to the police in London, Ontario by a forum member on 5 August, claiming that a Canadian trans activist had committed murders and was planning more, leading to her being swatted [54]. She and her family were then repeatedly tracked, doxxed, threatened, and generally harassed. In return, she launched a campaign on Twitter on 22 August under the hashtag #dropkiwifarms and organised a protest outside Cloudflare's headquarters to pressure the company to deplatform the site [55]. This campaign generated lots of attention and mainstream headlines, which ultimately resulted in several tech firms trying to shut down the forum. This is the first time that the forum was completely inaccessible for an extended period due to an external action, with no activity on any online places including the dark web. It attempted to recover twice, but even when it eventually returned online, the overall activity was roughly halved.

The majority of actions taken to disrupt the forum occurred within the first two months of the campaign. Most of them were widely covered in the media and can be checked against public statements made by the industry and the forum admins' announcements (see Figure 2). The forum came under a large DDoS attack on 23 August, one day after the campaign started. It was then unavailable from 27 to 28 August due to ISP blackholing. Cloudflare terminated their DDoS prevention service on 3 September – just 12 days after the Twitter campaign started – due to an "unprecedented emergency and immediate threat to human life" [26]. The forum was still supported by DDoS-Guard (a Russian competitor to Cloudflare), but that firm also suspended service on 5 September [27]. The forum was still active on the dark web but this .onion site soon became inaccessible too. On 6 September, hCaptcha dropped support; the forum was removed from the Internet Archive on the same day [56]. This left it under DiamWall's DDoS protection and hosted on VanwaTech – a hosting provider describing themselves as neutral and non-censored [57]. On 15 September 2022, DiamWall terminated their protection [28] and the '.top' domain provider also stopped support [29]. The forum was completely down from 19 to 26 September and

from 23 to 29 October. From 23 October onwards, several ISPs intermittently rejected announcements or blackholed routes to the forum due to violations of their acceptable use policy, including Voxility and Tier-1 providers such as Lumen, Arelion, GTT and Zayo. This is remarkable as there are only about 15 Tier-1 ISPs in the world. The forum admin devoted extensive effort to maintaining the infrastructure, fixing bugs, and providing guidance to users in response to password breaches. Eventually, by routing through other ISPs, the forum was able to get back online and remain stable, particularly following its second recovery.

III. Methods and Datasets

Our primary approach is data-driven, with findings supported by quantitative evidence derived from multiple longitudinal data sources. Where applicable, we enrich the findings with complementary qualitative content analysis of posts, tweets, announcements, and public statements. Our collection is maintained on a regular basis. All the data used are widely accessible and can be publicly scraped by anyone. We refrain from scraping images due to safety and legality concerns.

A. Forum and Imageboard Discussions

Besides common mainstream social media channels like Facebook and Twitter, independent platforms such as xenForo[4] and Infinity[5] have gained popularity as tools for building online communities. Despite being less visible and requiring more upkeep, these can offer greater resistance against external intervention as the operators have full control over the content and databases, thereby allowing easy backup and redeployment in case of disruption. These platforms typically share a hierarchical data structure ranging from bulletin boards down to threads linked to specific topics, each containing several posts. While facilitating free speech, these also increasingly nurture and disseminate hate and abusive speech. We have been scraping the two most active forums associated with online harassment for years due to their increasingly toxic content, as part of the ExtremeBB dataset [21]: Kiwi Farms and Lolcow Farm.

Our collection includes not only posts but also associated metadata such as posting time, user profiles, reactions, and levels of *toxicity*, *identity attack* and *threat* measured by the Google Perspective API as of January 2023.[6] Perspective API also offers other measures such as *insult* and *profanity* [58], but we exclude these due to lack of relevance to this paper's aim. We strive to ensure data completeness by designing our scrapers to visit all sub-forums, threads, and posts while keeping track of every single crawl's progress to resume incrementally in case of any interruption.

Kiwi Farms is built on xenForo, but the operators have been maintaining the forum by their own efforts since late 2021 when xenForo officially revoked their license. Our data covers the entire history of the forum from early January 2013 to the end of 2022 with 10.1M posts in 48k threads made by 59k active users, providing a full landscape through its evolution over time. While some extremist forums experienced fluctuating activity and rapid declines in recent years [21], Kiwi Farms has shown stable growth until being significantly disrupted in 2022 (see Figure 1). Our data precisely capture major reported suspensions, including those in 2017 and 2022.

According to Similarweb [59] and Semrush [60], the primary rival is Lolcow Farm, an imageboard built on Infinity. While Kiwi Farms discussions are largely text-based, Lolcow Farm is centred on descriptive images. While Kiwi Farms users adopt pseudonyms, Lolcow Farm users mostly remain hidden under the unified 'Anonymous' handle. We gathered a complete snapshot of Lolcow Farm from its inception in June 2014 to the end of 2022, encompassing 4.6M posts made in 10k threads. Lolcow Farm has much fewer threads, but each typically contains lots of posts. This collection brings the total number of posts for both forums to 14.7M (and still growing). We exclude Lolcow, a smaller competitor to Kiwi Farms (also based on xenForo), as it vanished in mid-2022 and had less than 30k posts in total. As Lolcow Farm is now the largest competitor, analysing it lets us estimate platform displacement when Kiwi Farms was down.

B. Telegram Chats

During periods of inaccessibility, the activity level increased in a Telegram group, which was mainly used to disseminate announcements and updates, particularly about where and when the forum could be accessed. This channel permits public access, allowing people to join and view historical messages. We used Telethon[7] to collect a snapshot of this channel during its lifespan from late August to the end of 2022, encompassing 525k messages, 298k replies, and associated metadata such as view counts and 356k emoji reactions made by 2 502 active users. The data is likely complete as messages and metadata are fully captured through the use of official Telegram APIs. As the forum operators are driven to keep users quickly informed, their announcements provide a reliable incident and response timeline.

C. Web Traffic and Search Trends Analytics

We found from announcements in the Telegram group that Kiwi Farms could be accessed through six major domains: the primary one is kiwifarms.net and four alternatives are kiwifarms.ru, kiwifarms.top, kiwifarms.is, and kiwifarms.st, while a Pleroma decentralised web version is at kiwifarms.cc.[8] To investigate how users navigated across these domains when the forum experienced disruption, we analysed traffic analytics towards all six domains provided by Similarweb – the leading platform in the market providing insights and intelligence into web traffic and performance.[9] Their reports aggregate

[4] The xenForo Platform: https://xenforo.com/

[5] The Infinity Imageboard: https://github.com/ctrlcctrlv/infinity/

[6] Google Perspective API: https://perspectiveapi.com/

[7] Telethon: https://telethon.dev/

[8] Other domains include kiwifarms.tw, kiwifarms.hk, and kiwifarms.pl, however they are either new or insignificant so their traffic data is trivial.

[9] Similarweb: https://similarweb.com/. Another popular web analytics is Semrush at https://semrush.com/, but it does not offer daily statistics.

anonymous statistics from multiple inputs, including their own analytic services, data sharing from ISPs and other measurement companies, data crawled from billions of websites, and device traffic data (both website and app) such as plugins, add-ons and pixel tracking. Their algorithm then extrapolates the substantial aggregated data to the entire Internet space. Their estimation therefore may not be completely precise, but reliably reflects trends at both global and country levels. In a separate paper, we tested the reliability of Similarweb data with a comparison to millions of ground truth traffic records collected from our own infrastructure over 6 months, showing that while Similarweb largely underestimates the amount of traffic, it is able to capture trends with a very high correlation (Pearson's coefficient > 0.9) [citation hidden]. Our analysis in the next section also suggests a high correlation between the traffic data and the forum activity.

As Similarweb does not offer an academic license, we use a free trial account[10] to access longitudinal web traffic and engagement data going back the past three months. This includes information about total visits, unique visitors, visit duration, pages per visit, bounce rate, and page views. It also provides figures on search activity, data for marketing such as visit sources (e.g., direct, search, email, social, referral, ads), and non-temporal insight into audience geography and demographics. These data, covering both desktop and mobile traffic, provide valuable perspectives. They span from July to December 2022, two months before and four months after the disruption; this time frame is sufficient as there was no significant industry intervention against the forum in the past (as shown in Figure 1), and the disruption campaign mostly ended after a few months (see §IV). In addition, we also collected search trends by countries and territories over time from Google Trends, covering the entire lifetime of the forum. Both of these datasets are likely to be complete as they were gathered directly from Similarweb and Google.

D. Tweets Made by the Online Community

The disruption campaign started on Twitter on 22 August 2022 with tweets posted under the hashtag #dropkiwifarms. We gathered the main tweets plus associated metadata, such as posting time and reactions (e.g., replies, retweets, likes, and quotes) using SNSCRAPE, an open-source Python framework for social network scrapers.[11] As they use Twitter APIs as the underlying method, the data are likely to be complete. We collected 11 076 tweets made by 3 886 users, spanning the entire campaign period. This data helps us understand the community reaction throughout the campaign, when the industry took action, and when the forum recovered. There might be more related tweets without the hashtag #dropkiwifarms of which we are unaware, but we believe the trend measured by our collection is reliable.

[10] A business subscription offers 6 months of historical data, but neither it nor the free trial provides access to longitudinal country-based records.

[11] SNSCRAPE: https://github.com/JustAnotherArchivist/snscrape/

E. Data Licensing

Our datasets and scripts for data collection and analysis are available to academics. However, as both researchers and involved actors such as forum members might be exposed to risk and harm [61], we decline to make our data publicly accessible. It is our standard practice at the Cambridge Cybercrime Centre to require our licensees to sign an agreement to prevent misuse, to ensure the data will be handled appropriately, and to keep us informed about research outcomes. We have a long history of sharing such sensitive data, and robust procedures to enable data sharing in multiple jurisdictions.

F. Ethical Considerations

Our work was formally approved by our institutional Ethics Review Board (ERB) for data collection and analysis. Our datasets are collected on publicly available forums and channels, which are accessible to all. We collected the forum when it was hosted in the US; according to a 2022 US court case, scraping public data is legal [62]. Our scraping method does not violate any regulations and does not cause negative consequences to the targeted websites e.g., bandwidth congestion or denial of service. It would be impractical to send thousands of messages to gain consent from all forum and Telegram members; we assume they are aware that their posting activity on public online places will be widely accessible.

In contrast to some previous work on online forums, we name the investigated forums in this paper. Pseudonymising the forum name is pointless because of the high-profile campaign being studied. Thus, we avoid the pretence that the forum is not identifiable and shift the focus to accounting for the potential harms to both researchers and involved actors associated with our research. We designed our analysis to operate ethically and collectively by only presenting aggregated behaviours to avoid private and sensitive information of individuals being inferred. This is in accordance with the British Society of Criminology Statement on Ethics [63].

Researchers may be at risk when doing work on sensitive data [61]. Studying extremist forums may introduce a higher risk of retaliation than other forums, resulting in mental or physical harm. We have taken measures to minimise potential harm to researchers and involved actors when doing studies with human subjects and at-risk populations [64], [65]. For example, we consider options to anonymise authors' names or use pseudonyms for any publication related to the project, including this paper, if necessary. We also refrain from directly looking at media; our data collection only scrapes text while discarding images and private/protected posts.

IV. THE IMPACTS ON FORUM ACTIVITY AND TRAFFIC

On 3 September, Cloudflare discontinued its DDoS prevention service, which attracted major publicity. This intervention led to a sudden and significant increase in global search interest about KIWI FARMS with a seven-fold spike, along with the web traffic to the six major domains doubling on 4 September (see Figure 3). This phenomenon, known as the Streisand effect, might be caused by people's curiosity

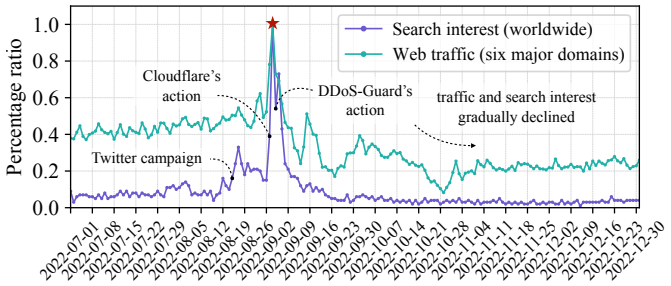

Figure 3: Global search trends and traffic to all forum domains during the disruption. The star indicates the Streisand effect.

about what happened to the platform, which is relatively rare but mainly seen with 'freedom of speech' issues [11]. It suggests that attempts at censorship may actually end up being counterproductive [66]: disruptive effort aiming to reduce user interactions instead led to the unintended consequences of increased attention, despite such effect lasting for only a few days before declining sharply. We now examine in detail the impacts of the disruption and the forum recovery on KIWI FARMS itself within 6 months from July to December 2022. This time frame provides a sufficient understanding, as the campaign was mostly over and the forum was growing stably before the disruption.

A. The Impacts of Major Disruptions

While some DDoS attacks were large enough to shut the forum down, their impact was temporary. For example, the DDoS attack on 23 August – which was probably associated with the Twitter campaign the previous day – led to a drop of roughly 35% in posting volume, yet the forum activity recovered the next day to a slightly higher level (see the first graph of Figure 4). The DDoS attack during Christmas 2022 was also short-lived. The ISP blackholing on 26 August was more critical, silencing the forum for two consecutive days, yet it again managed to recover quickly.

The most significant, long-lasting impact was caused by the substantial industry disruption that we analyse in this paper. While activity immediately dropped by around 20% after Cloudflare's action on 3 September, the forum was still online at kiwifarms.ru, hosting the same content. Activity did not degrade significantly until DDoS-Guard's action on 5 September, which took down the Russian domain. By 18 September, all domains were unavailable, including .onion (presumably their hosting was identified); forum activity dropped to zero and stayed there for a week. The operator managed to get the forum back online for the first time on 27 September 2022, after which it ran stably on both the dark web and clear web for roughly one month until Zayo – a Tier-1 ISP – blocked it on 23 October. This led to another silent week before the forum eventually recovered a second time on 30 October. It has been stable since then without serious downtime except for the ISP blackholing on 22 December which led to a 70% drop in activity. In general, although the forum is now back online, hosted on 1776 Solutions – a company also founded by the forum's owner – it has failed to bounce back to the pre-disruption level, with the number of active users and posting volume roughly halved. In short, the industry effort was much more effective than previous DDoS attacks, yet still could not silence the forum for long.

B. Platform Displacement

The natural behaviour of online communities when their usual gathering place becomes inaccessible is to seek alternative places or channels. The second graph in Figure 4 shows an initial shift of forum activity to Telegram that occurred on 27 August, right after the ISP blackholing. This was accompanied by thousands of emoji reactions on the admin's announcements since commenting was not allowed at that time. Community reactions (e.g., replies, emojis) seem to have been consistent with the overall Telegram posting activity, which increased rapidly afterwards and even occasionally surpassed the forum's activity, especially after the publicity given to the Cloudflare and DDoS-Guard actions. However, significant displacements only occurred when all domains were completely inaccessible on 18 September, and again when Zayo blocked the forum's second incarnation on 22 October. The shift to Telegram appears to be rapid yet rather temporary: users quickly returned to the forum when it became available, while activity on Telegram gradually declined.

There was no significant shift in activity from the forum to its primary competitor LOLCOW FARM (see the third graph of Figure 4), however, there was an increase in posting on LOLCOW FARM about the incident, indicating a minor change of discussion topic (see more in §V-D). It is unclear if these posting users migrated from KIWI FARMS, as LOLCOW FARM do not use handles, making user counts unavailable. LOLCOW FARM also experienced downtime on 17 and 18 September (the same day as KIWI FARMS) yet we have no reliable evidence to draw any convincing explanation. Another drop occurred around Christmas 2022 in sync with KIWI FARMS, perhaps because of the holiday. The activity of LOLCOW FARM returned to its previous level quickly after these drops, suggesting that the campaign did not significantly impact LOLCOW FARM or drive content between the rival ecosystems; the displacement we observed on KIWI FARMS was mostly 'internal' within its own ecosystem, rather than an 'external' shift to other forums.

C. Traffic Fragmentation

Before Cloudflare's action on 3 September, traffic towards KIWI FARMS (measured by Similarweb) was relatively steady, mostly occupied by the primary domain. However, we see the Streisand effect (as also seen in Figure 3) with an immediate peak in traffic of around 50% more visits and 85% more visitors once the site was disrupted. The publicity given by the takedown presumably boosted awareness and attracted people to visit both the primary and alternative domains. Traffic to the primary domain was then significantly fragmented to other previously abandoned domains, resulting in the kiwifarms.net accounting for less than 50% one day after Cloudflare's intervention, as shown in Figure 5.

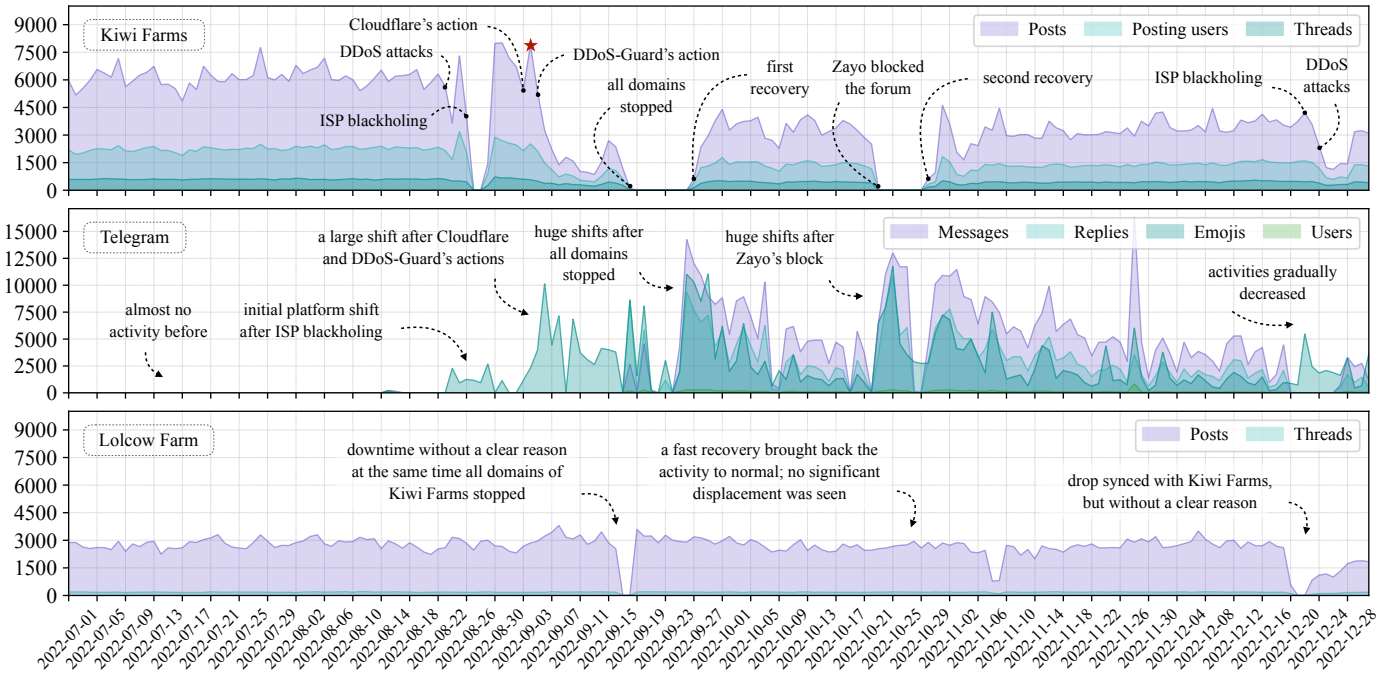

Figure 4: Number of daily posting activity, threads, and active users on KIWI FARMS, its Telegram channel, and LOLCOW FARM, as well as major disruptions and displacement from KIWI FARMS to other platforms. The red star indicates the Streisand effect.

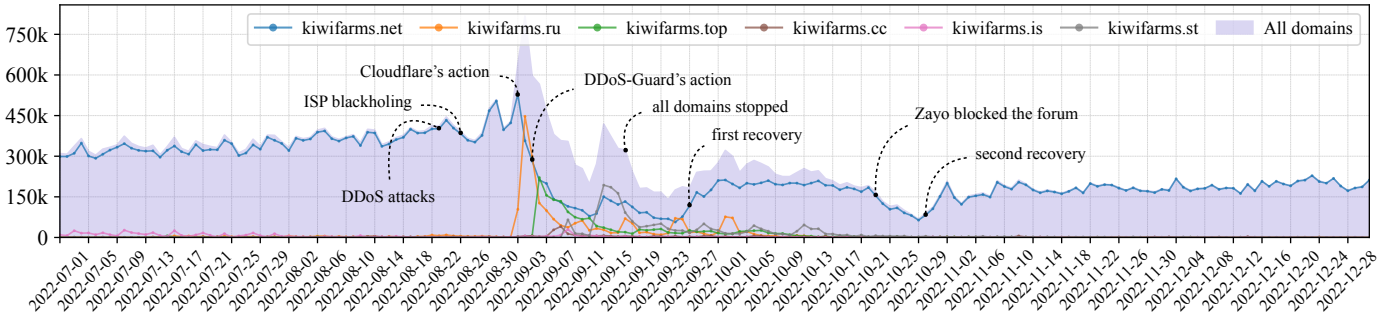

Figure 5: Number of daily estimated visits and the fragmentation from the primary domain to previously abandoned alternatives. We see non-zero traffic to the primary domain when the forum was down, presumably Similarweb counted unsuccessful attempts.

Following the unavailability of kiwifarms.net, most traffic was directed to kiwifarms.ru, which was under DDoS Guard's protection (accounting for around 60% total traffic on 4 September). The DDoS-Guard's action on 5 September reduced traffic towards kiwifarms.ru sharply, while traffic towards kiwifarms.top peaked. The suspension of kiwifarms.top on the following day led to increased traffic towards kiwifarms.cc (a Pleroma decentralised web instance), but it only lasted for a couple of days before traffic shifted again to kiwifarms.is. The seizure of kiwifarms.is later led to the traffic shifting to kiwifarms.st, but it was also short-lived.

The forum recovery on 27 September gradually directed almost all traffic back to the primary domain, and by 22 October, kiwifarms.net mostly accounted for all traffic, albeit at about half the volume. This effect is highly consistent with what has been found in our forum data, indicating a reliable pattern. Overall, our evidence suggests a clear traffic fragmentation across different domains, in which people attempted to visit surviving domains when one was disrupted.

V. THE IMPACTS ON RELEVANT STAKEHOLDERS

We have looked at the impacts of the disruption on KIWI FARMS itself. This section examines the effects on relevant stakeholders, including the harassed victim, the community leading the campaign, the industry, the forum operators, and active forum users who posted at least once. As our ethics approval does not allow the study of individuals, all measurements are conducted collectively on subsets of users. Besides quantitative evidence, we also qualitatively look at statements made by tech firms about the incident.

A. The Community who Launched the Campaign

There were 3 886 users in the online community involved in starting the campaign. Of these, 1 670 users (42.97%) were responsible for around 80% of tweets. There was a sharp

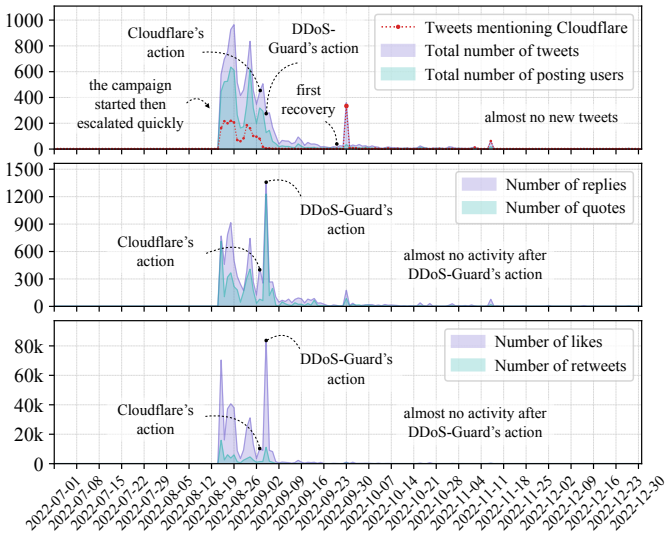

Figure 6: The number of daily tweets and reactions made by the community about the campaign. Figure scales are different.

increase in tweets and reactions at the beginning (see Figure 6). The first peak was on 25 August with nearly 900 tweets by around 600 users. However, this dropped rapidly to less than 100 per day after a few weeks when Cloudflare and DDoS-Guard took action, and almost to zero two weeks later. The number of tweets specifically mentioning Cloudflare (such as their official account, as well as those for jobs, help, and developers) was around 200 in the beginning but decreased over time, and dropped to zero after they took action. This lasted for roughly one month until after the forum recovered: we see around 400 tweets about Cloudflare, twice the previous peak, and accounting for almost all such tweets that day. However, having read through these tweets, they appeared to be mainly associated with another campaign namely #stopdoghate. We thus conclude this was a short-lived outlier instead of a genuine KIWI FARMS-related peak.

The trans activist who launched the campaign was engaged at the beginning but then became much less active in posting new tweets, although she still replied to people. Her posting volume was, however, trivial compared to the overall numbers: she made only four tweets on the day the campaign started, the number then dropped quickly to only one on 4th September after Cloudflare took action, and zero thereafter. It suggests that although she sparked the campaign, she might not be the primary maintainer. We see no notable peak of tweets after the forum was completely shut down, suggesting a clear loss of interest in pursuing the campaign, both from people posting tweets and people reacting to tweets. The community seemed to get bored quickly after a few weeks when they appeared to have gotten what they wanted – '*Kiwi Farms is dead, and I am moving on to the next campaign*', tweeted the activist.

B. The Industry Response

There is no quantitative data to cover the impact on industry actors, so we switch to qualitative analysis and read through their public statements. Cloudflare stated their abuse policies on 31 August without directly mentioning the Twitter campaign [67]. In summary, the firm offers traffic proxy and DDoS protection to lots of (mostly non-paid) sites regardless of the content hosted, including KIWI FARMS. The firm maintains that abusive content alone is not an issue, and the forum – while immoral – still deserves the same protection as other customers, as long as it does not violate US law.

Although Cloudflare are entitled to refuse business from KIWI FARMS, they initially took the view that doing so because of its content would create a bad precedent, leading to unintended consequences on content regulation and making things harder for Cloudflare. This could affect the whole Internet, as Cloudflare handles a large proportion of network traffic. They did not want to get involved in policing online content, but if they had to do it they would rather do so in response to a court order instead of popular opinion. The firm previously had dropped the neo-Nazi website Daily Stormer [12] and the extremist board 8chan [13] because of their links with terrorist attacks and mass murders, and a false claim about Cloudflare's secret support. They also claimed that dropping service for KIWI FARMS would not remove the hate content, but only slow it down for a while.

Nevertheless, Cloudflare did a U-turn a few days later on 3 September 2022, announcing that they would terminate service for KIWI FARMS [26]. They explained that the escalation of the pressure campaign led to users being more aggressive, which might lead to crime. They reached out to law enforcement in multiple jurisdictions regarding potential criminal acts, but as the legal process was too slow compared to the escalating threat, they made the decision alone [26], [30]. They still claimed that following a legal process would be the correct policy, and denied that the decision was the direct result of community pressure. Cloudflare's action also inadvertently led to the termination of a neo-Nazi group in New Zealand, as it was hosted by the same company as the forum [68].

DDoS-Guard's statements about the incident told a similar story [27]. Although they can restrict access to their customers if they violate the acceptable use policy, content moderation is not their duty (except under a court order) so they do not need to determine whether every website they protect violates the law. DiamWall took the same line; they claimed that they are not responsible, and are unable to moderate content hosted on websites [28]. They also maintained that terminating services in response to public pressure is not good policy, but the case of KIWI FARMS was exceptional due to its 'revolting' content. They also noted that their actions could only delay things but not fix the root cause, as the forum could find another provider to get back online. DiamWall's statement was removed afterwards, and it is now only accessible through online archives. It is understandable that infrastructure providers such as Cloudflare and DDoS-Guard do not want to get involved in content moderation the way Facebook and Google have to, as moderation is complex, contentious and expensive.

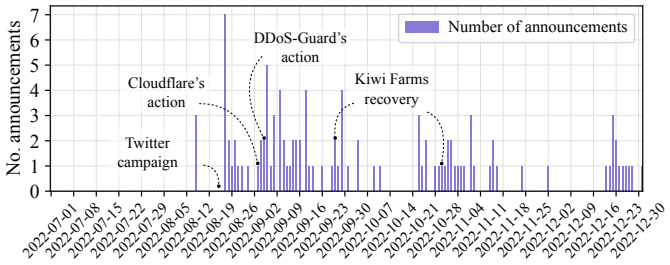

Figure 7: The number of Telegram announcements posted by the forum operators per day since the channel was created.

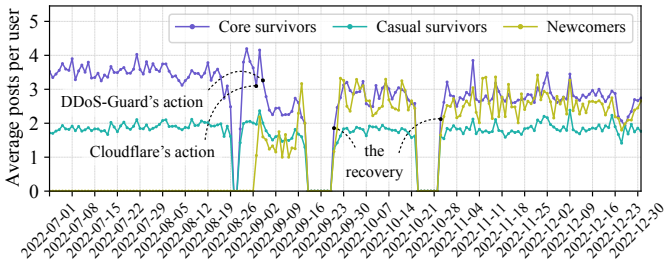

Figure 8: Number of average posts per day made by surviving actors and newcomers, who posted at least once after the event.

C. The Forum Operators

The disruption of KIWI FARMS led to a cat-and-mouse game where tech firms tried to shut it down by various means while the forum operators tried to get it back up. The forum needed DDoS protection to hide its original IP address and evade cyberattacks, so the operators first switched their third-party DDoS protection to DDoS-Guard, then DiamWall, yet these firms also resigned their business. They then attempted to build an anti-bot mechanism themselves based on HAProxy – an open-source software to stop bots, spam, and DDoS using proof-of-work [69] – and claimed to be resilient to thousands of simultaneous connections. They also changed hosting providers to VanwaTech and eventually their own firm 1776 Solutions, and attempted to route their traffic through other ISPs. They were actively maintaining infrastructure, fixing bugs, and giving instructions to users to deal with their passwords when the forum experienced a breach. The operators' effort seemed to be competent and consistent.

They posted 107 Telegram announcements during the period, mostly about when and where the forum was going to recover, the ongoing problems (e.g., DDoS attacks, industry blocks), and their plan to fix them (see Figure 7). This channel was activated after the Twitter campaign; the admins were very active, for example, sending seven consecutive messages on 23 August that mostly concerned the large DDoS attack on that day. The second peak was on 6 September after Cloudflare and DDoS-Guard's withdrawal of service, mostly about forum availability. The number of announcements then gradually decreased, especially after the second recovery, with many days having no messages. A DDoS attack hitting the forum during Christmas 2022 caught the admins' attention for a while. Their activity was inversely correlated with the forum's stability; they were less active when the site was up and running stably or when there were no new incidents.

D. The Forum Members

People sharing the same passion naturally coalesce into communities, in which some key actors may play a crucial role in influencing the ecosystem [70], [71], [72]. KIWI FARMS activity is highly skewed, with around 80%[12] of pre-disruption posts made by 8.96% most active users (5 158), while the remaining 20% posts were made by the 91.04% less active

[12] We make use of the 80/20 rule – the Pareto principle [73].

Figure 9: Average levels of toxicity, identity attack, and threat of survivors and newcomers before and after the disruption.

(52 417). There was around a 30% drop in the number of users after the disruption (as seen in Figure 4); around half of the key users (48.78%) remained engaged while only 13.05% of the less active stayed (86.95% left). There were 1 564 newcomers after the disruption. We focus on those active after the disruption, namely the 'core survivors', 'casual survivors', and 'newcomers'. On average, before the disruption, each 'core survivor' posted 22.2 times more than each 'casual survivor' (1800.99 vs 80.94 posts), while their active period (between their first post and last post) was around 2.5 times longer (1307.84 vs 516.90 days).

1) Posting Activity: Before the takedown, each core survivor made about 3.5 posts per day on average, while it was around 3 afterwards – see Figure 8. The activity of the other survivors appears consistent with the pre-disruption period; their average posts were at around 2 per day before the incident and almost unchanged afterwards. These figures suggest that the decreasing posting volume seen in Figure 4 was mainly due to users leaving the forum, instead of surviving ones largely losing interest – they engaged back quickly after the forum recovered. Newcomers posted slightly less than casual survivors before the forum was completely down on 18 September (less than 2 posts per day), yet their average

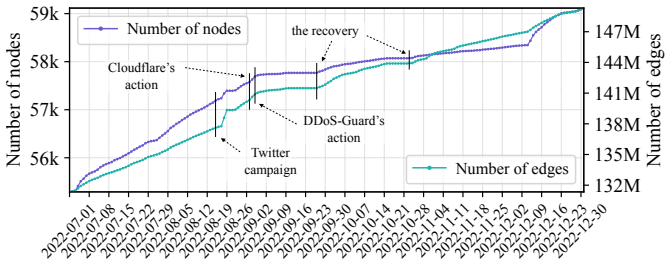

Figure 10: The number of nodes and edges in the social interaction network made by KIWI FARMS members over time.

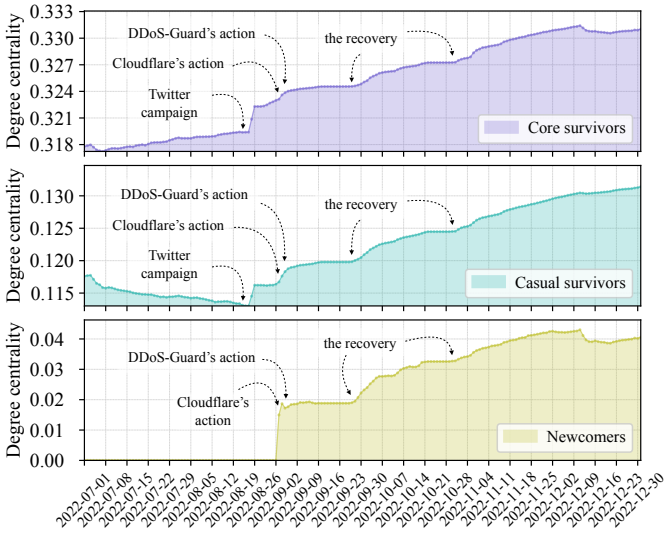

Figure 11: The degree centrality of survivors and newcomers in the network over time. Figures are in different scales.

posting volume then increased quickly. This suggests that the disruption, besides removing a very large proportion of old casual users, drew in many new users who then became roughly as active as the core survivors.

2) Toxicity Levels: KIWI FARMS has the most toxic posts among 12 extremist forums measured in previous work [21]. We thus further examine the toxicity of posts made by the surviving actors and newcomers, before and after the disruption. Figure 9 shows the average levels of *toxicity*, *identity attack* and *threat* of core survivors, casual survivors, and newcomers by days. We separate the pre-disruption and post-disruption by 3 September, when Cloudflare took action.

In general, the *toxicity*, *identity attack*, and *threat* scores were rather low as most postings are non-toxic (despite some having very high scores). There were small changes in the average scores of surviving actors, notably the peaks occurred 2 days after the campaign sparked on Twitter, with the average scores increasing significantly to around 30–50%, especially *toxicity* and *identity attack*. However, these dropped quickly a couple of days after and retreated to normal levels.

Newcomers, on the other hand, expressed a significant increase of *toxicity* and *identity attack* during the first two weeks after the disruption took place (about 2–2.5 times higher), largely surpassing surviving actors. Their scores for *threat* did not increase at that time but largely peaked after the forum first recovered on 27 September, with around 2 times higher. These activities suggest that while the surviving members were becoming more toxic when their community was under attack, new users became much more toxic for a few weeks after they engaged in the discussion before declining gradually to the same levels as old users. This is in line with the recent finding that users moving to other platforms can become more toxic than before [36].

3) Social Interactions: To measure how these survivors interact with each other, we build a social interaction network among KIWI FARMS members over time. We consider each active user as a node, with an edge between two users if they posted in the same thread (weighted by the number of such interactions) [74]. We then explore changes in the network structure with a focus on Degree Centrality, which indicates how well-connected a user is over the entire network [75].

The network had developed stably before the disruption, with around 55.3k nodes and 131.3M edges on 1 July, reaching to around 57.2k nodes and 137.6M edges just before the Twitter campaign started (see Figure 10). There was a rapid increase in both nodes and edges shortly after the Twitter campaign. It suggests that the campaign drew more actors involved in interacting with others. The Cloudflare and DDoS-Guard actions paused the network for a few weeks, yet it resumed shortly after the forum's recovery. Notably, the increasing level of edges was considerably faster than nodes (it was the opposite previously), indicating that people were getting more connected. As of 31 December 2022, the network size is 59.1k nodes and 149.3M edges.

Overall, core users are better connected than casual users. The Twitter campaign largely boosted the centrality of both core and casual survivors. Before that, while core survivors were getting more centralised over time, casual survivors were becoming less centralised. But after the campaign on Twitter, the centralisation of both steadily increased. Newcomers came into play quickly afterwards and the forum recovery also made them more centralised.

4) Discussion of the Incident: We examine how users talked about the two major involved parties (KIWI FARMS and Cloudflare) during the period by extracting posts containing case-insensitive keywords *'kiwifarm'*, *'kiwi farm'*, *'cloudflare'*, and *'cloud flare'* from KIWI FARMS, its Telegram channel, and LOLCOW FARM. Table I shows that discussions about the two parties were highly skewed and it significantly depends on the platforms. Telegram users appeared to discuss things related to KIWI FARMS far more than Cloudflare (13.3 times higher), while the ratios were less skewed for KIWI FARMS and LOLCOW FARM, with 6.7 and 7.6, respectively. Our qualitative look at messages posted on the channel reveals that people indeed cared more about recovering the forum, instead of solely blaming Cloudflare – although they did that when the disruption happened and when the forum recovered.

Although these posts accounted for a trivial contribution to the total posting volume on all three platforms as shown

Table I: Number of posts mentioning the two major involved parties during the period, with proportions of the total posts.

Platforms	Mentioning Kiwi Farms	Mentioning Cloudflare	Mentioning both parties
Kiwi Farms	10 096 (1.45%)	1 515 (0.22%)	300 (0.04%)
Telegram	3 794 (0.72%)	286 (0.05%)	44 (0.01%)
Lolcow Farm	1 494 (0.31%)	197 (0.04%)	44 (0.01%)

in Figure 4, most happened after the Twitter campaign, with almost no discussion before. The topic was popular for a short period, as shown in Figure 12. Users on both forums started discussing the incident shortly after the campaign started on 22 August. The topic was energised on both forums after Cloudflare's action on 3 September, peaking on 4 September on Kiwi Farms with over 400 and 600 posts about Kiwi Farms and Cloudflare (around 5% and 7.5% of all posts on that day), respectively. After Kiwi Farms activity was significantly reduced due to DDoS-Guard's action on 5 September, posts mentioning Kiwi Farms and Cloudflare on Lolcow Farm peaked at around 80 and 20, respectively.[13] Telegram activity regarding the incident was a bit different, as comments were only allowed after the forum was completely down; it followed the same trends as overall activity, with a peak of discussion about Kiwi Farms happening largely when the forum was inaccessible, as part of the forum discussion had moved here.

Discussion mentioning Kiwi Farms greatly exceeded those mentioning Cloudflare until the day Cloudflare took action (see the first graph in Figure 12). The pattern seen on Lolcow Farm suggests that the attention toward the incident was reflected there, although the peak did not correlate with the overall volume observed in Figure 4 as this contribution is trivial compared to the total. There were almost no posts about Cloudflare after Kiwi Farms became completely inaccessible, but there were still around 20 posts about Kiwi Farms seen on Lolcow Farm during that week. While nothing changed on Kiwi Farms during the second recovery, there was an increase in posts on Lolcow Farm about the incident, presumably as people there got the news.

Overall, attention on Kiwi Farms, Telegram, and Lolcow Farm was directed to the incident by the Twitter campaign, with posting volume peaking after the industry action. We believe it shows a genuine effect as none of the users there discussed Cloudflare and Kiwi Farms beforehand. However, the effect was temporary and almost dropped to the pre-disruption level after the second recovery: they lasted for a few days on Kiwi Farms, around one week on Lolcow Farm (partly due to many domains of Kiwi Farms being down while Lolcow Farm was still active), and a few weeks on Telegram. Users' interest was fleeting; they largely stopped talking about the incident after a few weeks.

[13] The numbers for Lolcow Farm are typically lower than Kiwi Farms as Lolcow Farm is smaller and centred on images instead of text. We do not collect images for safety and ethical reasons, but we believe the trends observed are likely indicative if not reliable.

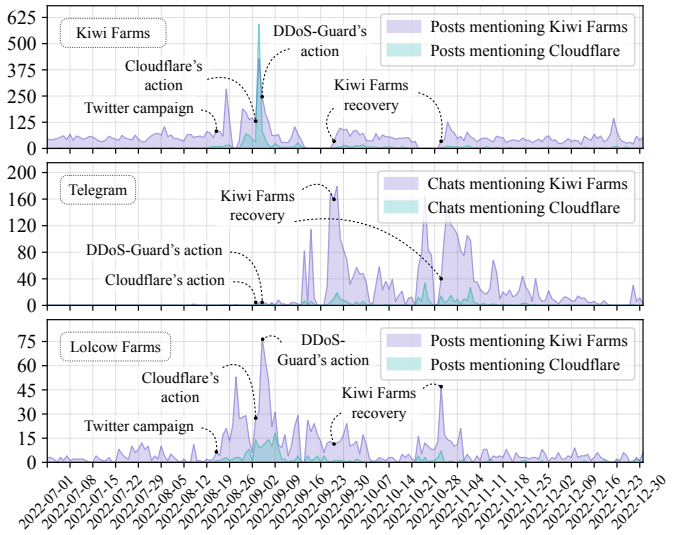

Figure 12: Discussion of the event on Kiwi Farms, its Telegram channel, and Lolcow Farm. Figure scales are different.

VI. Tensions, Challenges, and Implications

The disruption analysed in this paper could be the first time a number of infrastructure firms were involved in a collective effort to shut down a website. While deplatforming can reduce the spread of abusive content and safeguard people's mental and physical safety, and is already routine on social-media platforms like Facebook, doing so without due process raises a number of philosophical, ethical, legal, and practical issues. For this reason Meta set up its own Oversight Board.

A. The Efficacy of the Disruption

The disruption was more effective than previous DDoS attacks on the forum, as observed from our datasets. Yet the impact, although considerable, was short-lived. While part of the activity was shifted to Telegram, half of the core members returned quickly after the forum recovered. And while most casual users were shaken off, others turned up to replace them. Cutting forum activity and users by half might be a success if the goal of the campaign is just to hurt the forum, but if the objective was to "drop the forum", it has failed.

We are continuing to monitor the forum; it seems to be gradually recovering. There is a lack of data on real-world harassment caused by forum members, such as online complaints or police reports, so we are unable to measure if the campaign had any effect in mitigating the physical and mental harm inflicted on people offline.

One lesson is that while repeatedly disrupting digital infrastructure might significantly lessen the activity of online communities, it may just displace them, which has been also noted in previous work [76]. Campaigners can also get bored after a few weeks, while the disrupted community is more determined to recover their gathering place. As with the re-emergence of extremist forums like 8chan and Daily Stormer, Kiwi Farms is now back online. Deplatforming alone may be insufficient to disperse or suppress an unpleasant online

community in the long term, even when concerted action is taken by a series of tech firms over several months. It may weaken a community for a while by fragmenting their traffic and activity, and scare away casual observers, but it may also make core group members even more determined and recruit newcomers via the Streisand effect, whereby attempts at censorship can be self-defeating [11], [77].

B. Censorship versus Free Speech

One key factor may be whether a community has capable and motivated defenders who can continue to fight back by restoring disrupted services, or whether they can be somehow disabled, whether through arrest, deterrence or exhaustion. This holds whether the defenders are forum operators or distributed volunteers. So under what circumstances might the police take decisive action to decapitate an online forum, as the FBI did for example with Silk Road?

If some of a forum's members break the law, are they a dissident organisation with a few bad actors, or a terrorist group that should be hunted down? Many troublesome organisations do attract hot-headed young members, and activists from animal-rights activists and climate-change protesters through to trade union organisers do occasionally fall foul of the law. But whether they are labelled as terrorists or extremists is often a political matter. Taking down a website on which a whole community relies will often be hard to defend as a proportionate and necessary law-enforcement action. The threat of legal action can be countered by the operator denouncing whatever specific crimes were complained of. In this case, the KIWI FARMS founder denounced SWAT harassment and other blatant criminality [50]. Indeed, a competent provocateur will stop just short of the point at which their actions will call down a vigorous police response.

The freedom of speech protected by the US First Amendment [78] is in clear tension with the mental and physical security of harassment victims. The Supreme Court has over time established tests to determine what speech is protected and what is not, including clear and present danger [79], a sole tendency to incite or cause illegal activity [80], preferred freedoms [81], and compelling state interest [82]; however, the line drawn between them is not always clear-cut. Other countries are more restrictive, with France and Germany banning Nazi symbolism and Turkey banning material disrespectful of Mustafa Kemal Atatürk. In the debates over the Online Safety Bill currently before the UK Parliament, the Government at one point proposed to ban 'legal but harmful' speech online, while not making these speech acts unlawful face-to-face. These proposals related to websites encouraging eating disorders or self-harm. Following the tragic suicide of a teenage girl, tech firms are under pressure to censor such material in the UK using their terms of service or by tweaking their recommendation algorithms.

There are additional implications in taking down platforms whose content is harmful but not explicitly illegal. Requiring firms to do this, as was proposed in the Online Safety Bill, will drastically expand online content regulation. The UK legislation hands the censor's power to the head of Ofcom, the broadcast regulator, who is a political appointee. It will predictably lead to overblocking and invite abuse of power by government officials or big tech firms, who may suppress legitimate voices or dissenting opinions. There is an obvious risk of individuals or groups being unfairly targeted for political or ideological reasons.

C. The Role of Industry in Content Moderation

The rapid increase of cybercrime-as-a-service throughout the 2010s makes attacks easier than ever. A teenager with as little as $10 can use a DDoS-for-hire service to knock your website offline [83], so controversial websites depend on the grace and favour of a large hosting company or a specialist DDoS prevention contractor. This is just one aspect of a broader trend in tech: that the Internet is becoming more centralised around a small number of big firms, ranging from online social platforms, hosting companies, transit networks, to service providers and exchange points [84]. Many of them claim to be committed to fighting hate, harassment, and abuse yet some are disproportionately responsible for serving online bad content [76], and the effort they put into the fight is variable [85], [86]. Now that activists have pressured infrastructure providers to act as content moderators, policymakers will be tempted too. Some may stand up to political or social pressure, because moderation is both expensive and difficult, but others may fold from time to time because of political pressure or legal compulsion. This would undermine the end-to-end principle of the Internet, as enshrined for example in COPA s 230 in the USA and in the EU's Net Neutrality Law [87].

Private companies must comply and remove illegal content from their infrastructure when directed to do so by a court order. However, deplatforming KIWI FARMS or any other customers does not violate the principle of free speech. It is essentially a contractual matter; they have the right to cease their support for a website that violates their policies. Infrastructure providers may occasionally need to work expediently with law enforcement in the case of an imminent threat to life. Most providers have worked out ways of doing this, but the mechanisms can be too sluggish. Cloudflare attempted to collaborate with law enforcement to sort out the case of KIWI FARMS, yet the process could not keep up with the escalating threats and it ended up taking unilateral action, relying on its terms of service [26]. In an ideal world, we would have an international legal framework for taking down websites that host illegal content or that promote crime; unfortunately, this framework does not exist.

The Budapest Convention criminalises some material on which all states agree, such as child sex abuse images, but even there the boundaries are contested [88]. Online drug markets such as Silk Road and Hansa Market have been taken down because of other laws – drug laws – that also enjoy international standardisation and collaboration. Copyright infringement also gets the attention of international treaties and coordinated action by tech majors, though civil law plays a greater role here than criminal law. Then there is material about which some

states feel strongly but others do not; 'one man's freedom fighter is another man's terrorist'. And then there's a vast swamp of fake news, animal cruelty, conspiracy theories, and other material that many find unpleasant or distressing, and which social networks moderate for the comfort of both their users and their advertisers. Legislators occasionally call for better policing of some of this content.

D. Policy Implications

The UK Online Safety Bill proposes a new regulator who will be able to apply for a court order mandating that tech firms disrupt an objectionable online activity [43]. One might imagine Ofcom deciding to take down KIWI FARMS if their target had been a resident of Britain rather than Canada, and going to the various tech firms that were involved in the disruption we describe here, serving them one after another with orders signed by a judge in the High Court in London. Even if all the companies were to comply, rather than appealing or just ignoring the court altogether, it is hard to see how such an operation could be anything like as swift, coordinated or effective as the action taken on their own initiative by tech companies that we describe here. Where the censor's OODA loop – the process by which it can observe, orient, decide and act – involves a government agency assessing the effects of each intervention and then going to court to order the next one, the time constant would stretch from hours to months. And in any case, government interventions in this field are often significant but rather short-lived [14], [15]. One reason they can be effective is that the maintainer of a blatantly illegal website may be arrested and jailed, as happened with Silk Road. With a forum like KIWI FARMS, whose operator has denounced criminal acts perpetrated via his infrastructure [50], that option may simply not be available.

Previous work has also explored why governments are less able to take down bad sites than private actors [11]; that work analysed single websites with clearly illegal content, such as those hosting malware, phishing lures or sex-abuse images. This study shows why taking down an active community is likely to be even harder. Even when several tech firms roll their sleeves up and try to suppress a community some of whose members have indulged in crime and against whom there is an industry consensus, the net effect may be modest at best. Our case study may be the best result that could be expected for online censorship, but it only cut the users, posts, threads and traffic by about half. Our findings suggest that using content moderation law to suppress an unpleasant online community may be very challenging.

VII. CONCLUSION

Online communities may not only act as a discussion place but provide mutual support for members who share common values. For some, it may be where they hang out; for others, it may become part of their identity. Legislators who propose to ban an online community might consider precedents such as Britain's ban on Provisional Sinn Féin from 1988–94 due to its support for the Provisional IRA during the Troubles, or the bans on the Muslim Brotherhood enacted by various Arab regimes.[14] Declaring a community to be illegal and thus forcing it underground may foster paranoid worldviews, increase signals associated with toxicity and radicalisation [44], [36] and have many other unintended consequences. The KIWI FARMS disruption, which involved a substantial effort by the industry, is perhaps the best outcome that could be expected even if the censor were agile, competent and persistent. Yet this has demonstrated that merely trying to deplatform an active online community is not enough to deal effectively with online hate and harassment.

We believe the harms and threats associated with online hate communities may justify action despite the right to free speech. But within the framework of the EU and the Council of Europe which is based on the European Convention on Human Rights, such action will have to be justified as proportionate, necessary and in accordance with the law. It is unlikely that taking down a whole community because of a crime committed by a single member can be proportionate. For a takedown to be justified as necessary, it must also be effective, and this case study shows how high a bar that could be. For a takedown to be in accordance with the law, it cannot simply be a response to public pressure. There must be a law or regulation that determines predictably whether a specific piece of content is illegal, and a judge or other neutral finder of fact would have to be involved.

The last time a Labour government won power in Britain, it won on a promise to be 'Tough on Crime, and Tough on the Causes of Crime'. Some scholars of online abuse are now coming to a similar conclusion that the issue may demand a more nuanced approach [3], [21]: as well as the targeted removal of content that passes an objective threshold of illegality, the private sector and governments should collaborate to combine takedowns with measures such as education and psycho-social support [89]. And where the illegality involves violence, it is even more vital to work with local police forces and social workers rather than just attacking the online symptoms [88].

There are multiple research programmes and field experiments on effective ways to detox young men from misogynistic attitudes, whether in youth clubs and other small groups, at the scale of schools, or even by gamifying the identification of propaganda that promotes hate. But most countries still lack a unifying strategy for violence reduction [90]. In both the US and the UK, for example, while incel-related violence against women falls under the formal definition of terrorism, it is excluded from police counterterrorism practice, and the politicisation of misogyny has made this a tussle space in which political leaders and police chiefs have difficulty in taking effective action. In turbulent debates, policymakers should first ask which tools are likely to work, and it is in this context that we offer the present case study.

[14] During the Sinn Féin ban, it was illegal to transmit the voice or image of one of their spokesmen in Britain, so the BBC and other TV stations simply employed actors to read the words of Gerry Adams and Martin McGuinness.

ACKNOWLEDGMENTS

We are grateful to Richard Clayton, Yi Ting Chua, Ben Collier, Tina Marjanov, Konstantinos Ioannidis, Ilia Shumailov, and our colleagues at the Cambridge Cybercrime Centre for their useful feedback in the early draft of the paper.

REFERENCES

[1] M. Mondal, L. A. Silva, and F. Benevenuto, "A measurement study of hate speech in social media," in *Proceedings of the ACM Conference on Hypertext and Social Media (HT)*, 2017.

[2] S. A. Aghazadeh, A. Burns, J. Chu, H. Feigenblatt, E. Laribee, L. Maynard, A. L. Meyers, J. L. O'Brien, and L. Rufus, "Gamergate: A case study in online harassment," *Online harassment*, 2018.

[3] D. Kumar, J. Hancock, K. Thomas, and Z. Durumeric, "Understanding the behaviors of toxic accounts on reddit," in *Proceedings of the ACM Web Conference (WWW)*, 2023.

[4] M. Singhal, C. Ling, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "Sok: content moderation in social media, from guidelines to enforcement, and research to practice," *arXiv preprint arXiv:2206.14855*, 2022.

[5] E. de Keulenaar, A. Glyn Burton, and I. Kisjes, "Deplatforming, demotion and folk theories of big tech persecution." *Revista Fronteiras*, 2021.

[6] R. Rogers, "Deplatforming: Following extreme internet celebrities to telegram and alternative social media," *European Journal of Communication*, 2020.

[7] H. Habib, M. B. Musa, F. Zaffar, and R. Nithyanand, "To act or react: Investigating proactive strategies for online community moderation," *arXiv preprint arXiv:1906.11932*, 2019.

[8] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi, "The social world of content abusers in community question answering," in *Proceedings of the ACM World Wide Web Conference (WWW)*, 2015.

[9] The Verge, "Twitter will label COVID-19 vaccine misinformation and enforce a strike system," 2021. [Online]. Available: https://theverge.com/2021/3/1/22307919/twitter-covid-19-vaccine-labels-five-strike-system

[10] ——, "Facebook will remove COVID-19 vaccine misinformation," 2020. [Online]. Available: https://theverge.com/2020/12/3/22150425/facebook-covid-19-vaccine-coronavirus-misinformation-ban

[11] A. Hutchings, R. Clayton, and R. Anderson, "Taking down websites to prevent crime," in *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, 2016.

[12] Cloudflare, "Why We Terminated Daily Stormer," 16 August 2017. [Online]. Available: https://blog.cloudflare.com/why-we-terminated-daily-stormer/

[13] ——, "Terminating Service for 8Chan," 05 August 2019. [Online]. Available: https://blog.cloudflare.com/terminating-service-for-8chan/

[14] B. Collier, D. R. Thomas, R. Clayton, and A. Hutchings, "Booting the booters: Evaluating the effects of police interventions in the market for denial-of-service attacks," in *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2019.

[15] D. Kopp, M. Wichtlhuber, I. Poese, J. Santanna, O. Hohlfeld, and C. Dietzel, "Ddos hide & seek: on the effectiveness of a booter services takedown," in *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2019.

[16] Bleeping Computer, "FBI seized domains linked to 48 DDoS-for-hire service platforms," 2022. [Online]. Available: https://bleepingcomputer.com/news/security/fbi-seized-domains-linked-to-48-ddos-for-hire-service-platforms/

[17] Department of Justice, "United States Leads Seizure of One of the World's Largest Hacker Forums and Arrests Administrator," 12 April 2022. [Online]. Available: https://justice.gov/opa/pr/united-states-leads-seizure-one-world-s-largest-hacker-forums-and-arrests-administrator

[18] "United States v. Ross William Ulbricht." [Online]. Available: https://caselaw.findlaw.com/us-2nd-circuit/1862572.html

[19] K. Soska and N. Christin, "Measuring the longitudinal evolution of the online anonymous marketplace ecosystem," in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2015.

[20] M. Pless, "Kiwi farms, the web's biggest stalker community," *New York Magazine*, 2016.

[21] A. V. Vu, L. Wilson, Y. T. Chua, I. Shumailov, and R. Anderson, "Extremebb: Enabling large-scale research into extremism, the manosphere and their correlation by online forum data," *arXiv preprint arXiv:2111.04479*, 2021.

[22] Sam Ambreen, "Kiwi farms linked to at least 2 murders and 4 suicides," 2019. [Online]. Available: https://samambreen.wordpress.com/2019/03/08/fyi-kiwi-farms-linked-to-at-least-2-murders-and-4-suicides/

[23] Wired, "The End of Kiwi Farms, the Web's Most Notorious Stalker Site," 2022. [Online]. Available: https://wired.com/story/keffals-kiwifarms-cloudflare-blocked-clara-sorrenti/

[24] Daily Dot, "Pressure grows on Cloudflare to drop Kiwi Farms after latest doxing campaign," 2022. [Online]. Available: https://dailydot.com/debug/kiwi-farms-cloudflare-keffals/

[25] Vice Motherboard, "People Are Demanding That Cloudflare Drop Kiwi Farms," 2022. [Online]. Available: https://vice.com/en/article/z3434y/people-are-demanding-that-cloudflare-drop-kiwi-farms/

[26] Cloudflare, "Blocking Kiwifarms," 2022. [Online]. Available: https://blog.cloudflare.com/kiwifarms-blocked/

[27] DDoS Guard, "DDoS-Guard terminating services for Kiwi Farms," 2022. [Online]. Available: https://ddos-guard.net/en/info/blog-detail/ddos-guard-terminating-services-for-kiwi-farms/

[28] DiamWall, "Service Continuation of Kiwi Farms," 2022. [Online]. Available: https://diamwall.com/blog/service-continuation-of-kiwi-farms/

[29] David Covucci, "Kiwi Farms gets booted from another major domain," 2022. [Online]. Available: https://dailydot.com/debug/kiwi-farms-top/

[30] J. Menn and T. Lorenz, "Under pressure, security firm cloudflare drops kiwi farms website," *The Washington Post*, 2022.

[31] S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, "Evaluating the effectiveness of deplatforming as a moderation strategy on twitter," in *Proceedings of the ACM on Human-Computer Interaction (HCI)*, 2021.

[32] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," in *Proceedings of the ACM on Human-Computer Interaction (HCI)*, 2017.

[33] H. M. Saleem and D. Ruths, "The aftermath of disbanding an online hateful community," *arXiv preprint arXiv:1804.07354*, 2018.

[34] H. Innes and M. Innes, "De-platforming disinformation: conspiracy theories and their control," *Information, Communication & Society*, 2021.

[35] A. Rauchfleisch and J. Kaiser, "Deplatforming the far-right: An analysis of youtube and bitchute," *Available at SSRN 3867818*, 2021.

[36] S. Ali, M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Understanding the effect of deplatforming on social networks," in *Proceedings of the ACM Web Science Conference (WebSci)*, 2021.

[37] K. Bryanov, D. Vasina, Y. Pankova, and V. Pakholkov, "The other side of deplatforming: Right-wing telegram in the wake of trump's twitter ouster," in *Proceedings of the International Conference on Digital Transformation and Global Society (DTGS)*, 2022.

[38] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2021.

[39] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler, "Characterizations of online harassment: Comparing policies across social media platforms," in *Proceedings of the ACM International Conference on Supporting Group Work (GROUP)*, 2016.

[40] R. Jiménez Durán, "The economics of content moderation: Theory and experimental evidence from hate speech on twitter," *Available at SSRN*, 2022.

[41] B. Fishman, "Dual-use regulation: Managing hate and terrorism online before and after section 230 reform," *Brookings Institution*, 2023.

[42] F. Schauer, "The exceptional first amendment," *KSG Working Paper No. RWP05-021*, 2005.

[43] R. Anderson and S. Gilbert, "The online safety bill," *Policy Brief, Bennett Institute for Public Policy*, 2023.

[44] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West, "Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels," in *Proceedings of the ACM on Human-Computer Interaction (HCI)*, 2021.

[45] G. Russo, L. Verginer, M. H. Ribeiro, and G. Casiraghi, "Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning," *arXiv preprint arXiv:2209.09803*, 2022.

[46] C. Buntain, M. Innes, T. Mitts, and J. Shapiro, "Cross-platform reactions to the post-january 6 deplatforming," *Journal of Quantitative Description: Digital Media*, 2023.

[47] A. Mekacher, M. Falkenberg, and A. Baronchelli, "The systemic impact of deplatforming on social media," *arXiv preprint arXiv:2303.11147*, 2023.

[48] C. Monti, M. Cinelli, C. Valensise, W. Quattrociocchi, and M. Starnini, "Online conspiracy communities are more resilient to deplatforming," *arXiv preprint arXiv:2303.12115*, 2023.

[49] I. Goldstein, L. Edelson, D. McCoy, and T. Lauinger, "Understanding the (in) effectiveness of content moderation: A case study of facebook in the context of the us capitol riot," *arXiv preprint arXiv:2301.02737*, 2023.

[50] Kiwi Farms, "Principles of the Kiwi Farms," 2022. [Online]. Available: https://kiwifarms.net/threads/principles-of-the-kiwi-farms.127111/

[51] ——, "XenForo has revoked our license," 2021. [Online]. Available: https://kiwifarms.net/threads/xenforo-has-revoked-our-license.106654/

[52] Vice Motherboard, "Notorious website kiwi farms loses its domain registrar," 13 July 2021. [Online]. Available: https://vice.com/en/article/pkbmam/notorious-website-kiwi-farms-loses-its-domain-registrar/

[53] Heatst, "Notorious Forum Kiwi Farms Closed Following Alleged Harassment of Founder's Family," 24 January 2017. [Online]. Available: https://web.archive.org/web/20170129024109/http://heatst.com/tech/notorious-forum-kiwi-farms-closed-following-alleged-harassment-of-founders-family/

[54] "Trans Twitch streamer Keffals says she was swatted and arrested by police in Ontario," 2022. [Online]. Available: https://nbcnews.com/pop-culture/pop-culture-news/trans-twitch-streamer-keffals-says-was-swatted-arrested-police-ontario-rcna42533

[55] Keffals, "Keffals led a protest against Cloudflare to drop the Kiwi Farms forum," August 2022. [Online]. Available: https://twitter.com/keffals/status/1560730882012598274

[56] The Verge, "Kiwi Farms has been scrubbed from the Internet Archive," 2022. [Online]. Available: https://theverge.com/2022/9/7/23341051/kiwi-farms-internet-archive-backup-removal/

[57] Vice Motherboard, "QAnon's Jim Watkins Tried to Save Kiwi Farms. Now His Site 8Kun Is Down." 2022. [Online]. Available: https://vice.com/en/article/ake7q5/kiwi-farms-jim-watkins-8kun

[58] Perspective API, "Attributes & Languages," 2022. [Online]. Available: https://developers.perspectiveapi.com

[59] Similarweb, "Top Competitors of Kiwi Farms," 2023. [Online]. Available: https://similarweb.com/website/kiwifarms.net/#competitors

[60] Semrush, "Top Competitors of Kiwi Farms," 2023. [Online]. Available: https://semrush.com/analytics/organic/competitors/?sortField=&sortDirection=desc&db=us&q=kiwifarms.net&searchType=domain

[61] P. Doerfler, A. Forte, E. De Cristofaro, G. Stringhini, J. Blackburn, and D. McCoy, ""i'm a professor, which isn't usually a dangerous job": Internet-facilitated harassment and its impact on researchers," in *Proceedings of the ACM on Human-Computer Interaction (HCI)*, 2021.

[62] TechCrunch, "Web scraping is legal, U.S. Appeals Court reaffirms," 2022. [Online]. Available: https://cacm.acm.org/news/260233-web-scraping-is-legal-us-appeals-court-reaffirms/fulltext/

[63] British Society of Criminology, "Statement of ethics," 2015. [Online]. Available: http://britsoccrim.org/ethics/

[64] A. E. Marwick, L. Blackwell, and K. Lo, "Best practices for conducting risky research and protecting yourself from online harassment (data & society guide)," *New York: Data and Society Institute*, 2016.

[65] R. Bhalerao, V. Hamilton, A. McDonald, E. M. Redmiles, and A. Strohmayer, "Ethical practices for security research with at-risk populations," in *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022.

[66] S. C. Jansen and B. Martin, "The streisand effect and censorship backfire," 2015.

[67] Cloudflare, "Cloudflare's abuse policies & approach," 31 August 2022. [Online]. Available: https://blog.cloudflare.com/cloudflares-abuse-policies-and-approach/

[68] Keffals, "Cloudflare inadvertently de-platformed a neo-nazi group based in New Zealand," September 2022. [Online]. Available: https://twitter.com/keffals/status/1566335693663731712

[69] Thomas Lynch, "HAProxy Protection," 2023. [Online]. Available: https://gitgud.io/fatchan/haproxy-protection/

[70] J. Hughes, B. Collier, and A. Hutchings, "From playing games to committing crimes: A multi-technique approach to predicting key actors on an online gaming forum," in *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, 2019.

[71] S. G. van de Weijer, T. J. Holt, and E. R. Leukfeldt, "Heterogeneity in trajectories of cybercriminals: A longitudinal analyses of web defacements," *Computers in Human Behavior Reports*, 2021.

[72] A. V. Vu, J. Hughes, I. Pete, B. Collier, Y. T. Chua, I. Shumailov, and A. Hutchings, "Turning Up the Dial: the Evolution of a Cybercrime Market through Set-up, Stable, and Covid-19 eras," in *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2020.

[73] R. Sanders, "The pareto principle: its use and abuse," *Journal of Services Marketing*, 1987.

[74] I. Pete, J. Hughes, Y. T. Chua, and M. Bada, "A social network analysis and comparison of six dark web forums," in *Proceedings of the IEEE European Symposium on Security and Privacy workshops (EuroS&PW)*, 2020.

[75] M. Newman, *Networks*, 2018.

[76] C. Han, D. Kumar, and Z. Durumeric, "On the infrastructure providers that support misinformation websites," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2022.

[77] Y. T. Chua, S. Parkin, M. Edwards, D. Oliveira, S. Schiffner, G. Tyson, and A. Hutchings, "Identifying unintended harms of cybersecurity countermeasures," in *Proceedings of the APWG Symposium on Electronic Crime Research (eCrime)*, 2019.

[78] R. Casey, "John stuart mill and social media: Evaluating the ethics of de-platforming," *U. Cent. Fla. Dep't Legal Stud. LJ*, 2021.

[79] "Schenck v. United States, 249 U.S. 47." 1919.

[80] "Abrams v. United States, 250 U.S. 616," 1919.

[81] "Jones v. City of Opelika, 316 U.S. 584; 319 U.S. 103," 1942-1943.

[82] "Korematsu v. United States, 323 U.S. 214," 1944.

[83] A. Hutchings and R. Clayton, "Exploring the provision of online booter services," *Deviant Behavior*, 2016.

[84] T. Mirrlees, "Gafam and hate content moderation: Deplatforming and deleting the alt-right," in *Media and law: Between free speech and censorship*, 2021.

[85] D. Konikoff, "Gatekeepers of toxicity: Reconceptualizing twitter's abuse and hate speech policies," *Policy & Internet*, 2021.

[86] D. G. Heslep and P. Berge, "Mapping discord's darkside: Distributed hate networks on disboard," *New Media & Society*, 2021.

[87] "All you need to know about net neutrality rules in the eu." [Online]. Available: https://berec.europa.eu/en/all-you-need-to-know-about-net-neutrality-rules-in-the-eu

[88] R. Anderson, "Chat control or child protection?" *arXiv preprint arXiv:2210.08958*, 2022.

[89] C. Lally and R. Bermingham, "Online extremism," *UK Parliament Research Briefing*, 2020.

[90] L. Bates, *Men Who Hate Women – The Extremism Nobody is Talking About*, 2020.