Evaluating the Effects of Temporary Twitter Suspensions of High-Profile Twitter Users

Mika Desblancs-Patel

Master of Science
Cognitive Science
School of Informatics
University of Edinburgh
2023

Abstract

Heavy mediatization of high-profile Twitter suspensions, their controversial reception, the popularity of suspended users' initial post-reinstatement tweets, and the fear that suspending a user will galvanize their supporters and make their ideas more popular, have made people question whether suspensions, paradoxically, serve as catalysts for a targeted user's popularity instead of having the intended punitive effects. In this dissertation, we explore the effects that temporary suspensions of high-profile users have on their post-reinstatement popularity, along with their supporter demographic and activity. We present a case study of five famous influencers, who were *temporarily* deplatformed from Twitter in 2022 - Tucker Carlson, Jordan Peterson, Marjorie Taylor Greene, Charlie Kirk, and Dave Rubin.

Working with data from tweets posted before and after each user's suspension, we find that while the initial post-reinstatement tweet engagement metrics were the highest in our sample, their popularity isn't sustained over time. Our evidence shows their post-suspension tweet support is driven by new users, and not galvanized old supporters. Furthermore, supporters greatly reduce their discussions about the banned user during the suspension. Finally, their toxicity levels remain mostly stable, and on par with that of average Twitter users in periods before, during, and after the suspension.

Overall, we present novel research on the effects of temporary suspension measures, and contribute to the ongoing conversation about the effectiveness of deplatforming measures and influence of high-profile social media influencers.

Research Ethics Approval

This project obtained approval from the Informatics Research Ethics committee.
Ethics application number: rt #7591 ID 425077
Date when approval was obtained: 2023-08-11

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Mika Desblancs-Patel*)

Acknowledgements

I would like to, first and foremost, thank my supervisors Dr. Walid Magdy, Dr. Christopher Barrie, along with Dr. Youssef Al Hariri for the time, and immense amount of help they gave me during this entire project. I would have achieved only half of what was done without their suggestions, insights, and unwavering commitment to providing me with the help and guidance I needed throughout these past months. I am proud of our work, and my satisfaction was only possible thanks to their never ending support.

I am also indebted to my parents, Éric and Tara, for constantly checking in on my well-being, asking for ”signs of life” and giving me the opportunity to achieve my academic ambitions. I am also indebted to my brother Dorian for his guidance when helping me navigate the academic and professional sphere, along with his ever insightful hot-takes. He pushed me to take my first coding class and might be the reason I’m not unhappily in law school. For that, I will always be grateful. I want to also thank my sister Zoé for reminding me to take the time to love myself, take care of myself, and being my always-there emotional pillar. Your drive inspires me, and pushes me accomplish my dreams.

Finally, I would also like to thank my friends, those I met while in Edinburgh, and my cohort. Sofia, Marin, Léa-Mirana, Tiphaine and Victor, thank you for pushing me to being my most authentic self.

This dissertation is dedicated to my community of friends and my family.

Table of Contents

Chapter 1

Introduction

As mentioned in my project proposal, in recent years, social media platforms like Twitter have faced harsh criticisms from consumers and government authorities due to their inability to curb the spread of fake news[38], harboring of extremist content, calls for violence[3] and hosting toxic communities who spread hateful ideologies[19]. Facing growing demands and pressure to clean up harmful content on their platform, the Twitter have taken it upon themselves to introduce their own terms and conditions policies[1]. To punish users who violate their content policy, the platform have resorted to suspending their accounts either temporarily, permanently, or until a tweet has been deleted. Recently, Twitter suspended the accounts of numerous high-profile users like Jordan Peterson, Marjorie Taylor Green and Tucker Carlson.

But do Twitter suspensions **actually have** the intended punitive effects? All the previously mentioned suspensions were heavily mediatized in the days following their announcement. Their number of followers increased greatly after their reinstatement (see Figure A.1). Jordan B Peterson, Greene, and Carlsons's first tweets back were heavily liked retweeted and quoted. Furthermore, all saw the suspension as evidence they were censored [35][61][53][1].

Combined, the heavy mediatization of their suspension, their increase in Twitter followers, and their high engagement count of their initial tweets back suggests that temporary suspensions might not have the intended punitive effect. In fact, the opposite might be true: a temporary suspension might serve as a catalyst for a user's popularity at their return.

[1] Twitter's content policy: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate

1.1 Project Goals & Research Questions

The goal of this project is to evaluate the effect that temporary Twitter suspensions have on high-profile political influencers which were *temporarily* suspended from the platform. We are especially interested in determining whether suspensions increased their popularity and whether it galvanized and made their supporter base more toxic. Below, we formalize our research questions and articulate our initial hypothesis.

Given the surge in popularity that the initial tweets from suspended high-profile users received right after the ban and users positioning themselves as martyrs, we can ask ourselves whether suspensions in fact made users more popular. Importantly, because we focus exclusively on *temporary* suspensions, we have data about their tweet engagement metrics before and after their suspension. How does it evolve? We notice the suspensions had a large amount of media attention and user's initial tweets back were popular.

RQ1: How does a suspended user's tweet engagement change for tweets posted before the suspension and those posted after?

H1: Suspended user's tweet popularity increases after their reinstatement.

The evolution of a suspended user's tweets' engagement metrics only tell a partial story. We are also interested in understanding whether their suspension galvanized support among their supporters. Many suspended users claimed they were being censored, and their suspensions might be proof of that for their supporters.

RQ2: Do suspensions increase the level of support that a user receives from their pre-suspensions supporter base ?

H2: Supporters will increase their level of support after a user's reinstatement compared to before.

For tweets posted before and after a user's suspension, engagement signals are quite clear; we have access to the number of quoters, likers and retweeters. However, during their suspension, supporters are no longer able to engage in these ways. It is unclear whether they even talk about them at all.

RQ3: Is the amount of chatter about a suspended user among supporters affected by the suspension?

H3: Here, we hypothesize first about the evolution of the number of mentions overtime and which supporters continue to mention a user during their suspension.

1. The overall number of mentions of a user among supporters will increase around the time of their suspension, decrease during their suspension, and increase again after their reinstatement to higher than pre-suspension mention levels.
2. Furthermore, the most ardent supporters will mention a user during their suspension more than casual supporters.

Finally, one of the big goals of a suspension due to violating content policies against hateful rhetoric, is to limit the amount of toxicity on the platform. Previous research[33] covered in Section 2 suggests that after a user's suspension, toxicity levels among their supporters decreased.

RQ4: Do suspensions change the level of toxicity among supporters before, during and after the suspension ?

H4: We hypothesize first about the evolution of hate-speech levels around a suspension, then about their general levels.

1. Toxicity levels will decrease after the suspension and increase again after the suspended user's reinstatement
2. Toxicity levels among supporters of a high-profile user are generally higher than random Twitter users' levels.

1.2 Structure of the Dissertation

The remaining content of this dissertation is separated into an additional four chapter. In **Chapter 2** we present background knowledge of Twitter's role in shaping public discourse along with relevant suspended users. We also present previous studies on social media user and community bans. Finally, we discuss the prevalence and ways to measure online toxicity. In **Chapter 3** we detail the data collection procedure and the work carried out in the experimental design portion of the project. In **Chapter 4** we present the results of the experiments and note the findings most relevant for the evaluation of our hypothesis. In **Chapter 5** we attempt to provide a general narrative explaining our findings. Finally, in **Chapter 6**, we provide a general summary of our results, limitations, and speculate on future work.

Chapter 2

Background and Related Work

In this section I present some background information on the role of Twitter in shaping public discourse, relevant suspended users, and Twitter's content moderation policy I will then address previous work on the effects of suspensions on social media platforms.

2.1 New Media and Online Political Influence

One of the central reasons Twitter suspensions have garnered so much attention and understanding their efficacy is so important, is that the platform plays a central role in shaping public discourse. In the US, in 2022, Twitter has around 77.75 million active users[62]. Furthermore, one of the biggest reasons users visit the platform is to stay up to date with the news[48] with 83% of users tweeting about the news and 55% of users coming to Twitter for the news over other social media platforms. Given its popularity and almost non-existent barrier to entry, the website attracts contentious figures who argue their ideas are blocked or misrepresented by mainstream news outlets and who turn to Twitter to spread their beliefs.

In this section I will present five high-profile political figures who were suspended for violating Twitter's content policy, and whose suspension was heavily mediatized.

Tucker Carlson is a conservative American political commentator, host of 'Tucker Carlson Tonight' on Fox News between 2016 and 2023. It was one of the most watched cable news programs in the US over that period and described by the New York Times as maybe "the most racist show in the history of cable news - and also, by some measure, the most successful", pushing far-right and fringe ideas into the mainstream[17]. He has criticized Twitter for censoring conservative voices[18] and expressed disdain for mainstream media outlets[63]. He currently has over 9.3 million Twitter followers[1].

He was suspended from Twitter around the 23d of March for relaying screenshots and expressing support to two transphobic tweets from ’Babylon Bee’, a conservative satire media, and Charlie Kirk, another conservative political commentator. Both targeted Rachel Levine, a transgender U.S official. At the time of his suspension, he had just under 5 million followers. He heavily criticized his suspension, stating on his show that he was silenced for sharing ”factual statements”[35]. He was reinstated around the 25^{th} of April with the tweet in question deleted from his Twitter timeline[54]. In the month following his reinstatement, he gained around 300,000 followers (see Figure A.1a for the evolution of his Twitter following).

Charlie Kirk is an American conservative activist and radio talk show host famous for co-founding Turning Point USA along with subsidiaries like Turning Point Action and Students for Trump[2]. He amassed a large gathering after founding Turning Point USA, an American conservative youth group which raised almost 39.8 million dollars in 2020[64]. Charlie Kirk has been a vocal about his beliefs that the 2020 election was rife with voter fraud, was accused of promoting the January 6 rally and of spreading misinformation about COVID-19[64]. He currently has over 2.4 million Twitter followers on his personal account[3].

His Twitter account was suspended for the second time around the 22^{nd} of March 2022 for the transphobic tweet Tucker Carlson later shared[5]. About his suspension, he claimed he would ”NEVER apologize for speaking the truth” and that ”this type of censorship that will ultimately destroy Twitter”[42]. He was reinstated 5 weeks later around the 28th of April, 2022[54]. In the month following his reinstatement, he gained around 50,000 followers (see Figure A.1b for the evolution of his Twitter following).

Marjorie Taylor Greene Marjorie Taylor Greene is an American far-right politican who rose to prominence after being elected to be a US representative for Georgia’s 14th congressional district in 2021, and then again in 2022. She is a supporter of the QAnon conspiracy theory, suggested wildfires were caused by Jewish space lasers and that 9/11 was a hoax[24]. As a US representative, she has both a personal and congressional Twitter account, from which she posts infrequently. She currently has over 820,000 Twitter followers on her personal account[4].

Her personal account was permanently suspended around the 2^{nd} of January for

[1] https://twitter.com/TuckerCarlson

[2] https://www.tpusa.com/bio/charliekirk

[3] https://twitter.com/charliekirk11

[4] https://twitter.com/mtgreenee

repeatedly spreading COVID-19 misinformation[1]. Before that, she had been temporarily suspended three times[12][11][52]. In a statement she posted on Telegram right after, she claimed ”social media platforms can’t stop the truth from being spread far and wide” and accused Twitter of supporting ”a Communist revolution”[1].

While her last suspension was supposed to be permanent, her personal account was reinstated by Elon Musk around the 21st of November 2022 as part of a wave of reinstatements of banned accounts which saw Donal Trump, Kanye West and the satire site Babylon Bee return to the platform[25]. In the month following her last reinstatement, her account gained around 82,000 followers (see Figure A.1c for the evolution of her Twitter following).

Jordan Peterson is a Canadian psychologist turned media figure. He rose to prominence in 2016 when criticizing political correctness and the newly proposed Bill C-16[10] which introduced gender expression and gender identity as protected grounds to the Canadian Human Rights Act[23]. In his lectures and interviews, widely popular on the video streaming platform YouTube, and in his books, he gives advice to young men about how to live a more fulfilling life[10] and argues that political correctness is threatening freedom of speech. He currently has over 4.5 million followers on Twitter[5].

He was suspended from Twitter around the 29th of June after tweeting transphobic comments against Eliott Page, a trans actor. His reinstatement was conditioned on him deleting the tweet in question[61]. In a video posted on his YouTube channel in reaction to his suspension, he threatened the decision makers by stating ”up yours woke moralists, we’ll see who cancels whom”. He further described it as a ”badge of honor” in his battle against ”leftist ideology”[53]. The video has since garnered over 3.5 million views. Like Marjorie Taylor Greene, he was unsuspended by Elon Musk around November 21st 2022[25]. In the month after his return to Twitter, he gained over 526,000 followers (see Figure A.1d for the evolution of his Twitter following).

Dave Rubin is an American political commentator and host of *The Rubin Report*, a political talk show on YouTube. It currently has over 2 million followers[6]. On his Twitter account he has over 1.4 million followers[7]. Starting his career espousing progressive views, Rubin became conservative when he became disillusioned with ”the left” which he argued was regressive, threatened religious freedom, and against free speech. In 2018 he frequented the ”Intellectual Dark Web”, a group of media figures

[5] https://twitter.com/jordanbpeterson

[6] https://www.youtube.com/@RubinReport

[7] https://twitter.com/RubinReport

who's central tenets were "there are fundamental biological differences between men and women. Free speech is under siege. Identity politics is a toxic ideology that is tearing American society apart"[67].

Around the 5th of July, his Twitter account was suspended after he shared screenshots of the same transphobic tweets that got Jordan Peterson suspended[26]. After his suspension, he accused company executives of being "a bunch of Woke activists" and asked that Elon Musk, then in the process of buying Twitter, "blow up [Twitter's] servers so humanity can move past this pervasive, twisted, self-imposed mental institution"[26]. He was reinstated only a couple days later after deleting the tweet. Unlike the others, Dave Rubin did not gain a mass amount of followers following his reinstatement. In fact, he lost around 3,000 followers that month (see Figure A.1e for the evolution of his Twitter following).

2.2 Deplatforming as Content Moderation

In their early years, social media platforms like Twitter sought simply to break the barriers to online expression and circumvent gatekept traditional media outlets[36][21]. In line with this promise of offering a platform to everyone, they imposed very little content moderation. However, in recent years they were heavily criticized and have started to moderate more heavily by broadening its conception of unacceptable content and increasing their enforcement techniques[21].

Twitter's official stance on content moderation "is to serve the public conversation" by ensuring "all people can participate in the public conversation freely and safely" by curbing, for example, violent tweets or openly affiliating with "violent or hateful entities"[8]. Users violating the content policy will now face three forms of punishment: they can see their tweets removed, their account can be put in 'read-only' mode until the user deletes the tweet, or their account can be suspended, either temporarily or permanently. During the duration of the suspension, content they posted isn't be accessible to anyone[9].

These moderation methods and policies have been criticized. Critics argue there is a lack of transparency in what counts as unacceptable behaviour[50]. Furthermore, in the U.S, conservatives frequently claim they were targeted by suspensions and their right to free speech violated. For example, when Tucker Carlson was suspended in

[8] Twitter's content moderation policy: https://help.twitter.com/en/rules-and-policies/twitter-rules
[9] https://help.twitter.com/en/rules-and-policies/notices-on-twitter

March of 2022, he argued he was silenced for ”not having the right opinion”, that ”there was nothing harmful” about the two tweets he had shared[35]. Before purchasing the platform, Elon Musk had claimed the platform was silencing conservatives and bought the platform with the promise of restoring free speech[45]. Finally, detractors also argue suspensions might not actually be useful for two reasons. First, suspended figures and their supporters might simply move to fringe platforms with little to no moderation[49][59]. For example, after his Twitter suspension, Trump moved Truth Social[15]. Second, censoring users might paradoxically draw attention to their ideas[32] and ”makes forbidden ideas attractive”[16].

Thus, given Twitter’s important in shaping public discourse (see Section 2.1), and the controversies around suspensions as a form of content moderation, it is important to know whether they achieve their intended punitive effects and goals of limiting spread of a user’s unacceptable content, signalling to others that hateful rhetoric is not welcome, and showing that posting inadmissible tweets has consequences.

2.3 Previous work

In this section I present previous work on the effects of suspensions on social media platforms. Previous studies have taken two types of approaches: cross and within-platform user migrations. Finally, I will provide background information on toxicity and its measure in online communities.

2.3.1 Suspensions and User Migration

Previous work on the effect of social media bans have focused on whether banning users or communities might trigger mass user migration both within and cross-platform. The rational is that suspensions of users or communities to reduce toxic content on the platform might fail in two ways: first it might push users to other parts of the platform, second users might move to other platforms with little to no content moderation.

In the first event, users move to other communities within the same platform, making previously non-problematic communities toxic or recreating the same community elsewhere. Chandrasekharan et al.[13] and Saleem and Ruths[47] study the effects of Reddit’s ban of subreddits on hate-speech levels of users previously belonging to the banned subreddits, and communities they joined or created. They find that user-level hate-speech reduced and toxicity levels didn’t increase in communities receiving a heavy influx of users affected by the ban. Furthermore, newly created subreddits were quickly neutralized[13][47].

In the second event, affected users might leave the platform altogether and move towards fringe platforms with little to no content moderation. In many of those, the lax approach to moderation is marketed as the platform's appealing feature. Ribeiro et al. evaluate the effect on activity levels and toxicity signals of user migrations from banned subreddits to standalone websites created in the aftermath of the ban. They find that while activity levels on the standalone websites tend to decrease, relative activity increases. Furthermore, toxicity signals indicate a stark increase in user radicalization[27]. Other studies look at the post-suspension user migrations from Reddit and Twitter to Gab, an alternative platform to Twitter. They note an increase in toxicity among migrated Reddit users, a decrease in toxicity for the majority of Twitter users but find a strong increase for a small proportion. Furthermore, users from both platforms tended to engage more[4].

In our work, we do not study user migration as we the effects of a suspension on the same platform, and restrict ourselves to Twitter.

2.3.2 Suspensions and within-platform effects

Another way to study the effects of suspension on Twitter is to study its effects on the suspended user's post-ban influence and supporter activity, within the same platform.

Similarly to our project, Jhaver et al. evaluated the claim that a consequence of deplatforming famous Twitter users is that it ”draws attention to [the suspended users] or their ideas” and thus fails to achieve its intended goal of limiting the spread of the harmful content[33]. They chose to focus on Alex Jones, Milo Yiannopoulos and Owen Benjamin, three high-profile Twitter political figures suspended for spreading offensive ideas and speech. The author's goal was to a) evaluate whether the amount of conversations about the suspended user, b) the dissemination of their ideas, along with c) the toxicity and activity level of a suspended user's supporters, is in fact subdued by the ban, or whether the opposite, intended effect occurs.

For a), they identified key words, hashtags and expressions present in Tweets used to reference the suspended users. Then they measured the number of tweets containing these expressions and tweeters over time. For b), they collected relevant n-grams of words appearing in the collected tweets as proxies for ideas spread by a user and plotted their usage in the period before and after the suspension. For c), they identified supporters by identifying users who frequently tweeted expressions collected in b) and trained a classifier to identify them from their list of collected tweets. Then, they measured their activity levels and toxicity scores in the time preceding and following the high-profile user's suspension. They found suspensions were followed by a significant decline in conversations about a suspended user, their ideas, along with their supporter's activity levels. The authors found that though most users didn't significantly reduce their toxicity levels, there was a slight decrease in overall toxicity levels[33].

Importantly, because the high-profile users were permanently banned, the authors could not identify users which had liked, quoted and retweeted their tweets. As a consequence, in order to identify supporters, they were unable to isolate users which had retweeted and liked a suspended account's tweets before their ban. Instead, they had to rely on noisier signals of support like the frequent use of keywords and expressions associated to the suspended account. Limiting themselves to these signals meant the authors could not find users who engaged with the suspended user's tweets but didn't themselves talk about their ideas in tweets. This might be the vast majority of supporters as Twitter users are primarily consumers of content. In 2019 in the US, 80% of tweets were posted by 10% of users[68]. This also has limits when studying whether suspensions reduce the amount of harmful content on a platform. These missed supporters might have started massively engaging with and promoting other harmful content creators on Twitter. The newly shared content might not contain the same n-grams related to the suspended user's ideas, but is nonetheless toxic. Finally, because the suspended users were never reinstated, work presented in this section can't tell us what happens after they make their way back on Twitter.

To the best of our knowledge, our work is the first of its kind to study the effects of user suspensions *and* their reinstatements, and which uses data from a suspended user's Twitter timeline before their suspension.

2.4 Online Toxicity and Radicalization

2.4.1 The state of toxicity online

At the center of many criticisms of online platforms is the accusation that social media platforms harbor and amplify toxicity by design. Exposure to toxic content has almost become an inherent part of the user experience with around 40% of American users experiencing some from of harassment online[2]. Toxicity is hard to define and research suggests platforms like Twitter have built its conception by adapting to external shocks, criticisms and shifting speech norms[21]. Research also suggests it comes under many forms such as incivility[9], harassment[8], trolling[14] and cyberbullying[37]. Frequent targets of toxicity are women and minority groups[31][46].

Platform designs have also been accused of pushing certain political views, amplifying extreme content[20] and radicalizing users[65]. Some research does find evidence for algorithmic amplification of extreme views[43], right-wing views on Twitter[29] and left-wing views on YouTube[30]. However, other studies have been less conclusive[57][22][39].

2.4.2 Measuring online toxicity

In order to detect toxicity at scale, numerous automatic toxicity detectors have been developed alongside novel research in NLP. Bianchi et al., developed a Transformer model to classify sentiment about immigrants in the US and the UK with six different types of toxicity[7]. Barbieri et al. also leverage a transformer model to build a binary classifier of hatespeech[6]. Furthermore, Vidgen et al. also build a transformer model to classify comments in five categories of hatespeech[66]. Finally, Google made available their Persective API, a production-ready toxic comment detector capable of scoring comments over 7 different degrees of toxicity: toxicity, severe toxicity, insult, profanity, identity attack, threat, and sexually explicit[40]. It was also used in all previous studies mentioned before[27][33][13][47].

Previous research has investigated the suitability of the Perspective API, with researchers finding its performance is comparable to human annotators on Reddit political communities[56], and outperforms many competing models while remaining light weight and scaleable. Furthermore, given that it leverages subwords instead of a fixed vocabulary, the model should be highly resilient to internet slang[40]. Research has found a correlation between uses of African American English (AAE) and higher scores[60]. However, given that we are mostly interested in uncovering fluctuations instead of raw toxicity scores, we believe the model fits our needs.

Chapter 3

Materials, Methods, and Experimental Setup

3.1 Data collection

The entire data collection portion of the project started earlier than anticipated, and lasted a little over a month, from early May to mid-June. After Elon Musk, Twitter's owner, announced he would be ending the Academic Access to the Twitter API in May, we were under huge pressure to collect what was necessary for the project. While we were not able to collect to the full extent of our ambitions, we adjusted our research questions accordingly and believe what we collected provides sufficient empirical support to evaluate our hypothesis. In this section I will go through our methodology for choosing suspended high-profile users of interest, collecting their tweets and supporter timelines.

3.1.1 Choosing high-profile users

In order to study the effects of temporary Twitter suspension on high-profile users, we first put together a list of suspended users. When looking at suspended users, our criterias were a) they had at least 500,000 followers at the time of suspension, b) their pre-suspension Tweets could be scraped, and c) had gotten unsuspended after 2020. We required a) because we were specifically interested in the effects of suspending high-profile users. Given the media attention given to high-profile suspensions and their large following, we believe it is fair to assume the effects of their suspension are more widespread, larger, and overall have different dynamics to those of smaller accounts.

We required b) because our approach described in section 3.2 hinges on being able to scrape the interaction numbers and timeline of retweeting users before the suspension. This data is only available if pre-suspension Tweet are available. We wanted c) as we didn't want differences in observed effects to be due to time sensitive shifts in user habits.

To gather a list of candidates we used a list of high-profile Twitter suspension on Wikipedia [1] along with news articles that came up mentioning users which had gotten suspended. Our original list contained all users described in Section 2.1 along with Candace Owens[41], Andrew Tate[55], h3h3productions[28] and Matt Walsh[51]. Since, we were under time pressure, we didn't keep Candace Owens and Matt Walsh as the others had a bigger following. However, given their cultural relevance, they would've been a great addition. While we were heavily interested in evaluating the effects of Andrew Tate's suspension and subsequent reinstatement, we realized only a handful of pre-suspension Tweets could be scraped and thus no meaningful pre-suspension trend could be measured. Finally, given that all the previous users were conservative voices, we had also originally collected information on h3h3productions as it was one of the only politically left-wing accounts with a wide following that had gotten temporarily suspended. However, upon closer inspection, despite their high follower count, their tweets had extremely low user engagement.

Our final list consisted of the users described in section 2.1 and further described in Table 3.1: Tucker Carlson (@tuckercarlson), Marjorie Taylor Greene (@mtgreenee), Charlie Kirk (@charliekirk11), Jordan Peterson (@jordanbpeterson), and Dave Rubin (@rubinreport).

3.1.2 Data collection

3.1.2.1 Tweet engagement information

We relied on the Twitter API to collect all our data while our Academic API key was still valid. For each user (see Table 3.1 and Section 2.1), we scraped the last 100 pre-suspension tweets and first 100 post-suspension tweets. For each tweet, we collected the number of retweeters, likers and quoters, the text, tweet ID and author ID. Furthermore, we also collected the author id, number of followers and following for each user who had retweeted one of the collected tweets. Ideally, we would have also collected information about users who liked the tweet. However, given that some of our

[1] https://en.wikipedia.org/wiki/Twitter_suspensions

Twitter username	Follower count at time of suspension[2]	Post-suspension follower count[3]	Suspension date[4] (dd/mm/yyy)	Reinstatement date[5] (dd/mm/yyyy)
@mtgreenee	465,739	548,654	02/01/2022	22/11/2022
@charliekirk11	1,700,434	1,762,537	21/03/2022	28/04/2022
@jordanbpeterson	2,836,556	3,502,497	28/06/2022	18/11/2022
@rubinreport	1,436,814	1,441,508	04/07/2022	06/07/2022
@tuckercarlson	4,993,711	5,231,105	23/03/2022	25/04/2022

Table 3.1: Table of high-profile user accounts. (2) measured at the time of the last scraped pre-suspension tweet. (3) measured two weeks after the reinstatement date. (4) date of the last pre-suspension tweet that was scraped. (5) date of the first post-suspension tweets that was scraped

chosen high-profile users' tweets reached over 50 million likes and our scraping rate was limited, we concluded we wouldn't have time.

We ran into significant trouble trying to collect tweets for Jordan B Peterson. At the time, the Twitter API was unstable and we believe it ran into internal errors when returning his Twitter activity before his suspension. After confirming he had posted tweets before his suspension using the Wayback Machine [6], we realized these tweets were still on the Twitter platform and accessible by searching their URLs directly. To overcome this challenge we used the Twitter API's historical search endpoint to return any tweet either referencing or posted by him in the weeks before their suspension. After parsing through the results, we scraped the pre-suspension tweets and their engagement information.

3.1.2.2 User timelines

For each user who had retweeted tweets collected in section 3.1.2.1, we also collected their entire timeline and account information. Their timeline contained information about any tweet they had retweeted, tweeted, or quoted. The account information had the number of followers and following.

[6] http://web.archive.org/web/20230000000000*/https://twitter.com/jordanbpeterson

3.1.2.3 Data Management

In total, we collected user engagement information and retweeter user data for 1,000 tweets (200 for each user), scraped the timelines of over 585,456 users, and collected user engagement metrics and meta-data for around 525,958,455 tweets. User engagement information and retweeter user data were stored in zipped JSON files. Originally, retweeter timelines were also stored in zipped JSON files. Later, high-profile tweets along with their corresponding retweeter timeline tweet information were stored in their own SQL databases (one for each high-profile user) to improve data access efficiency. The project's storage requirement was of 135GB for the zipped JSON files and 250 GB for the SQL databases for a total of 385 GB of storage. Everything was stored on the University's encrypted Hawksworth server.

Due to the sheer volume of data, the whole process of writing the scripts to transfer user timelines from a list of zipped JSON files to databases, transferring the zipped files to the remote server and transferring the data took about three weeks.

3.2 Experimental Setup

3.2.1 Data for RQ#1: Suspended user's tweet popularity evolution

To measure a high-profile user's tweet's popularity, we first get the number of likes, retweets and quotes for each tweet whose information we've collected. In Section 3.1.2.1 we used the last pre-suspension tweet and earliest post-suspension tweet to estimate each user's suspension date. Before having access to these dates, we used news reports and gaps in the graph of tweet posting dates to estimate suspension dates. We then considered all tweets before the dates as pre-suspension tweets and all those after as post-suspension tweets. We then graphed the like, retweets and quotes counts for each tweet to visualize the popularity of their tweets. Finally, we also calculated the average and standard deviation for engagement metrics for pre-suspension tweets, post-suspension tweets, the first twenty post-suspension tweets, and the last eighty post-suspension tweets. All these scores are shown in Table A.2.

3.2.2 Data for RQ#2: Suspended user's supporter engagement evolution

To evaluate RQ#2 we first need to clarify what we mean by a supporter. When presented a tweet, users can engage with it by either liking, retweeting, quoting or commenting on the tweet. When a user quotes or comments under a tweet, they could be signalling to their followers, the tweeter, or other users in the comment section, that they endorse or oppose the tweet. For example, they could quote a tweet and argue its content is fallacious, or they could say it's inspiring. Furthermore, they could reply to the tweet by praising or opposing its content. Alternatively, the reply and quote could also be neutral. Thus, the simple act of replying or quoting does not in itself mean a user supports the tweet or user.

We believe that liking and retweeting a tweet both signal that a user actively endorses the views expressed in the tweet. Unlike retweets, a user's likes won't appear on their followers Twitter feed. If a user is public, it is still possible to see which tweets they've liked. However, you would need to actively search for their like history. It is a 'semi-private' endorsement of the views expressed in a tweet. On the other hand, retweeted tweets will automatically appear in a followers feed. It is a 'public' endorsement of a tweet's view. We believe that retweeting and giving public support to a tweet is a strong signal of endorsement, and a much stronger one than liking which indicates only 'semi-private' support. Thus, we define a supporter as a user who has retweeted a user at least once, with ardent supporters retweeting multiple times and casual supporters retweeting less frequently.

We understand that retweeting a tweet only signals support for the tweet and does not entail that the retweeter endorses the entirety of a tweeter's views. We believe, however, that it signals support for at least a portion of their views. Notably, those expressed in the tweet. Thus, when we argue that a group of users are bigger supporters, or a higher level than another group, we are only arguing that we have evidence that they endorse a higher proportion of tweets than others, and, consequently, endorse a bigger proportion of the tweeter's views.

Now that we've clarified what we mean by a 'supporter', we present how we measured the evolution of support among a suspended user's supporters. We measured two aspects about the evolution of supporter support (i.e. support from retweeters):

- *Distribution of retweeter retweet frequency*: We took the raw counts for the number of times retweeters had retweeted a high-profile user's tweets before and after their suspension. To identify every individual retweeter we simply used their author ID. Then, we plotted the retweet frequency distribution on the same plot. Because the resulting distribution followed an exponential distribution with an extremely steep curve of high frequency retweeters and long tail of low frequency retweeters, we represented both on a logarithmic scale with base 10. Finally, we simply compared the two curves graphically. This showed whether supporters tended to retweet more, or less frequently before, or after the suspension (i.e. whether they gave more ardent support before or after the suspension).

- *Distribution of retweet frequency among supporter levels.* We took the retweet frequency of each retweeter of tweets before and after the suspension. Then, for pre-suspension and post-suspension retweeters separately, we grouped them into 5 percentiles: those in the 0-0.5% percentile, 0.5-2% percentile, 2-10% percentile, 10-30% percentile, 30-100% percentile. For example, we have one group with the 0-0.5% of pre-suspension retweeters, and another group with the 0-0.5% of post-suspension retweeters. We also categorized post-suspension supporters according to their 'seniority': those that had retweeted before the suspension (i.e. 'old supporters'), and those that had retweeted for the first time after the suspension (i.e. 'new supporters'). We then plot for each suspended user, in a single graph, side by side box-and-whisker plots of the distributions of supporters' (i.e. retweeters') retweet frequency for each percentile. Box-and-whisker plots enables us to compare the median of the distribution (locality), the minimum and maximum values (spread), along with the the first and third quantile (spread) of the retweet frequency distribution for each percentile group before and after the suspension.

Finally, we also measure characteristics of the retweeters in the top percentile groups for pre- and post-suspension tweets. Here, we measured:

- *Distribution of follower count.* We use all percentile groups both before and after the suspension along with the the old supporters and new supporters groups described previously. Then, we simply used box-and-whiskey plots to represent the distribution of follower count among retweeters before and after the suspension for each percentile group and old/new supporters groups.

- *Retweeter account metric statistics.* Finally for each percentile group before and after the suspension, we get the number of users in the group, average number of retweets, follower count, following count, number of tweets and the corresponding standard deviation.

3.2.3 Data for RQ#3: Evolution of the number of mentions from supporters

First we filtered through all tweets in the timeline of retweeting users to keep those posted between the posting date of the first and last scraped tweet of the associated account to created the dataset $\mathcal{D}^i_{\texttt{filtered}}$. Here, i stands for the i^{th} high-profile user. All tweets in a supporter j's timeline includes all tweets that they have posted, retweeted, and quoted.‘ We filter them this way to remain coherent with the other experiments and measure an effect in the same time-frame as other experiments.

For each high-profile user, we tried to capture four types of mentions: mentions of their name, their Twitter handle, their Twitter handle when appearing in a retweet, and any other twitter account heavily linked to them. For example, to find mentions of Charlie Kirk, we would be looking for the name 'Charlie Kirk', handle '@charliekirk11' in a tweet and 'RT @charliekirk11' in a retweet of his tweet. Charlie Kirk is also a founding member and face of a number of non-profit organisations like Turning Point USA, Action, Faith and Endowement. We try to find mentions of his subsidiaries by finding mentions of their organization names, and their twitter handle. To provide some flexibility in how we captured mentions, we wrote Regex patterns for each suspended user, for each mention type. These are summarised in Table A.1.

Once the tweets were filtered by date and regex patterns written, we measured:

- *Mention counts per day.* We counted for each day the number of tweets with mentions from one of the four mention types described earlier. Then, we plotted the mention count for each mention type and observed its evolution. '
- *Name and handle mention count per retweeter retweet frequency percentile.* We measured the distribution of the number mentions for users in each supporter percentile group. For each percentile group, we get the mention frequency distribution among pre-suspension supporters before the suspension, pre-suspension supporters during the suspension, post-suspension supporters during the suspension, and post-suspension supporters after the suspension. Here, we only considered mentions of a suspended user's name or Twitter handle outside retweets. Again, we use box-and-whisker plots to represent the distribution.

3.2.4 Data for RQ#4: Supporter hate-speech usage evolution evolution

We evaluated hate-speech levels among supporters before, during and after the suspension. We measured hate-speech levels using Google's Perspective API, which we described in Section 2.4. We used three types of hate-speech available with the Perspective API: toxicity, severe toxicity, identity attack and insult. The developer documentation gives the following definitions of each score[7]:

1. **toxicity**: A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.

2. **severe toxicity**: A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words.

3. **identity attack**: Negative or hateful comments targeting someone because of their identity.

The Perspective API gives a score between 0 and 1. In Figure A.2 we show the Perspective API's breakdown of what the scores mean.

We choose to score based on 'toxicity' and 'severe toxicity' as those are general measures covering a broad range of incivility and harmful comments. Furthermore, we wanted to remain coherent with previous studies[27][33][13][47] mentioned in Section 2.3.1 and 2.3.2. We also collected 'identity attack' because 4 out of the 5 suspended users we choose to study were banned for making, or sharing transphobic tweets which would have high 'identity attack' scores. Thus, while the scores won't tell us whether supporters tweets are transphobic, we would be able to speculate that an evolution in 'identity attack' scores could be attributed to an evolution in the number of transphobic tweets.

One problem was that the Perspective API model is hosted on servers and not locally on our machine. As a consequence, the model's inference times was restricted by the server's request handling time which allowed processing a maximum of 10-13 tweets per second. This limited the speed at which we could score tweets. Furthermore, our access was subject to a maximum of scoring 60 tweets per second, greatly reducing our

[7] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US

ability to bypass the slow inference time by scoring tweets in parallel. Combined, the dataset of tweets between the first and last tweet post date of each associated high-profile user, presented in section 3.2.3 $\mathcal{D}^i_{\texttt{filtered}}$, contained a total of 188,620,108 million tweets. Scoring all tweets would have lasted a minimum of 2,183 days.

Originally, to circumvent this issue, for each dataset $\mathcal{D}^i_{\texttt{filtered}}$, we simply sampled 10% of their tweets. However, it appeared that the retweet timelines of Marjorie Taylor Greene, Jordan Peterson and Rubin Report supporters had a lot more tweets during the time of her ban than before or after their ban. Consequently, we did not have a large enough sample of pre-, during-, and post-suspension tweets to establish trends for each period.

Instead, for each high-profile account and their corresponding database, $\mathcal{D}^i_{\texttt{filtered}}$, we sampled 10% of the tweets posted in each period: before, during and after high-profile user i's suspension. We name this new database of tweets $\mathcal{D}^i_{\texttt{sampled}}$, where i again refers to the i^{th} suspended high-profile user. It is important to note that among tweets in each supporter j^i's timeline we did not filter out the retweets or quotes. Thus, any effect we measure, is a trend among 'tweets appearing in the timeline of user i's supporters' and not among 'tweets posted by user i's supporters'.

My supervisor Dr. Christopher Barrie supplied me with a dataset $\mathcal{D}_{\texttt{rand}}$ of tweets from 'random' Twitter users. The Twitter API doesn't allow us to sample tweets randomly. Instead, he approximated a random generation of tweets. First, he collect $100,000$ tweets containing the generic words "who", "where", and "why" posted from users in the US between 01/01/2011 (1^{st} of January, 2011) and 15/03/2023 (15^{th} of March, 2023). He used these generic words as they are commonly used, and have low semantic content. Together, this ensures the tweets he collected were not thematically linked to any event or time frame. For each tweet, he identified the tweeting account's ID. Then, he collect a maximum of 1,000 tweets that weren't retweets from their timelines. In total, he collected 1,190,804 tweets from 1,257 different users.

In Tucker Carlson's time frame, we have 198,394 tweets from 1,789 users, in Marjorie Taylor Greene's time frame we have 430,215 tweets from 1017 users, in Charliek Kirk's time frame we have 68,160 tweets from 655 users, in Jordan Peterson's time frame we have 235,554 tweets from 916 users, and, finally, for Rubin's time frame we have 21,804 tweets from 681 users. Finally, we gave toxicity, severe toxicity, and identity attack scores for each tweet in our five time frames. We treat this as the control set of Tweets. Using only $\mathcal{D}^i_{\texttt{sample}}$ we can study fluctuations in hate-speech level before, during and after the suspension. With $\mathcal{D}_{\texttt{rand}}$, we can measure whether hate-speech levels are relatively high compared to a random set of tweets. We then measure:

- *Average daily toxicity scores.* We measure the hourly median toxicity, severe toxicity and identity attack scores along with the first (Q1) and third quartile (Q3) given by the Perspective API over all tweets in $\mathcal{D}^i_{\texttt{sample}}$ for all i, and associated time frame in $\mathcal{D}_{\texttt{rand}}$. We also fit a linear regression through the medians before the suspension, during the suspension, and after reinstatement for each time frame. To study their evolution, we plot both sets of scores in side-by-side graph and the slope coefficient of the regression lines.

- *Toxicity levels distribution per supporter percentile.* We measure the distribution of toxicity scores for the entire period before, during and after a high-profile user i's suspension. This time, we only measure scores of 'Toxicity' (i.e. not severe toxicity or identity attack). Importantly, because we only scored tweets from a random sample of tweets from supporters before, during and after a high-profile user i's suspension, we are only able to find the toxicity score for a smaller portion of users in each supporter percentile.

Chapter 4

Results

4.1 RQ#1: How does a suspended user's tweet engagement change for tweets posted before the suspension and those posted after?

We had hypothesized in Section 1.1 that a suspended user's tweet popularity would increase after their ban. We present the evolution of the evolution of the number of likes, quotes and retweets for tweets posted by a suspended user *i* in Figure A.2. Because the number of likes is orders of magnitude bigger than the number of retweets and quotes for each suspended user *i* and would've made their evolution impossible to see, we have included the number of retweets for the tweets of a suspended user *i* in a separate graph. We, we also present the normalized evolution of the number of likes, quotes and retweets so they can all be shown in one graph. Finally, we present the averages and standard deviations of each user's tweet popularity metrics in Table A.2. The table calculates the average for four different groups of tweets: those posted before the suspension, after the suspension, the first twenty, and the last eighty tweets posted after the suspension.

Overall, we find that a) for all users, the initial post-suspension tweets receive a relatively large amount of attention compared to the rest of the tweets. For Tucker Carlson, Jordan Peterson and Marjorie Taylor Greene this phenomenon is particularly prominent. Tucker Carlson's average number of likes pre-suspension was $9,771$ and his average number of likes for his first 20 post-suspension tweets was $37,963$. For his number of retweets, those numbers were $2,444$ and $5,000$. Respectively, that's a four-fould and two-fold increase. For Jordan Peterson, we saw a seventeen-fold and

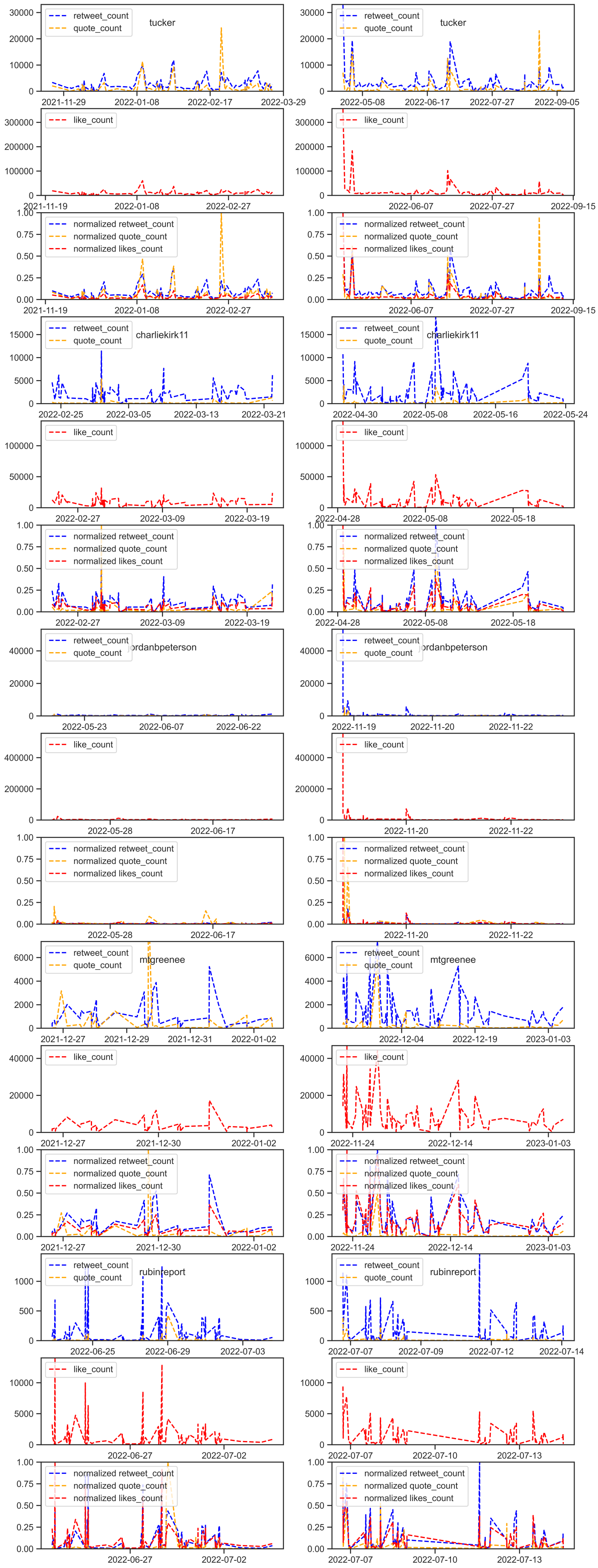


Figure 4.1: Evolution of the like, retweet, and quote counts per tweet per suspended user before their suspension (left) and after (right) their reinstatement. Every third graph also shows the normalized tweet like, retweet and quote count.

sixteen-fold increase. Finally, for Marjorie Taylor Greene, we have a four-fold and close to three-fold increase (see Table A.2).

We also find that b) while the initial popularity metrics are not sustained over time, all users but Dave Rubin (@rubinreport) saw an increase in their tweet popularity metrics. Regarding the average like count before the suspension, and after the suspension (excluding the first 20 tweets), Tucker Carlson saw a 30% increase, Charlie Kirk saw a 30% increase, Marjorie Taylor Greene saw a strong increase of 200%, and Jordan Peterson saw a 150% increase. Dave Rubin, on the other hand, saw a 30% decrease. For retweets, we saw a + 16% increase of Tucker Carlson, + 19% increase for Charlie Kirk, a 180% increase for Marjorie Taylor Greene, a 130% increase for Jordan Peterson, and a 90% decrease for Dave Rubin.

4.1.1 Critical Evaluation

Thus, we can validate **H1** as our findings suggest that while the initial post-reinstatement tweet popularity isn't sustained over time, they do see sustained moderate to significant increases in engagement in later tweets.

4.2 RQ#2: Do suspensions increase the level of support that a user receives from their pre-suspensions supporter base ?

He had hypothesized in Section 1.1 that supporters will increase their levels of support for a suspended user after their reinstatement compared to before. In Section 4.2.1 we evaluate whether supporters as a whole were more supportive after the suspension, in Section 4.2.2 we evaluate whether different degrees of supporters had shifted their levels of support, and, finally, in Section **??**, we briefly overview follower counts, and general account information for each supporter level.

4.2.1 RQ#2.1 Distribution of retweeter retweet frequency

In Figure 4.2 we show the retweet frequency among supporters of each user before and after their suspension. We use a logarithmic transform on both the y and x axis for easier visualization as the curve follows an exponential distribution with a long tail. Overall, we find a slight increase in retweet frequency among supporters in the lower

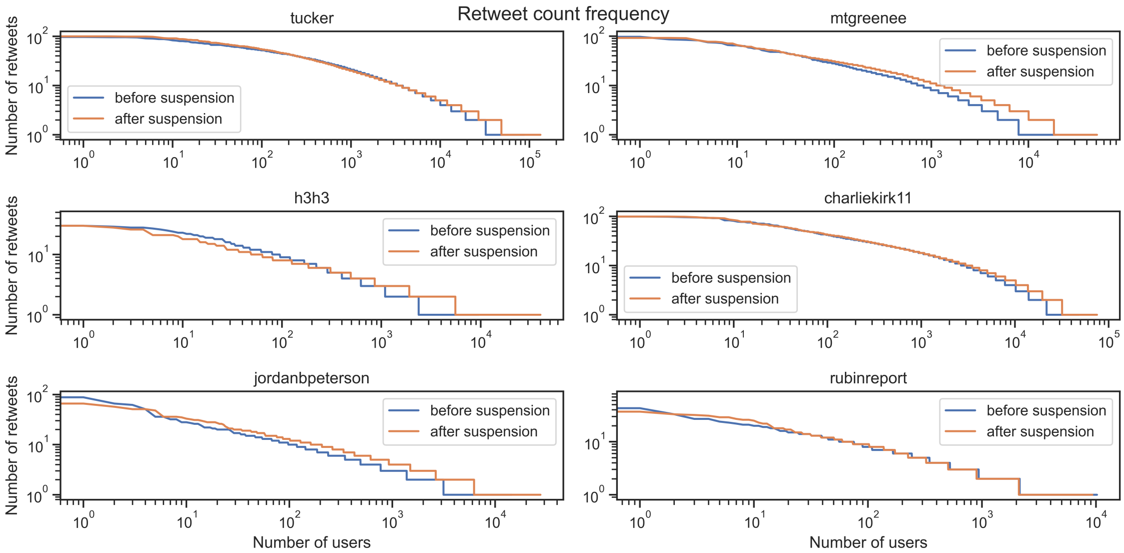


Figure 4.2: Distribution of the log number users (x-axis) for each log retweet count (y-axis).

end of the curve who don't retweet as much. However, as whole, retweet frequency among pre-suspension supporters remains largely identical to that of post-suspension supporters.

4.2.2 RQ#2.2 Distribution of retweet frequency among supporter levels

While there might not have been a general increase in support, maybe ardent or casual supporters increased their level of support. In Figure 4.3 we show supporter retweet distribution before and after the suspension per supporter percentile level and among new and old post suspension supporters.

Overall, among pre- and post-suspension supporter percentiles, we find very little difference between the retweet frequency distribution among percentile groups before and after the suspension. When there are slight differences, they are very minimal. For example, among the top 0.5% of supporters for Tucker Carlson, we find that the median retweet count was bigger by 5 tweets. Among the top 0.5-2% of post-suspension supporters, however, while the median retweet count did not change, pre-suspension supporters of that percentile had a bigger range of support. Among supporter percentiles for other users, the pre and post-suspension differences are just as small, or even smaller.

We do, however, find a slight, but consistent decrease in the median and first quartile number of retweeters among new post-suspension supporters compared to their older counterparts. It seems that new, post-suspension supporters tend to retweet less, and are thus generally less supportive of a user after their suspension than their older counterpart. This would suggest that even though there is little to no difference between the intensity of the support from the different degrees of supporters, old supporters retweet more than new supporters after the reinstatement of the suspended user.

4.2.3 RQ#2.3 Supporter characteristics

As additional information, we offer a slight characterization of each focal user's supporters.

In Figure A.3 we show the log box-plot distribution of the number of followers per supporter percentile and among new and old post-suspension supporters. For all focal users except Marjorie Taylor Greene, the distribution of the number of followers per supporter percentile group is almost identical. Importantly, this suggests that for all but Marjorie Taylor Greene, their supporters retain the same follower distribution before and after the suspension and the increase in retweet numbers doesn't seem to be attributed to mass retweeting on the part of bots.

We also note that for all, the median, first and third quartile is smaller for new post-suspension supporters than their old counterparts. This effect is the most prominent for Marjorie Taylor Greene. This difference could be due to the fact that new supporters are simply newer Twitter users and thus have not had time to collect the same amount of followers as their older counterparts.

In Table A.3 we present the number of users along with the average and standard deviation retweet follower and following count, for all percentiles and new supporter types, before and after the suspension. We notice that all suspened users have many more new supporters than old supporters in their group of post-suspension supporters. However, unlike other suspended users which tend to have between 2 and 3 times more new supporters than old supporters among their post-suspension supporters, Jordan B Peterson and Marjorie Taylor Greene have over 10 and 7 times more new supporters than old supporters respectively.

4.2.4 Critical Evaluation

We find that there was no general, supporter-wide increase in retweet frequency between before and after a focal user's suspension. Furthermore, there is also no general

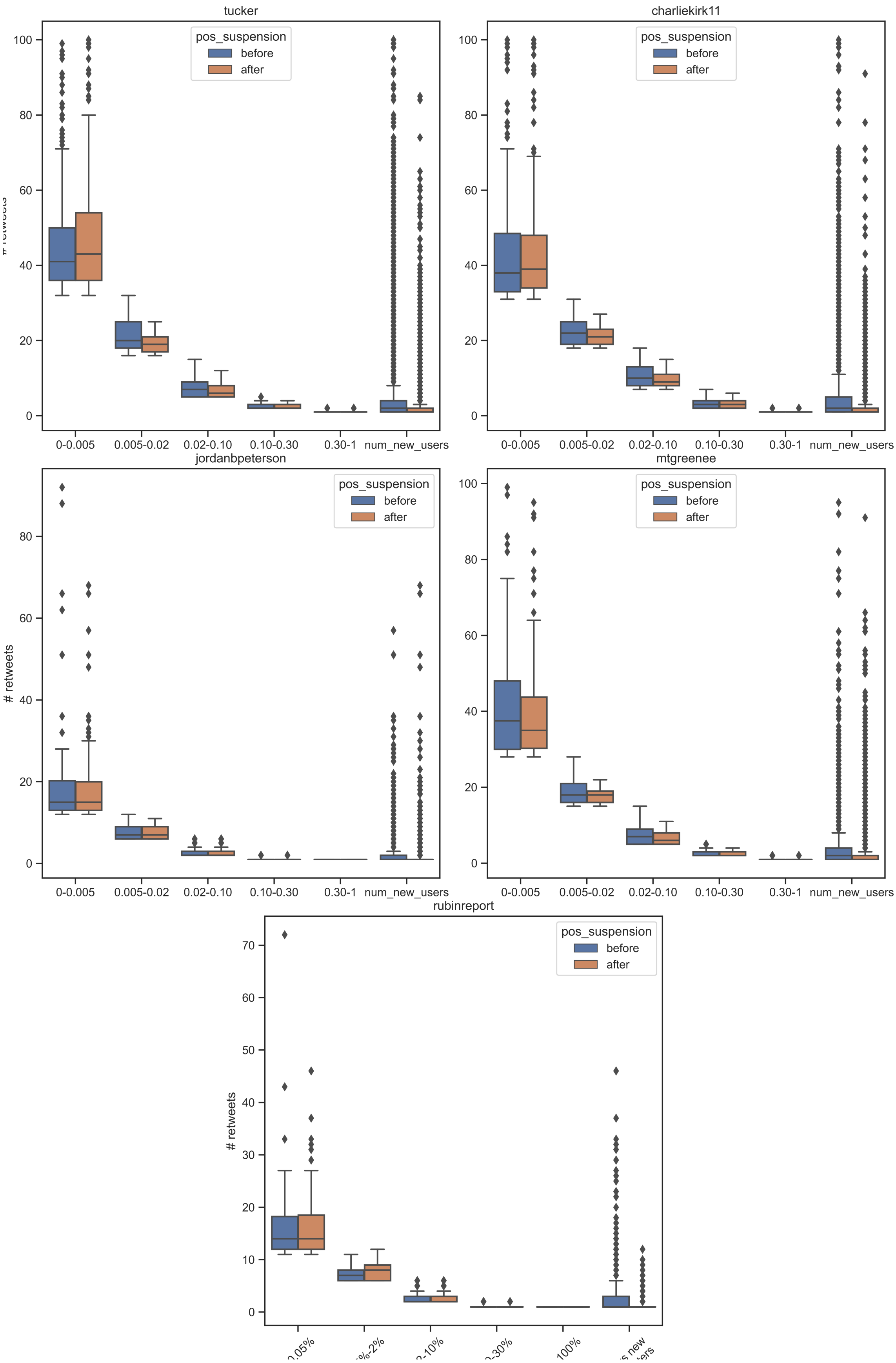


Figure 4.3: Pre and post-suspension box-plot distributions of raw retweet counts (x-axis), per supporter percentile and post-suspension supporter seniority (y-axis).

substantial difference between retweeting frequency of different supporter levels before and after a focal user's suspension. Both these conclusions suggest that after their suspension, focal users don't receive more intense support from their supporters. Instead, they appear to receive the same amount. We thus reject **H2** on the grounds that supporter intensity appears to remain the same before and after a focal user's suspension. Even tho old post-suspension supporters retweet more than their new counterpart, the difference is very small. Furthermore, as extra information, we find that all focal users have a lot more new supporters than old supporters among their post-suspension supporter base.

4.3 RQ#3: Is the amount of chatter about suspended user among supporters affected by the suspension?

We had hypothesized in Section 1.1 that 1) the overall number of mentions of a user among supporters will increase around the time of their suspension, decrease during their suspension, and increase after the suspension to higher than pre-suspenion mention levels, 2) only the most ardent supporters will continue to mention a user during their suspension.

4.3.1 RQ#3.1: Number of mentions among their supporter for each user over time.

In Figure 4.4 we graph the number of mentions of suspended users among their supporters per type of mention for each suspended user. Unfortunately, because Marjorie Taylor Greene posted a large number of tweets in a very short period of time before her suspension, and Jordan Peterson after his, we also had to plot a zoomed in version of Greene's pre- and Peterson's post-suspension mention evolution plot in Figure A.4. The counts shown in both versions of Marjorie Taylor Greene's and Jordan Peterson's are identical; Figure A.4 was created by limiting the date range shown in Figure 4.4.

For Charlie Kirk and Dave Rubin, we notice significant spikes in name and handle mentions around the time of their suspension. The number of twitter handle and name mentions reaches almost 700 and 600 mentions each. For Jordan Peterson and Marjorie Taylor Greene, while there is small increase in mentions at the time of suspension, large spikes mostly occurred after their suspension with retweet and handle mentions reaching 1,500 and 3,000 mentions respectively. For Jordan Peterson, Charlie Kirk

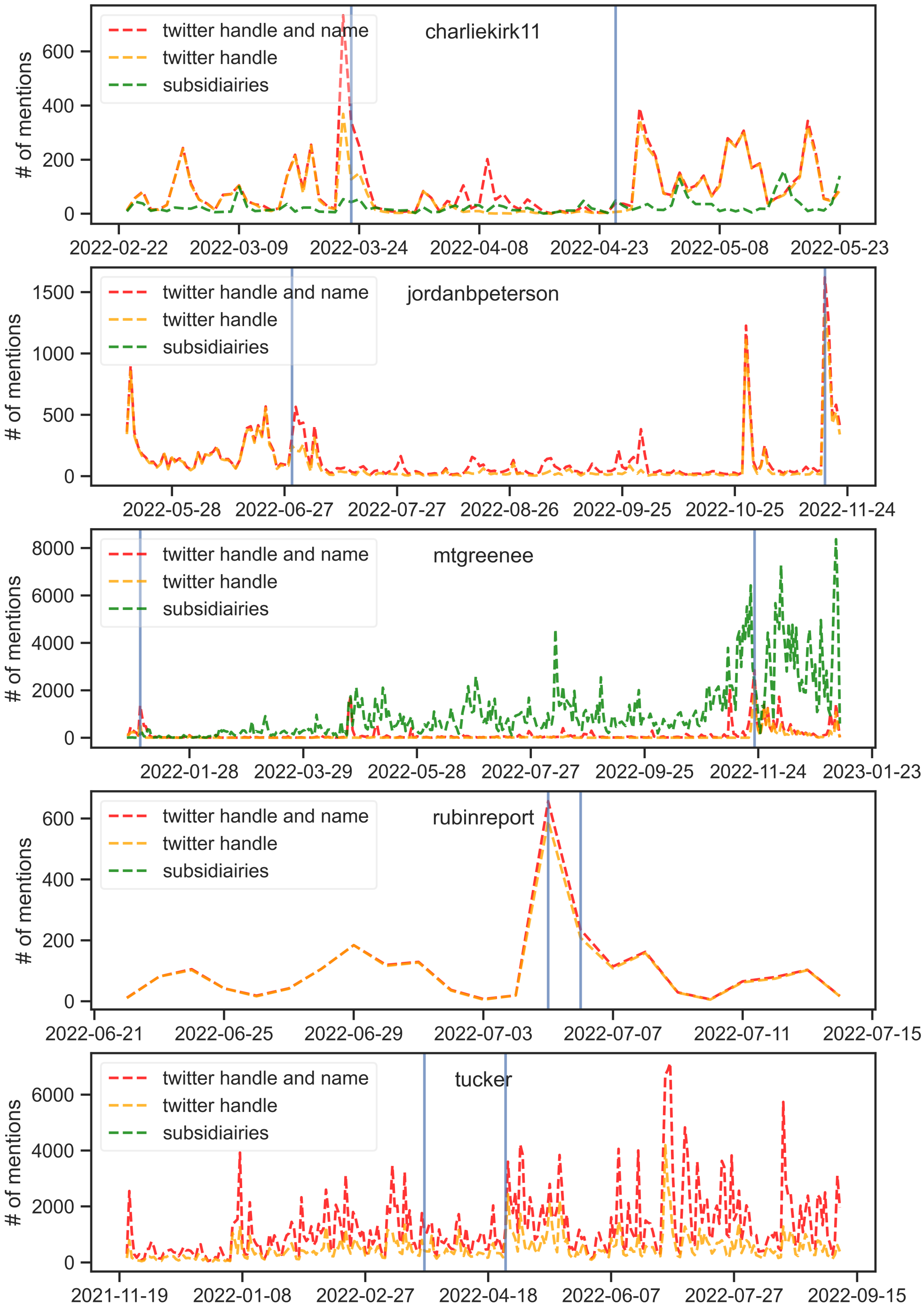


Figure 4.4: Evolution of the number of mentions per mention type among their supporters (y-axis) for each user user over time (x-axis). Vertical bars represent each suspended user's suspension and reinstatement dates.

and Marjorie Taylor Greene, name and handle mentions crash during the suspension. For Tucker Carlson, we find increases around the time of suspension, drops during the suspension and another spike after the reinstatement. However, none of the effects are very drastic.

We note a large peak in the number of mentions for Jordan Peterson around the 28^{th} of October. We believe it is because Elon Musk, having just purchased Twitter, had suggested that Jordan Peterson's account would be reinstated[44].

Interestingly, for Dave Rubin, mention counts are small around his reinstatement. We presume that unlike the other three, because Dave Rubin was only suspended for a couple of days, there wasn't much more chatter about him among his supporters after his reinstatement. This suggests the number of mentions after a suspension is heavily influenced by the amount of time a high-profile user is suspended for.

Interestingly, even though both Charlie Kirk and Marjorie Taylor Greene have subsidiaries accounts, only Greene's supporters started to mention her other account, her government representative account, whose mention count exploded at the time of her reinstatement. This suggests that her supporters changed their mentioning method by referring to her other account, because it was her's and not an organization's account.

We also notice that in most cases, supporters don't merely consistently start using a user's name instead of their handle. For Charlie Kirk, Jordan Peterson, and Marjorie Taylor Greene, once the handle could no longer be used to mention them, supporters did not turn to using their name. Interestingly, it seems this was not the case for Tucker Carlson as his supporters simply started using his name to talk about him.

4.3.2 RQ#3.2: Mentions per supporter percentile and post-suspension seniority before, during and after the suspension.

In Figure 4.5 we show the number of name or handle mentions per supporter percentile and post-suspension supporter seniority for each user to verify whether there's a link between supporter level and number of mentions of a focal user before, during, and after their suspension.

We find that during their suspension, high-profile users are mentioned slightly more by their highest supporter percentile than all lower percentiles. However, for all users but Jordan Peterson, this difference is only by a couple of tweets. Jordan Peterson's 0-0.5% post-suspension supporter percentile's Q3 retweet frequency is the highest by far compared to other users, reaching around 120 mentions while, for all other suspended

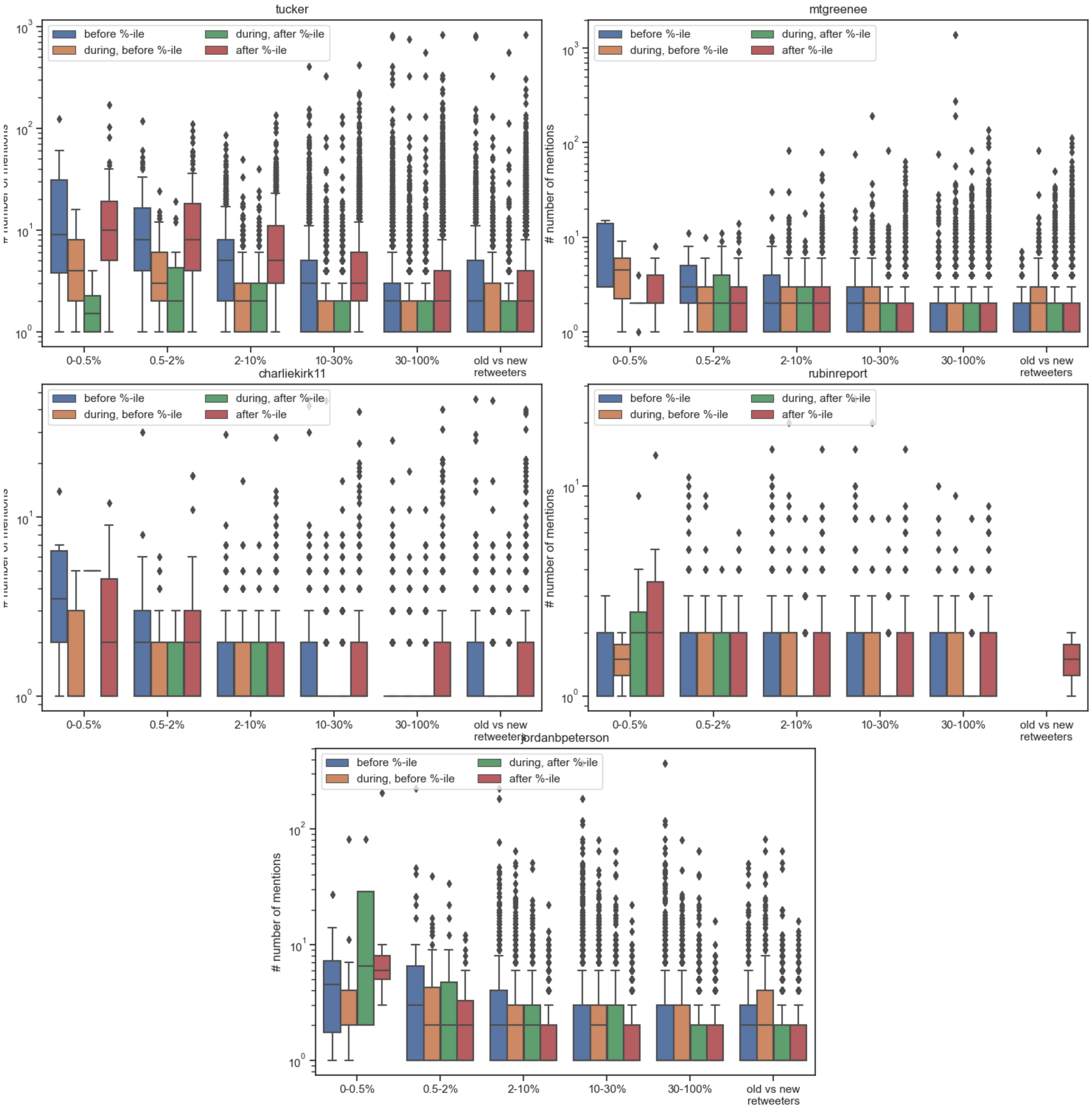


Figure 4.5: Distribution of the number of handle or name mentions per supporter percentile and post-suspension supporter seniority for each focal user.
Note that '*before %-ile*', and '*after %-ile*' refer to the supporters in the n^{th} percentile group before the user's suspension, and those after their suspension respectively. Similarly, '*during, before %-ile*' and '*during, after %-ile*' refers to the number of mentions from the n^{th} percentile group for pre- and post-suspension supporters respectively.

users, the number of mentions is under 10.

We also note that, for all but Dave Rubin, the highest percentile (i.e. top 0.5%) supporters tend to have higher mention counts than lower percentiles. However, this tends to be reserved for the highest percentile of supporters; only Tucker Carlson's higher supporter percentiles consistently mention him more.

4.3.3 Critical Evaluation

Considering results in this section, it seems we can can validate **H2.1** on the grounds that, apart for accounts reinstated only a few days after their suspension, mentions do spike at the time of suspension, drop during the suspension, and increase again around the time of reinstatement. The effect isn't as drastic for all but is there nonetheless.

Furthermore, we can also slightly validate **H2.2** on the grounds that the highest percentile supporters, generally, tend to mention suspended users more than supporters in lower percentile. The hypothesis is only partially validated however because this is only true for the highest percentile supporters (i.e. top 0.5% of supporters), and, generally, the difference is of a couple of mentions.

4.4 RQ#4: Do suspensions change the level of toxicity among supporters before, during and after the suspension?

We had hypothesized in Section 1.1 that toxicity levels would decrease after the suspension and increase after the reinstatement, and were overall higher than levels among random Twitter users. In Subsection 4.4.1 we look at the evolution over time of toxicity levels for both supporters and random Twitter users. In Subsection 4.4.2 we look at the distribution of Toxicity scores among supporter levels.

4.4.1 Evolution of toxicity levels before, during and after the suspension.

In Figure 4.6 we show the hourly median toxicity, severe toxicity, identity attack, and insult scores along with the first (Q1) and third quartile (Q3) of each score. For each user, we also fit a linear regression through the medians before the suspension, during the suspension, and after reinstatement. The regression lines slope value for the Toxicity score is displayed above the curves. The left column shows the scores for suspended users' supporters' timelines. In the right column, we show the score of tweets from a random selection of twitter users. As in Section 4.3.1, we also provide Figure A.5 with a limited time frame before and right after Green's suspension, and right before and after Peterson's reinstatement.

We find that medians scores of Toxicity among supporter tweets are mostly stable while Q3 scores fluctuate a little during the period before, during, and after each focal user's suspension. Overall, for Q3 scores among for supporters, slope coefficients were small, under 0.005. We do note that this is in part due to scores have a small range, between 0 and 1. If scores ranged between 0 and 100, Q3 regression coefficients would be under 0.5 and hover around 0.2. Furthermore, we also find that Insult scores follow Q3 Toxicity scores. Finally, Severe Toxicity scores don't fluctuate. This suggests that any effect that the suspension may have, mostly affects the prevalence of the tweets with extreme Toxicity and Identity Attack scores.

We find that even when regression slope coefficient were relatively high among supporters, they were also high among random users. During Dave Rubin's suspension, the median and Q3 Toxicity score's regression slope coefficient was around -0.001 and -0.005 respectively. For random users, they were a lot stronger at -0.007 and -0.01. Thus, we can't conclude that dips in Toxicity scores around Dave Rubin's suspension is caused by the suspension and not a particularly toxic period on overall on Twitter. We reach a similar conclusion looking at supporter and random user slope coefficients during Tucker's suspension. Finally, we make the same observation about Identity Attack scores.

Generally, we note that only Charlie Kirk's time frame seems to match our hypothesis **H4.1**; we find that among supporters, Q3 Toxicity scores increase before the suspension, decrease during the suspension, and increase again after the reinstatement. All Q3 slopes were around 0.001. Furthermore, these slope coefficients weren't matched by random users. This suggests his Twitter suspension did seem to have an affect on the Toxicity levels of his most toxic users. However, the effect is small. We make the same observation about his supporter's Identity Attack score levels.

We also find that supporter toxicity scores are, on all measures, in line with those from a random set of Twitter users. For the tweets of all supporters of our focal users and the tweets from a random set of users, on average, their Toxicity scores are around 0.1, their Severe Toxicity scores around 0, their identity attack around 0.02.

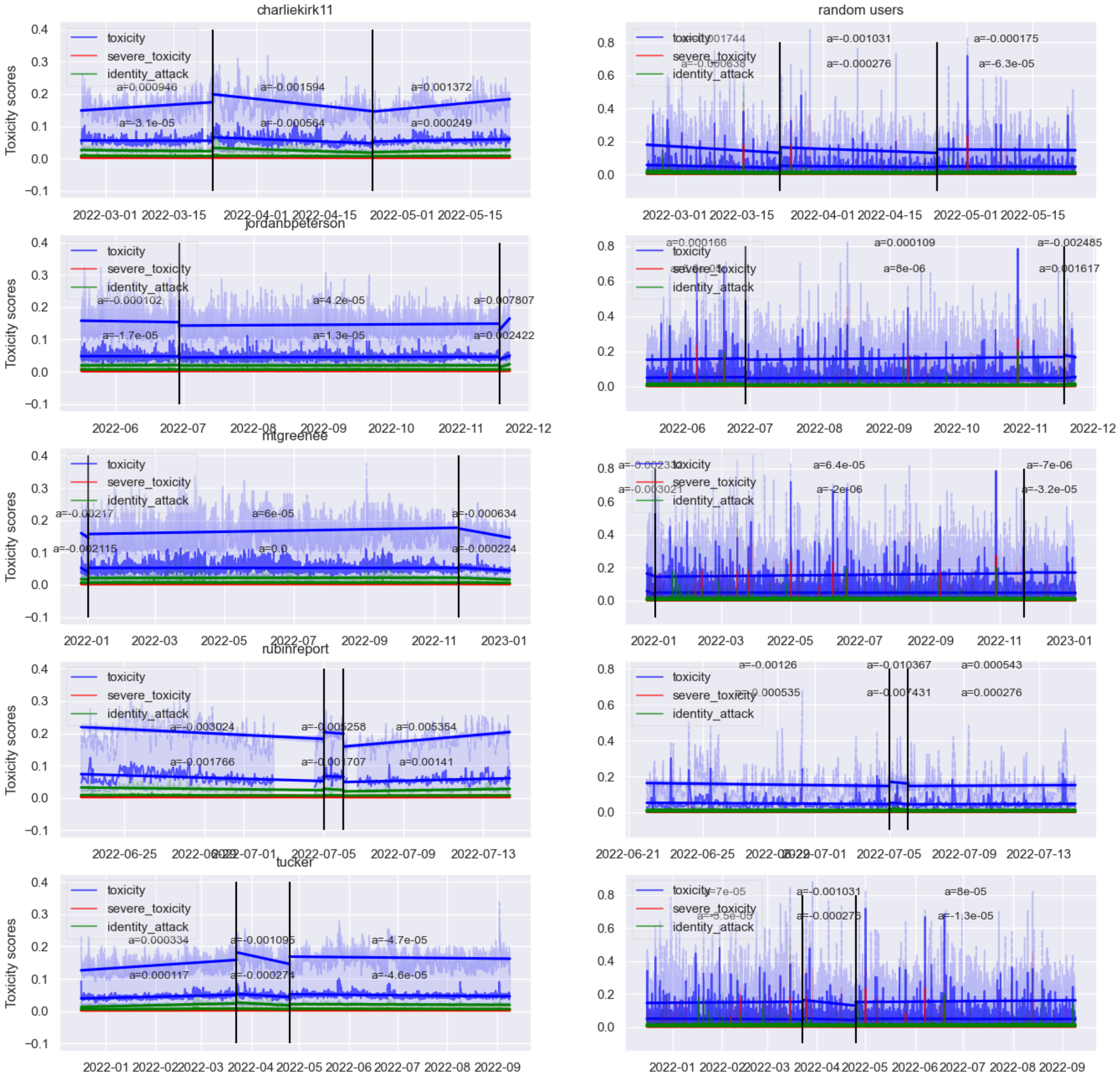


Figure 4.6: Evolution of the median toxicity, severe toxicity, identity attack, and insult scores (dark color lines along the y-axis) over time (x-axis). We also plot each score's first (Q1) and third quartile (Q3) and fill the area in between (lighter filled in colored regions along y-axis). Finally, we fit a regression line through each score for each period before, during, and after the suspension. Tweets are from a random sample of 10% of tweets from supporters' timelines (left) and random Twitter users (right). Black vertical lines represent suspended user's suspension and reinstatement date.

4.4.2 Toxicity levels per supporter percentile and post-suspension seniority.

As additional information, we verify whether toxicity scores differ between different levels of supporters for each focal user. In Figure A.6 we present the distribution of daily average scores of tweet toxicity per supporter percentile for each focal user. We only provide the distribution of daily scores in the category Toxicity (i.e. not severe toxicity, insult, or identity attack). Because sample sizes are not as big as previous scores per percentile, we provide population counts for each pre- and post-suspension supporter group. We also note that for some supporter percentiles before and after the suspension, we did not score a single tweet from their timeline. As a consequence, the distribution is not shown.

Because the sample sizes are much smaller than in other experiments, our data is much less conclusive. However, we do find that the distribution of daily average toxicity scores do not change depending on the supporter percentile. Furthermore, they also stay the same whether they are pre- or post-suspension supporters.

4.4.3 Critical Evaluation

Based on the findings presented in this section, we mostly reject **H4.1** on the grounds that toxicity levels don't seem to follow the evolution pattern we had hypothesized. Even for Charlie Kirk's supporters, the evolution was very small. Furthermore, we also reject **H4.2** on the grounds that toxicity scores for the sample of tweets of supporters of our suspended users are on par with from random Twitter users. Finally, we note that toxicity levels do not seem to be influenced by supporter levels.

Chapter 5

Discussion

In this project we try to present a first of its kind, comprehensive picture of the effects of temporary Twitter suspension on a suspended user's pre, during, and post suspension popularity, along with its effects on their supporters. Bringing our results together, it seems we can extract a common narrative.

Explosions in popularity and support around a suspended user's reinstatement seem to be caused by the suspension. From results presented in Section 4.1, we find that suspensions are heavily linked with temporary explosions in tweet engagement metrics at the time of their reinstatement for all suspended users. These engagement numbers are so far from the user's typical trends that the temporary explosion in popularity around the time of user's reinstatement seems to be caused by the news of their reinstatement. A perfect example of that are the two spikes: first, around Musk's hinting at Jordan Peterson's reinstatement, and then in his tweet announcing he was back. However, while the initial popularity is never sustained completely, our evidence does suggest that users who were suspended for long enough (at least longer Dave Rubin was suspended), do in fact become more popular. This is shown by the sustained increase in engagement metrics shown in Section 4.1. But what drives this increase in user engagement ?

Sustained post-reinstatement support is driven by an influx of new supporters. Since we were only able to collect data on retweet levels, we can only explore hypothesis about their supporter's demographic. We identify three reasons that could explain this sustained increase in popularity: 1) supporters retweet more frequently as a group, 2) different percentiles of supporters retweet more frequently, and 3) there are a lot of new supporters. Our results provide some evidence that 1) casual (low-percentile) supporters retweet slightly more. However, the difference is very small. We find that 2) different percentiles of supporters don't change their behaviour much. Our hypothesis had assumed that enthusiastic supporters (part of the top supporter percentiles) would become even more enthusiastic. That wasn't the case. Given our evidence for 2) and 3), we can't attribute the sustained increases in popularity to an increase in retweet frequency, whether by ardent supporters, casual supporters, or fans in general. What we have clear evidence for is 3) users received an influx of new supporters. In fact, those which saw the largest sustained popularity for their tweets, Jordan Peterson and Marjorie Taylor Greene, also had the highest relative numbers of new supporters among their post-suspension supporters. Thus, increases in tweet support (as measured by retweet levels), is driven by new supporters, not old supporters galvanized by the suspension.

At this stage, it would be interesting to know whether these new supporters previously followed each user, or whether they are also new followers. In the former case, this would suggest that the post-suspension popularity is driven by ”radicalized” users who were on the fence regarding their support, before the suspension. In the latter case, we would talk about a popularity increase driven by users whose contact with the suspended user came after their reinstatement. However, we can't provide a general conclusion about the effects of suspension on user radicalization, and strongly contribute to the discussion about social media and radicalization (see Section 2.4.1)

Regardless, our evidence cannot conclude there is a causal link between suspensions and sustained increase in post-suspension popularity; users could have simply become more popular through other platforms or related Twitter accounts. Marjorie Taylor Greene's supporters mention her US representative account a lot more and her government position could have contributed to her post-reinstatement increase in popularity. Tucker Carlson's show on Fox News was among the most watched shows in the United States[34]. Jordan Peterson signed a media deal with the Daily Wire at the time of his suspension[69], continued his book tour[1] and posted on his YouTube channel. Finally, Charlie Kirk also posted on his YouTube channels[2] and continued his talk show, The Charlie Kirk Show on Salem Radio Network[58].

In the event that their sustained post-reinstatement popularity is due to an increase in notoriety outside Twitter, our findings would corroborate with this explanation; those who gained the most supporters, Jordan Peterson and Marjorie Taylor Greene were suspended the longest, and thus had the longest time to garner extra support outside Twitter. On the other hand, the one suspended for the shortest amount of time, Dave

[1] https://www.jordanbpeterson.com/events/

[2] https://www.youtube.com/@RealCharlieKirk/videos

Rubin, lost supporters.

Political influencers might that influential. When measuring a suspended user's supporters' toxicity levels before, during, and after their suspension, we had hypothesized it would find a significant impact. Jhaver et al. had found suspensions were followed by a slight overall decrease after the suspension, though it only strongly affected a few users[33]. Our evidence did not find any general patterns before, during, and after a suspension. Furthermore, the effects were always very small, and, when relatively large, were reflected in the scores of random users. Thus, according to our findings, it doesn't seem that our focal user's suspensions had any general influence on their supporter's toxicity levels.

We do, however, note that our definition of a supporter did not match previous work[33], and that both our identification procedures were different. Consequently, we might have tracked different trends. This might also explain why toxicity levels among our focal user's supporters were much smaller than levels in previous studies[33]. Furthermore, unlike Jhaver et al., we don't heavily investigate changes in toxicity levels among high toxicity users[33].

We had also found that apart for Dave Rubin who wasn't suspended for long, and Tucker Carlson, focal users weren't talked about much among their supporter while away from Twitter. In fact, even in Tucker Carlson's, their most ardent supporters only talked about them slightly more than casual supporters. When suspended users were mentioned, like in Marjorie Taylor Greene's case, it was only because they had another account that represented them. This can be explained in two ways: 1) supporter simply move their conversations onto other platforms (ex. YouTube), or 2) supporters just stop caring about the focal user once they're away from the platform. Regardless, we find that generally, in a focal user's absence, their most ardent supporters don't continue expressing interests in them on Twitter.

The effects of Twitter suspensions. Our original motivation for this project was to evaluate whether suspensions, paradoxically, made users more popular or excited their supporters. Drawing on our findings and discussion above, it seems that temporary Twitter suspensions don't stop users from attracting a bigger audience and, while user restatements draw lots of attention, it is not sustained in the long run. Furthermore, they don't encourage supporters to mention suspended users during their suspended. In fact, they don't express much interest in them at all. Finally, they don't galvanize supporters into posting or sharing harmful content, but neither do they dissuade them.

Chapter 6

Conclusion, Limitations & Future Work

6.1 Conclusion

We examined the effect that temporary Twitter suspensions have on high-profile political influencers which were temporarily suspended from the platform, and their supporters. We collected data on the last 100 pre-suspension and first 100 post-suspension tweets, along with the timelines of all retweeting users. We argued that 'retweets' are a good indicator of support and qualified retweeters as supporters. We find while the initial post-reinstatement tweet popularity isn't sustained over time, suspended users do see sustained moderate to significant increases in engagement in later tweets. We also find that suspensions weren't followed by a supporter-wide increase in retweet frequency after a user's reinstatement, nor a substantial difference between retweeting frequency of different supporter levels. Furthermore, we find, supporter's barely mention high-profile users during their suspension, talk about them a lot at the time of suspension and reinstatement, and that ardent supporters talk about the high-profile users only slightly more than casual supporters during their suspensions. Finally, we do not find a meaningful links between both suspensions and supporter levels, and supporter toxicity levels, before and after the suspension.

6.2 Limitations

In our work, we were unable to collect the follower list for suspended users before and after their suspension. This would have enabled us to determine whether the influx of new supporters we find for most users were previous followers, or only discover suspended user's content during or after their suspension. Furthermore, we could not collect the list of likers, which would have enabled us to track even more casual supporters than retweeters. Another limitation is we did not use causal inference strategies, used in other studies[33][27]. They would have helped us measure causal effects of suspensions. Instead, we are only able to notice correlations and interesting trends. For example, we were not able to separate Twitter-wide toxicity score increases from changes in supporter scores, or statistically significant tweet engagement metrics in a period. Finally, our sample size is limited to five users.

6.3 Future work

We would suggest extending our analysis to include causal inference strategies to determine whether suspensions had a causal effect on the various variables we tracked (i.e. engagement metrics, toxicity scores, mention counts etc...). Furthermore, we believe a lot of work could be done to delve deeper into which supporter characteristics influence the number of mentions of a user, retweet levels, and toxicity scores. While we focused on retweet frequency, we believe a user's Twitter activity, seniority on the platform, communities in which they interacted with, all could have influenced the variables we tracked. Finally, an in-depth analysis of a restricted number users with out of the ordinary mention counts, toxicity scores, or retweet values is warranted to understand whether suspensions affect users with extreme characteristics- differently.

Bibliography

[1] *BBC News*, Jan 2022.

[2] https://www.adl.org, Jun 2022.

[3] Mar 2023.

[4] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Understanding the effect of deplatforming on social networks. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci '21, page 187–195, New York, NY, USA, 2021. Association for Computing Machinery.

[5] Justin Baragona. Twitter suspends charlie kirk for repeatedly misgendering rachel levine, Mar 2022.

[6] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics.

[7] Federico Bianchi, Stefanie HIlls, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8093–8099, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[8] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW), dec 2017.

[9] Porismita Borah. Interaction of incivility and news frames in the political blogosphere. *Advances in human and social aspects of technology book series*, page 407–424, Jan 2014.

[10] Nellie Bowles. Jordan peterson, custodian of the patriarchy. *The New York Times*, May 2018.

[11] Kevin Breuninger. Twitter says rep. marjorie taylor greene suspended “in error”; dems push to expel her from congress, Mar 2021.

[12] Matthew Brown, David Jackson, and Rebecca Morin. Live politics updates: Twitter temporarily suspends account of rep. marjorie taylor greene, Jan 2021.

[13] Eshwar Chandrasekharan, Adam Glynn, Jacob Eisenstein, Eric Gilbert, Umashanthi Pavalanathan, and Anirudh Srinivasan. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact*, 1(CSCW):31, 2017.

[14] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. *Antisocial Behavior in Online Discussion Communities*. May 2016.

[15] James Clayton and Sam Cabral. Truth social: Banned from twitter, trump returns with a new platform. *BBC News*, Feb 2022.

[16] Nathan Cofnas. Deplatforming won’t work - quillette, Jul 2019.

[17] Nicholas Confessore. How tucker carlson stoked white fear to conquer cable. *The New York Times*, Apr 2022.

[18] Andrew Court. Tucker carlson slams big tech companies after twitter flags his post, Jun 2020.

[19] Valentine Crosset, Samuel Tanner, and Aurélie Campana. Researching far right groups on twitter: Methodological challenges 2.0. *New Media Society*, 21(4):939–961, Dec 2018.

[20] Jacob Davey and Julia Ebner. We analyzed how dangerous far right ideas spread online, Jul 2019.

[21] Emillie de Keulenaar, João C Magalhães, and Bharath Ganesh. Modulating moderation: a history of objectionability in twitter moderation practices. *Journal of Communication*, 73(3):273–287, Jun 2023.

[22] Mika Desblancs and Joseph Vybihal. *Analysis of the Impact of Algorithms on Siloing Users: Special Focus on YouTube*. Taylor Francis Group, Oct 2022.

[23] Nina Dragicevic. Cbc docs pov, 2018.

[24] Catie Edmondson. Marjorie taylor greene's controversies are piling up. republicans are quiet. *The New York Times*, Jan 2021.

[25] Robert Hart. Elon musk is restoring banned twitter accounts—here's why the most controversial users were removed and who's already back, Nov 2022.

[26] Gabriel Hays. Dave rubin suspended from twitter for tweeting about jordan peterson's twitter suspension, asks musk for help, Jul 2022.

[27] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise content moderation? evidence from r/the$_{d}onaldandr/incels.ProceedingsoftheACMonHuman-ComputerInteraction,5(CSCW2)$ 1–24,Oct2021.

[28] https://www.facebook.com/thesunshowbiz. Who is ethan klein?, Oct 2022.

[29] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

[30] Hazem Ibrahim, Nouar AlDahoul, Sangjin Lee, Talal Rahwan, and Yasir Zaki. Youtube's recommendation algorithm is left-leaning in the united states. *PNAS Nexus*, 2(8), Aug 2023.

[31] Amnesty International. Why twitter is a toxic place for women, Mar 2018.

[32] Sue Curry Jansen and Brian Martin. The streisand effect and censorship backfire. *Faculty of Law, Humanities and the Arts - Papers (Archive)*, 9:656–671, Jan 2015.

[33] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021.

[34] Mark Joyella. Tucker carlson has most-watched show in cable news as fox leads basic cable for 17 straight weeks, Jun 2021.

[35] Gustaf Kilander. Tucker carlson mocked for live rant after twitter removes comment, Mar 2022.

[36] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6):1598–1670, 2018.

[37] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3739–3748, New York, NY, USA, 2015. Association for Computing Machinery.

[38] Katie Langin. Fake news spreads faster than true news on twitter—thanks to people, not bots, Mar 2018.

[39] Mark Ledwich and Anna Zaitsev. Algorithmic extremism: Examining youtube's rabbit hole of radicalization, 2019.

[40] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3197–3207, New York, NY, USA, 2022. Association for Computing Machinery.

[41] Clare Malone. The gospel of candace owens, Apr 2023.

[42] Carly Mayberry. Conservative charlie kirk slams twitter over suspension, Mar 2022.

[43] Jesse McCrosky and Brandi Geurkink. *YouTube Regrets A crowdsourced investigation into YouTube's recommendation algorithm*. Jul 2021.

[44] Dan Milmo. Twitter could split into strands allowing users to stage rows, elon musk says, Oct 2022.

[45] Dan Milmo and Johana Bhuiyan. The price of free speech: why elon musk's $44bnvisionfortwittercouldfallapart,Apr2022$.

[46] Karsten Muller and Carlo Schwarz. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3):270–312, July 2023.

[47] Edward Newell, David Jurgens, Haji Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on

reddit during a period of community unrest. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):279–288, Aug 2021.

[48] Twitter News. How many people come to twitter for news? as it turns out, a lot, Sep 2022.

[49] Jack Nicas and Davey Alba. How parler, a chosen app of trump fans, became a test of free speech. *The New York Times*, Jan 2021.

[50] Abby Ohlheiser. Here's what it takes to get banned from twitter. *The Hamilton Spectator*, Jul 2016.

[51] Indigo Olivier. Transphobe of the year: Matt walsh, Dec 2022.

[52] Donie O'Sullivan and Paul LeBlanc. Twitter temporarily suspends rep. marjorie taylor greene for vaccine misinformation — cnn politics, Jul 2021.

[53] Jordan Peterson. Twitter ban, Jul 2022.

[54] Kieran Press-Reynolds. A wave of banned far-right influencers and extremists tried to rejoin twitter after musk announced his buyout, Apr 2022.

[55] Antoinette Radford. Who is andrew tate? the self-proclaimed misogynist influencer. *BBC News*, Dec 2022.

[56] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):557–568, May 2020.

[57] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 131–141, New York, NY, USA, 2020. Association for Computing Machinery.

[58] Nick Robins-Early. The christian radio network working to reelect trump, Nov 2020.

[59] Kevin Roose. On gab, an extremist-friendly site, pittsburgh shooting suspect aired his hatred in full. *The New York Times*, Oct 2018.

[60] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[61] Vanessa Serna. Jordan peterson is suspended from twitter after elliot page comment, Jun 2022.

[62] Jack Shepherd. 22 essential twitter statistics you need to know in 2022, Feb 2022.

[63] Ben Smith. Tucker carlson calls journalists “animals.” he’s also their best source. *The New York Times*, Jun 2021.

[64] Peter Stone. Money and misinformation: how turning point usa became a formidable pro-trump force, Oct 2021.

[65] Zeynep Tufekci. Youtube, the great radicalizer. *The New York Times*, Mar 2018.

[66] Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online, August 2021. Association for Computational Linguistics.

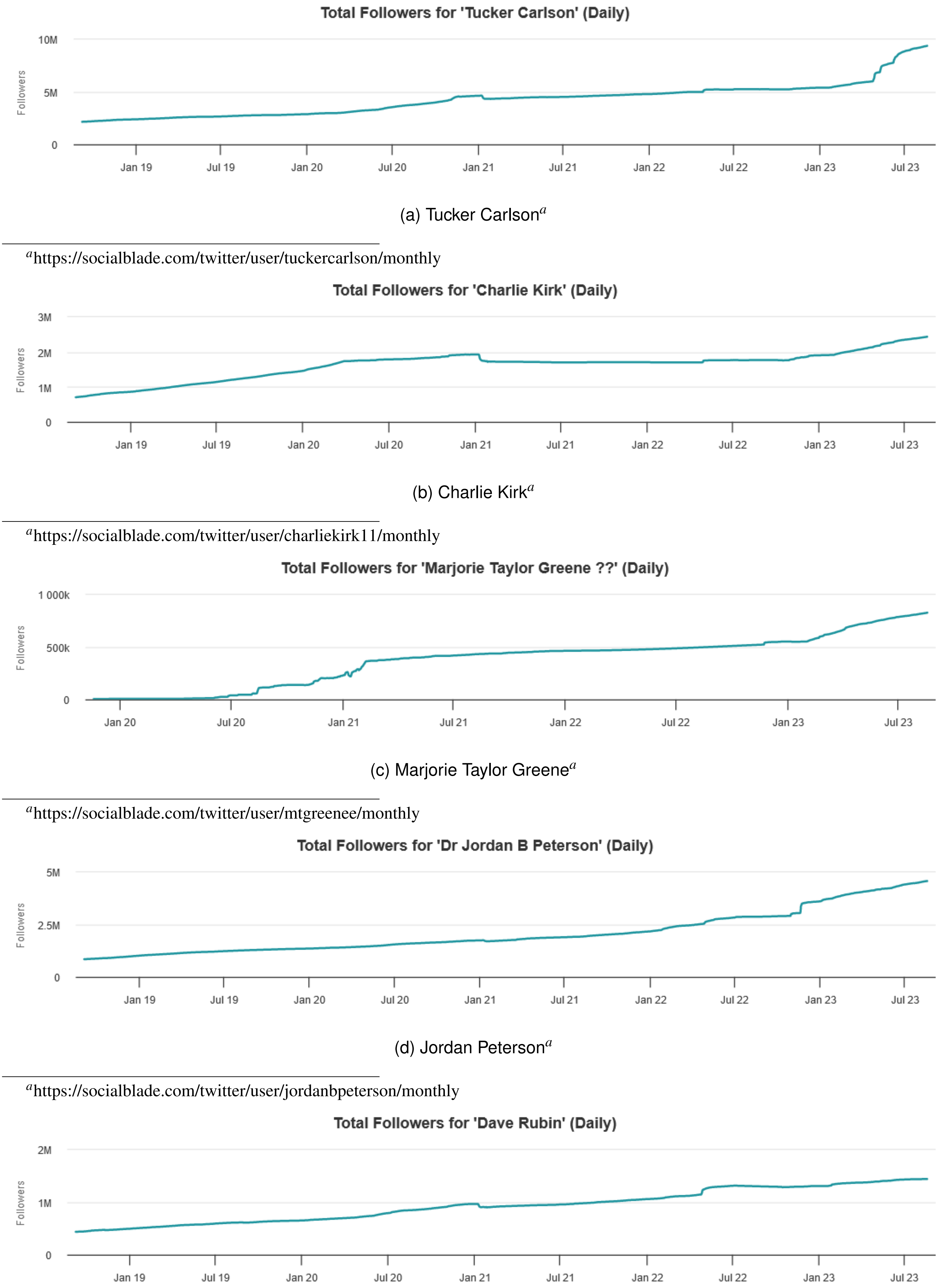
[67] Bari Weiss. Opinion — meet the renegades of the intellectual dark web. *The New York Times*, May 2018.

[68] Stefan Wojcik and Adam Hughes. Sizing up twitter users, Apr 2019.

[69] Ariel Zilber. Daily wire signs jordan peterson to podcast deal, Jun 2022.

Appendix A

Extra Tables, Graphs and Figures



(a) Tucker Carlson[a]

[a] https://socialblade.com/twitter/user/tuckercarlson/monthly

(b) Charlie Kirk[a]

[a] https://socialblade.com/twitter/user/charliekirk11/monthly

(c) Marjorie Taylor Greene[a]

[a] https://socialblade.com/twitter/user/mtgreenee/monthly

(d) Jordan Peterson[a]

[a] https://socialblade.com/twitter/user/jordanbpeterson/monthly

(e) Dave Rubin[a]

[a] https://socialblade.com/twitter/user/rubinreport/monthly

Figure A.1: Daily follower count for each suspended user studied in this dissertation.

	Marjorie Taylor Greene	Charlie Kirk	Dave Rubin	Jordan Peterson	Tucker Carlson
Name (example)	mar?jor(ie—y)(\s[a-z]*)?(\s*(taylor—greene*))	charlie?\s*kirk— charlie\sk— c kirk	(dav(e—id)\s*r(u—oo)bin)— (\brubin\s*report\b)	jordan\s*(b(ernt)?)?\s*peterson— \bjbp\b— jbpeterson— jordanbp— jordanbpeterson	(tucker(\s[a-z]*)?\s*carle*son)— (\btucker(?!\s*carlson)\b)
Twitter handle (example)	\@?mtgreenee	\@?charliekirk11	\@rubinreport	\@jordanbpeterson	\@tuckercarlson
Retweet (example)	rt\s*\@mtgreenee	rt\s*\@charliekirk11	rt\s*\@rubinreport	rt\s*\@jordanbpeterson	rt\s*\@tuckercarlson
Subsidiairies	\@?repmtg[1]	\@?tp((usa)—(action_?)— (usafaith))— turning\s*point\s*(usa—action—endowement)[2]	None	None	None

Table A.1: High-profile tweet mentions regex patterns with examples

	Suspended Accounts														
	@tuckercarlson			@jordanbpeterson			@mtgreenee			@charliekirk11			@rubinreport		
	Like	RT	Quote	Like	RT	Quote	Like	RT	Quote	Like	RT	Quote	Like	RT	Quote
Before	9771.6	2444.4	1359.7	2307.2	245.9	63.4	2205.9	543.8	264.9	7001.7	1620.9	159	1481	165.4	19
After	17235.6	3276.9	1203.4	12519.8	1158.9	115.8	6902.7	1164.2	185.5	10460.7	2143.1	205.3	1178.2	153.8	19.15
After, only first 20	37963	5000	1481.5	40785	3976	477.5	9610.6	1393.9	205.7	18975.3	2741.3	391.8	1903.6	238.1	36.9
After, w/ first 20	12117.7	2851.4	1134.7	5272.3	436.5	23.1	6250.3	1108.8	180.6	8384.1	1997.2	159.9	996.9	132.7	14.7

Table A.2: Average number of likes, retweets (RT), and quotes of suspended user tweets before their suspension, after their suspension, for the first twenty post-suspension tweets, and for post-suspension tweets excluding the first twenty.

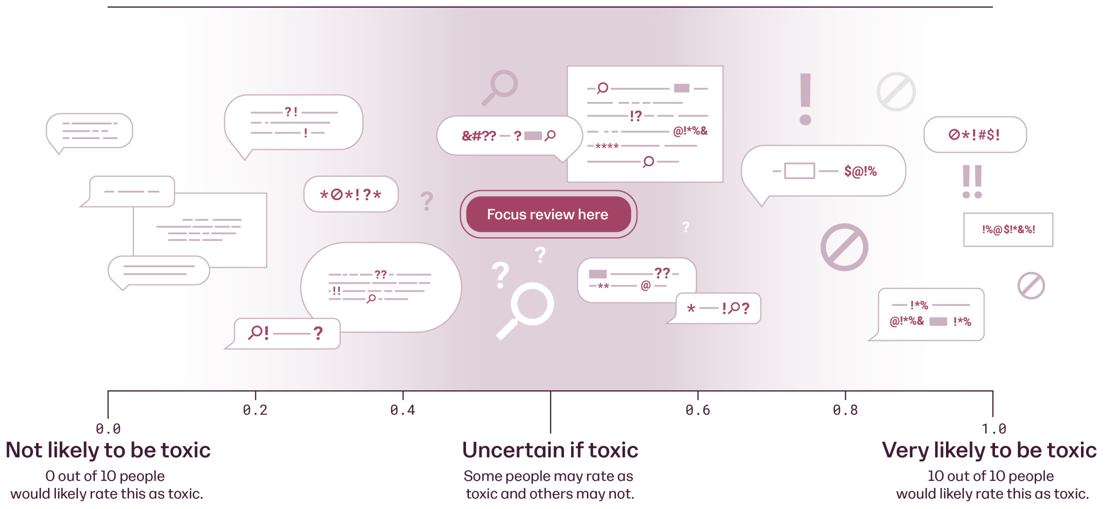

Focus review here
0.0
0.2
0.4
0.6
0.8
1.0
Not likely to be toxic
0 out of 10 people would likely rate this as toxic.
Uncertain if toxic
Some people may rate as toxic and others may not.
Very likely to be toxic
10 out of 10 people would likely rate this as toxic.

Figure A.2: Breakdown of Perspective API score by severity from the Perspective API documentation page[3].

[3] https://developers.perspectiveapi.com/s/about-the-api-score?language=en_US)

		# users		# RT		# Followers		# Following	
Account	Group	Before	After	Before	After	Before	After	Before	After
@tuckercarlson	0-0.5 %	453	414	45	47	3152	3905	2462	2671
	0.5-2.0 %	1330	970	21.5	19.3	3391.2	3540.6	2443.5	2703.3
	2-10 %	8277	9602	7.8	6.9	2818.8	3082.5	2324.4	2602
	10-30 %	24508	36588	2.8	2.6	3013.4	2551.4	1904.1	1976.8
	30-100 %	68351	106158	1.2	1.2	2159.6	2158.6	1444.9	1458.4
	new supporters %	34610	98408	4.8	1.6	3024.4	2066.7	2203.9	1472.4
@mtgreenee	0-0.5 %	104	138	42.2	40.4	3124.6	1628.1	3203.9	1598.2
	0.5-2.0 %	326	403	19	17.8	4991.8	1724	4274.9	1902.4
	2-10 %	2074	3507	7.7	6.7	4090.1	1791.5	3319.1	1926.6
	10-30 %	6030	13838	2.8	2.5	3383.3	2496.5	2786.9	1890.9
	30-100 %	14504	40810	1.2	1.2	3035.4	1825.9	2381.3	1578.7
	new supporters %	6384	44482	3.5	2.0	4613.3	1591.2	2947.5	1462.8
@charliekirk11	0-0.5 %	259	272	43.9	44.2	1820.2	2810.5	2072.7	2868
	0.5-2.0 %	822	726	22.6	21.4	2552.5	3608.5	2323.7	3105
	2-10 %	4199	4824	10.5	9.6	2534.5	3692	2797.8	2746.4
	10-30 %	17441	25605	3.3	3.0	2315.5	2331.5	2378.5	2077.1
	30-100 %	34246	55438	1.2	1.2	2386	2007.3	1869.1	1599.7
	new supporters %	23954	51074	4.6	1.8	3141.3	1724.2	2551.7	1470.2
@jordanbeterson	0-0.5 %	80	129	20.6	19.2	816.4	849.4	1444.3	1336.5
	0.5-2.0 %	280	488	7.6	7.5	1301.5	1077	1520.5	1269.4
	2-10 %	2907	5805	2.7	2.7	3435.9	1253.5	1214.6	1222.3
	10-30 %	13127	24867	1.1	1.1	2020.7	1568.6	1087.7	1128.3
	30-100 %	11363	21278	1	1	1589.7	1656.9	1081.2	1121.3
	new supporters %	2386	25127	2.5	1.4	1397.7	1577.2	1529.7	1108.5
@rubinreport	0-0.5 %	56	60	17	17.2	1422.4	1239.8	1652.5	1376.8
	0.5-2.0 %	198	186	7.4	8	2066.2	1863.3	2119	1498.2
	2-10 %	1975	1932	2.7	2.7	6460	6066.8	1828.4	1824.5
	10-30 %	9259	8387	1.1	1.1	3070.4	4244.5	1598	1685.6
	30-100 %	8054	7178	1	1	2588.3	4395.9	1578.3	1662.2
	new supporters %	2425	6858	2.5	1.2	5478.4	4422.2	1894.2	1619.7

Table A.3: User metrics information for supporters of suspended users per percentile and post-suspension seniority.

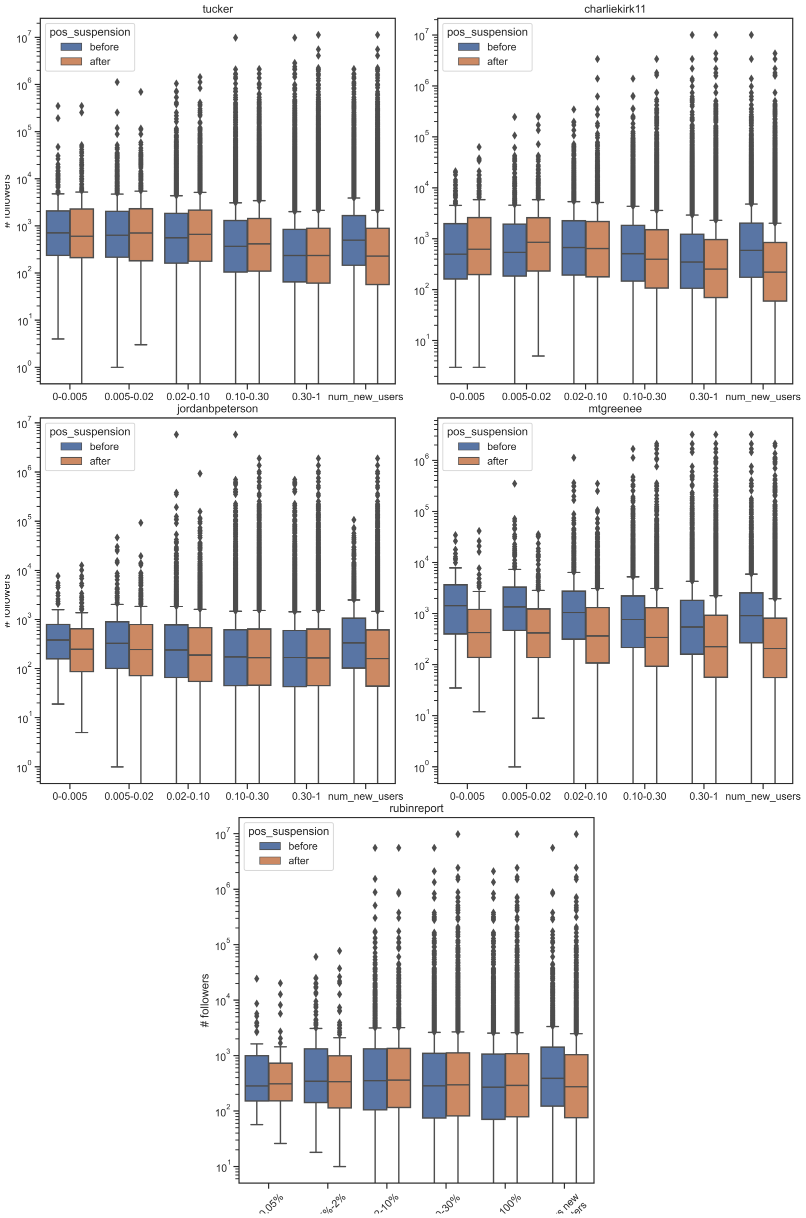


Figure A.3: Pre and post-suspension box-plot distributions of follower counts (x-axis), per supporter percentile and post-suspension supporter seniority (y-axis).

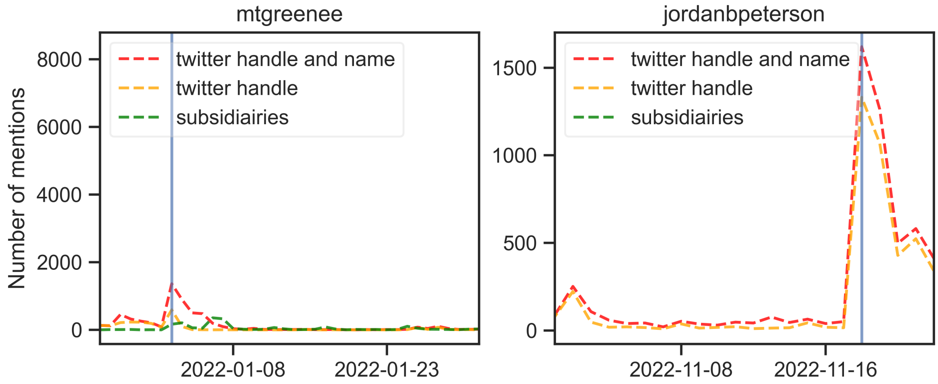


Figure A.4: Evolution of the number of mentions per mention type among their supporter (y-axis) for each user user over time (x-axis) for Marjorie Taylor Greene (right) and Jordan Peterson with their x-axis cut-off to include only the time around the suspension of Marjorie Taylor Greene and reinstatement of Jordan Peterson.

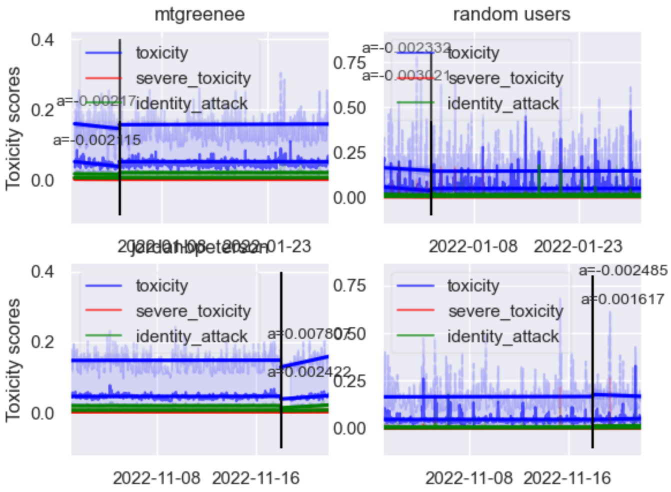


Figure A.5: Evolution of the median toxicity, severe toxicity, identity attack, and insult scores (dark color lines along the y-axis) over time (x-axis). We also plot each score's first (Q1) and third quartile (Q3) and fill the area in between (lighter filled in colored regions along y-axis). Finally, we fit a regression line through each score for each period. We only show scores for Jordan Peterson supporter timeline tweets (left) right before and after his suspension, and scores for Marjorie Taylor Greene supporters (left) before and right after her suspension.

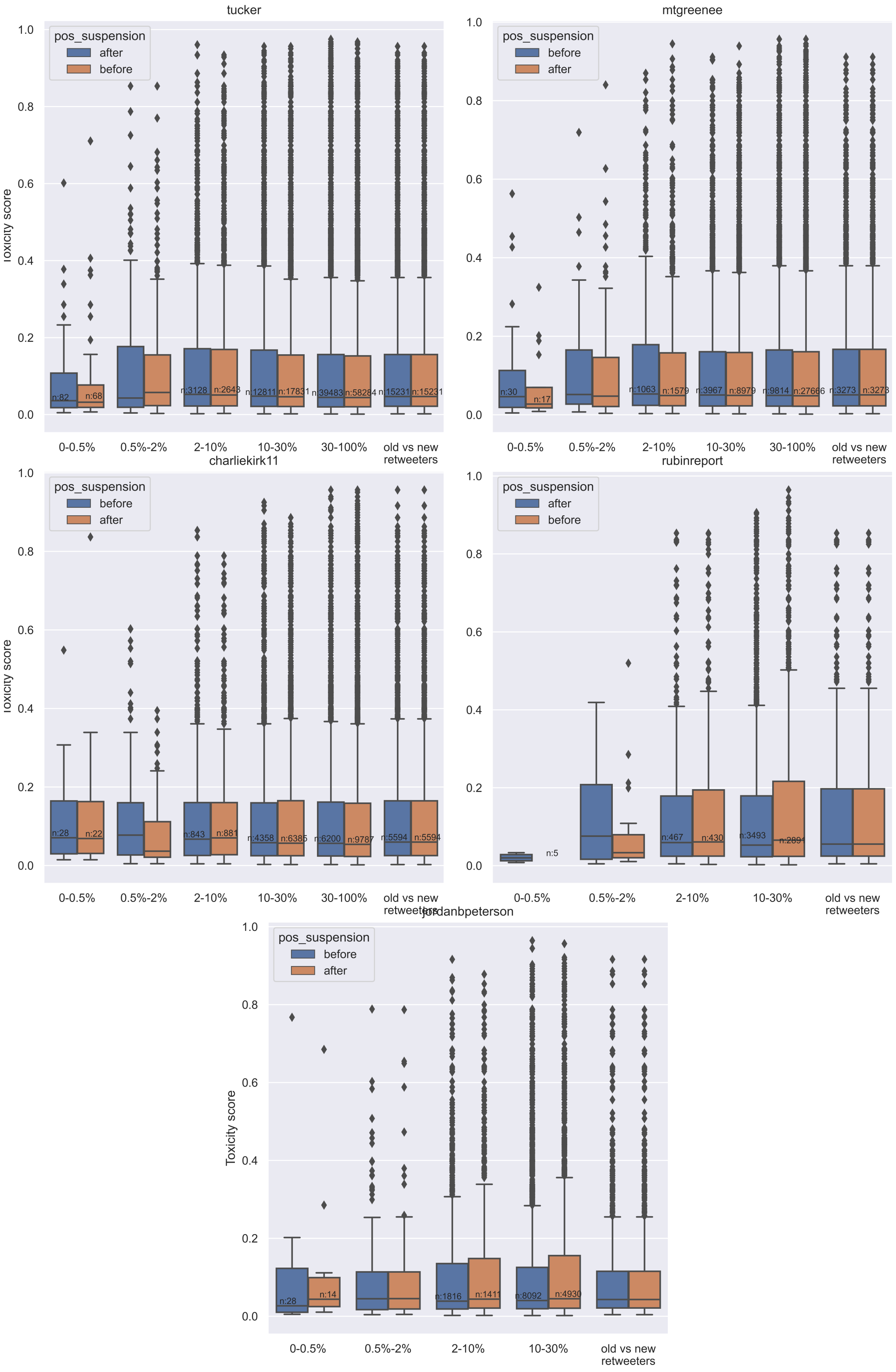


Figure A.6: Evolution of the average Toxicity score for each percentile and supporter seniority for each suspended user before and after their suspension. For groups for which we could calculate it, we also include the number of users in the sample.