Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

The Effects of Group Sanctions on Participation and Toxicity: Quasi-experimental Evidence from the Fediverse

Carl Colglazier[1], **Nathan TeBlunthuis**[2], **Aaron Shaw**[1]
[1]Northwestern University
[2]University of Michigan
carlcolglazier@u.northwestern.edu

Abstract

Online communities often overlap and coexist, despite incongruent norms and approaches to content moderation. When communities diverge, decentralized and federated communities may pursue group-level sanctions, including defederation (disconnection) to block communication between members of specific communities. We investigate the effects of defederation in the context of the Fediverse, a set of decentralized, interconnected social networks with independent governance. Mastodon and Pleroma, the most popular software powering the Fediverse, allow administrators on one server to defederate from another. We use a difference-in-differences approach and matched controls to estimate the effects of defederation events on participation and message toxicity among affected members of the blocked and blocking servers. We find that defederation causes a drop in activity for accounts on the blocked servers, but not on the blocking servers. Also, we find no evidence of an effect of defederation on message toxicity.

Introduction

Content moderation in response to toxic and anti-social behavior is pervasive in social media. In general, moderation interventions strive to balance the value of wide and active user bases with the threats posed by conflicts and hate speech (Gillespie 2018). Websites that host user-generated content and sub-communities apply many kinds of policies and interventions. However, when norms diverge across interconnected, independent communities in the absence of a single (corporate or not) parent or owner, governance and moderation pose acute challenges.

A growing empirical literature has investigated social media content moderation and governance. Moderation actions most frequently target individual posts and accounts, but other group-level sanctions affect entire communities or websites. For example, Reddit has banned subreddits and Discord has blocked servers, reducing the prevalence of unwanted behavior within and sometimes beyond the targeted groups (Chandrasekharan et al. 2017, 2022; Ribeiro et al. 2021, 2023; Russo et al. 2023; Zhang and Zhu 2011). Most prior work on group-level sanctions focuses on sanctions applied by central actors such as commercial social media platform staff. However, autonomous community administrators can also enact group-level sanctions such as when a sub-community restricts contributions from members of another sub-community. These decentralized group-level sanctions are distinct in that the targeted sub-community remains part of the larger network. To our knowledge, the effects of such sanctions remain unexplored empirically.

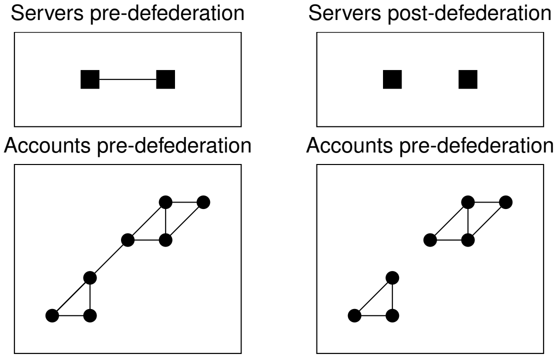

Servers pre-defederation
Servers post-defederation
Accounts pre-defederation
Accounts pre-defederation

Figure 1: Illustration of how defederation disconnects two servers and thereby disconnects the subnetworks of people using each server. The top row shows the network of servers before (left) and after (right) defederation. The bottom row shows the corresponding networks of users. On the left, an edge connects a user on one server with a user on a different server. Defederation (right) disconnects them so they can no longer exchange messages.

To investigate the effects of decentralized group-level sanctions, we analyze *defederation events* in the Fediverse, a decentralized social media system which consists of independently managed servers that host individual accounts and pass messages using shared protocols. Communication between servers can happen only when the administrators of both servers permit it. Server administrators can revoke such permission by "defederating" from (blocking all interactions with) specific servers. Defederation is one of the few tools administrators in a decentralized system have to protect against bad actors or enforce norms from beyond their own servers. While many defederation events on the Fediverse occur between servers with no known interaction history, many also come in response to norm violations and toxic interactions across server boundaries with a history of previous interactions. Defederations that cutoff cross-server interactions provide an opportunity to identify the effects of these group sanctions on accounts most likely to be directly affected.

We collect data from 214 defederation events between January 1, 2021 to August 31, 2022 that involved 275 servers and 661 accounts which had previously communicated across subsequently defederated inter-server connections. Using a combination of non-parametric and parametric methods, we estimate the effects of defederation on two outcomes: posting activity and toxic posting behavior among affected accounts. We find an asymmetric impact on posting activity: Accounts on blocked servers reduce their activity, but not accounts on blocking servers. By contrast, we find that defederation has no effects on post toxicity on either the blocked or blocking servers.

These findings suggest that defederation, although a common group-level sanction on the Fediverse, has mixed effectiveness: Despite the risks of severing communication channels, communities implementing group-level sanctions do not lose activity. This implies that defederation may avoid some of the costs associated with other moderation techniques such as account requirements or group sanctions like geographic blocks (Hill and Shaw 2021; Zhang and Zhu 2011). Although defederation reduces activity by blocked accounts, we did not find evidence that it made their posts less toxic. This suggests that defederation may not improve adherence to broadly held norms. Our study contributes to knowledge of content moderation on social media in that it (1) describes defederation, a novel form of group-level sanction as instantiated on the Fediverse, (2) derives hypotheses regarding the effects of defederation from prior literature, (3) creates a novel dataset of defederation events, (4) conducts a quasi-experimental analyses to quantify effects of defederation on parties affected on the blocked and blocking servers and (5) finds that defederation has asymmetric effects on activity and no measurable effect on toxicity.

Background

Social Media Content Moderation and Governance

Challenges of decentralized governance and content moderation in networked communication systems are ubiquitous. Threats to public safety, trust, and health resulting from toxicity, misinformation, and incivility online are now widely perceived and addressed by various content moderation and governance tools (Gruzd, Soares, and Mai 2023). National and supra-national governments impose disparate legal regimes such as the GDPR in Europe or the DCMA in the United States. Simultaneously, corporate platforms employ many strategies, from algorithmic filters to user and content monitoring systems, to monitor and shape content visibility. Leaders of communities such as subreddits and Facebook groups institute their own rules and norms and seek to hold community participants to them (Gillespie 2018). Interdependence among these components makes effective policymaking, design, and analysis of social media governance especially difficult.

Large corporate platforms and autonomous communities alike must grapple with incongruent norms about online behavior and the appropriate role of the above governance components. For instance, today's regulatory assemblage governing social media departs from earlier eras characterized by utopian notions of freedom of expression on the "electronic frontier" (Kaye 2019). Whether governance interventions are understood as "censorship" or "content moderation" depends not only on one's perspective, but also the on their origin within the institutional hierarchy (Myers West 2018). At the highest level, state-led interventions prevent access to sites, domains, services, IP blocks, or protocols (King, Pan, and Roberts 2013). Similarly, network service providers and other intermediaries also engage in blocks or website takedowns justified by terms of service violations or legal risks (Vu, Hutchings, and Anderson 2023; Ribeiro et al. 2023; Han, Kumar, and Durumeric 2022). Lower in the hierarchy, owners and administrators of specific sites or platforms set their own policies and can enforce them through sanctions applied to individuals as well as whole (sub-)communities (Jhaver et al. 2021; Chandrasekharan et al. 2022; Ribeiro et al. 2021). Devolving policy setting and enforcement to increasingly low-level actors poses a tradeoff. Low-level governance can institute local norms and advance a community's specific goals, but may lack the necessary means to prevent important categories of harm.

A growing literature in the social and computational sciences evaluates the empirical effects of specific content moderation and governance techniques at various institutional levels. Previously studied techniques include deplatforming, quarantines, censorship and geographic blocks, banning, account requirements, and pre-publication filtering (Ribeiro et al. 2023; Chandrasekharan et al. 2022; King, Pan, and Roberts 2013; Ribeiro et al. 2021; Zhang and Zhu 2011; Hill and Shaw 2021; Tran et al. 2022). In general, this prior work finds that barriers to specific kinds of contributions from specific kinds of actors cause immediate decreases in the targeted activity among the targeted accounts. As examples, requiring accounts on wikis decreased the quantity of low-quality contributions (Hill and Shaw 2021); mainland China's block of Chinese language Wikipedia decreased contributions from editors inside and outside of China (Zhang and Zhu 2011), and pre-publication moderation increased quality contributions (Tran et al. 2022).

Some governance interventions spill-over or have indirect effects (intended or not) beyond their immediate targets. Banning hateful communities on Reddit, for example, reduced speech characteristic of the banned communities by both the active participants in those communities and within the other communities they participated in following the ban (Chandrasekharan et al. 2022). However, when Reddit banned two toxic communities, they subsequently reorganized on other platforms where they experienced declining activity and one had increasing radicalization and toxicity (Ribeiro et al. 2021). Banning whole sub-communities may also create spillovers of antisocial and toxic behavior when the former members of banned communities migrate their activities elsewhere (Ribeiro et al. 2021, 2023).

The majority of prior work in this area has identified effects of governance and moderation interventions on the largest "mainstream" social media platforms such as Facebook, Twitter, or Twitch (Jhaver et al. 2021; Mitts, Pisharody, and Shapiro 2022; Seering, Kraut, and Dabbish 2017). A few have focused on more decentralized or autonomous community environments like subreddits or independently managed wikis (Chandrasekharan et al. 2022; Hill and Shaw 2021). Many of the interventions studied include group-level sanctions such as political blocks and censorship events affecting whole geographic regions (Zhang and Zhu 2011), as well as decisions and enforcement actions by "local" administrators or moderators (Srinivasan et al. 2019). Several more recent studies have considered alternative social media sites, such as alt-right communities Voat, Parler, or Gab, often with an eye towards understanding how interventions on mainstream platforms may have impacted participation in less regulated environments (Ribeiro et al. 2023; Stocking et al. 2022). The current project expands this literature by analyzing the effects of a previously unstudied group-level sanction in a decentralized and federated social media network.

Governance and Moderation in Federated Communication Networks

The architectures of federated communication systems create distinct opportunities and obstacles to governance and moderation. Federated communication networks, such as email, are defined by independently operated servers which let users pass messages both within and across servers. Protocol-based interoperability underpins this design and affords cross-server communication. The resulting server-level autonomy may also offer benefits in terms of scalability and resilience. For instance, new servers may join or leave the network without coordinating with other servers. Such decentralized and complex systems of communication tend to possess a scale-free structure with superior robustness to the failure of single nodes over other classes of large, complex graphs (Albert, Jeong, and Barabási 2000).

Scalability and resilience are particularly important qualities for communication networks, where the value of the network is often a superlinear function of its size (Van Hove 2016). Expanding a communication network can produce collective benefits to members driven by both the increased space of potential connections (*connectivity*) as well as the increased opportunities for pooled information (*communality*) (Fulk et al. 1996). Federated networks also offer profuse opportunities for decentralized action, including modular extensions and localized content and norms (Datta et al. 2010).

At the same time, the decentralized structure of federated networks also makes it difficult to centrally control or apply uniformity. Technical features typically require a sufficient level of adoption across servers to be useful. The overall result is that the user experience in a federated communication network can vary enormously depending on how one connects to it.

This variability extends to governance. No server can directly dictate the rules or decisions of another. The administrators of each server hold complete autonomous control over their internal affairs. In choosing which server to connect to the network through, individuals also choose a governance regime (intentionally or not). If unsatisfied with one regime, they can (at least in theory) move to another and take their data and social ties with them. Among dissatisfied members of a community, the opportunity for "exit" may be especially easy when it costs little time or effort to join another subnetwork (Frey and Sumner 2019; Frey and Schneider 2021). Individuals wishing to behave anti-socially—to harass, troll, or make offensive posts—may select a server that allows them to do so. Other servers within a federated social media network must address this to protect their own users.

Governance in federated communication networks is not limited to servers' authority over individuals: servers also implement sanctions against each other (Anaobi et al. 2023). If one server hosts anti-social behavior, other servers can implement a server-level intervention. Such interventions, for instance group-level bans or blocks, also happen among semi-autonomous communities within platforms or more centralized environments like Reddit, Wikipedia, and multiplayer video game servers. Yet, to our knowledge, prior work on governance in social media has not analyzed the effects of these sorts of group-level interventions. Doing so offers a valuable opportunity to expand the empirical knowledge in this domain.

The Effects of Community-Level Defederation as a Group Sanction

Servers in a federated network can enact sanctions against other servers in several ways, all of which regulate interactions among their users. For example, servers may filter, limit or slow the propagation of messages from another server through their subnetwork. Of the possible approaches, *defederation*—which completely blocks all communication across server boundaries—is the most extreme. Defederation is widely used when server administrators determine that another server is causing problems to such a degree that managing continued interactions is not worth the trouble. For example, email spam filters often include rules blocking all messages originating from a given email domain or server.

Based on prior content moderation research and the characteristics of defederation as a specific type of governance intervention in decentralized social media, we pursue the following research questions:

1. How does defederation impact the activity levels for affected accounts on (a) the *defederated* instance (blocked server); and (b) the *defederating* instance (blocking server)?
2. How does defederation impact toxic posting behavior among the affected accounts on either (a) the blocked or (b) blocking servers?

When one Fediverse server blocks another, the blocked server no long propagates messages to the blocking server. This reduces the size of the audience that can be reached from the blocked server and thereby may decrease the utility of posting on that server. As with interventions studied in prior research such as bans, quarantines, and deplatforming, we expect that defederation reduces activity among affected accounts on the blocked servers.

The likely consequences of defederation for the blocking server are less clear. Disconnecting from the blocked server decreases the potential audience of the blocking server's users. As a result, individuals who lose valuable connections might have fewer reasons to use the network. On the other hand, server administrators who initiate defederation likely have an informed rationale. They may anticipate the costs and benefits of defederation by observing messages between their server and the servers they consider blocking. Perhaps more importantly, some volunteer administrators struggle to effectively protect their users from anti-social behavior. By enacting defederation, administrators demonstrate competence in server governance, eliciting increased trust and commitment from members of their communities. If users tend to value such competence more than their connections to the blocked server then defederation is unlikely to decrease their activity and could even increase it.

Similarly, it is unclear how defederation will affect toxic posting behavior. Some past studies of deplatforming have found that toxicity increases after some deplatforming events (Ali et al. 2021; Buntain et al. 2023; Chandrasekharan et al. 2017). Such increases may happen when group-level sanctions provoke blowback or revolts against the imposition of sanctions in the form of non-compliance with sanctions or increasing anti-social behavior (Heckathorn 1988). However, other studies have found that group-level sanctions decrease or do not change toxicity (Jhaver et al. 2021; Chandrasekharan et al. 2022).

Therefore, we anticipate that defederation will have limited or negligible effects on toxic posting behavior. Accounts on blocked servers may not receive any indication of the reasons for the intervention. While server administrators may choose to publicize their defederation decisions, no mechanism propagates the decisions to the users of targeted servers (much less the users whose behavior may have triggered the block in the first place). Users on the blocking serer who had interacted across the severed connection likewise experience no other direct effects and retain access to the rest of their communication networks. In the absence of more targeted or focused information about the reasons behind the block, we do not expect it to impact toxic behavior substantially.

Data, Measures, and Methods

We pursue an observational, quasi-experimental research design to identify effects of defederation on the activity and toxic posting behavior in the Fediverse. We collected longitudinal trace data from 7,445 publicly listed defederation events and about 104 million public posts that occurred in the Fediverse on either the Mastodon or Pleroma networks between April 2, 2021 and May 31, 2022. Using this data, we analyze activity of user accounts (for RQ1) and the toxicity of their messages (for RQ2) on the blocking and blocked servers impacted by these events in comparison to matched control accounts. We apply a difference-in-differences approach and present both non-parametric and parametric estimates of the effects of defederation.

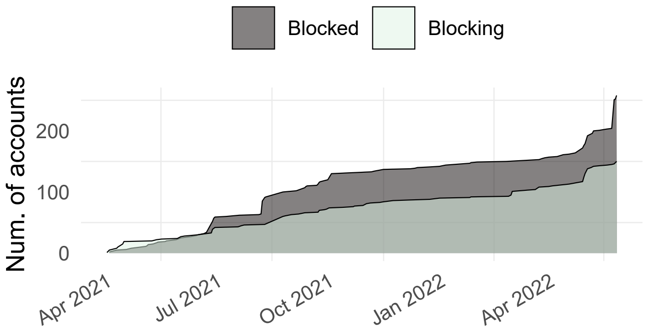


Figure 2: The y-axis shows the cumulative number of blocked and blocking accounts included in our analysis over our study period.

Empirical Setting

The Fediverse is a decentralized social network comprised of many servers which each institute their own rules, policies, and moderation practices. Each account connects to the network through a home server and can send and receive posts from accounts both within and outside the home server via the ActivityPub protocol. Mastodon and Pleroma are two popular, interoperable ActivityPub-based microblogging systems with overlapping but distinct features.

Fediverse administrators have nearly complete control over their servers, from choosing and configuring software to setting rules. However, administrators only have direct control over their own server. Although users of one server can report rule-breaking posts by another server's accounts, only the administrator of an account's home server can sanction the user, such as by removing the post or sending a warning. Consequently, when two servers come into conflict, administrators must either use persuasion, ignore the problem, or block or filter the other server.

Defederation is the process of severing ties between servers to restrict interactions, typically due to differences in policies, norms, or content. Administrators are responsible for defederation decisions, and they can choose to defederate from specific servers. Defederation helps maintain the autonomy of servers and allows them to adhere to their own values and rules. Individuals with accounts on Fediverse servers do not tend to receive direct notifications about defederation events, though they may either see an announcement by their server administrator. On Mastodon, defederations delete records of the defederated server, meaning that affected accounts lose followers from the defederated server and no longer follow accounts on the defederated server.

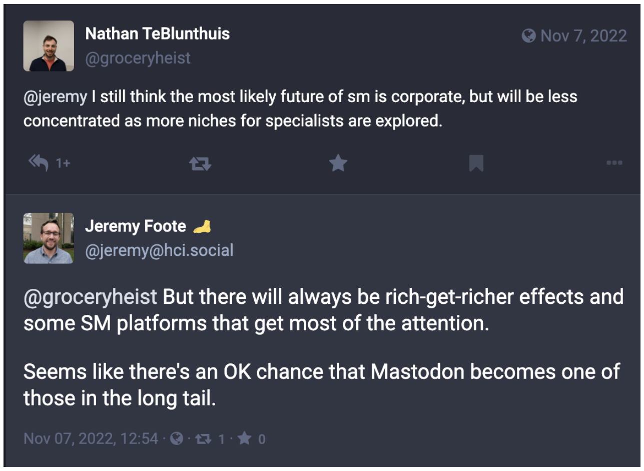
Data Collection

We first identify a list of Fediverse servers involved in defederation by compiling a set of known servers which federate with other servers on the list, starting with the set of servers on mastodon.social[1]. We then iteratively add servers with

[1] https://mastodon.social/

Term	Description
Fediverse	A group of independently operated social media servers which interoperate with shared protocols.
Defederation	A server-to-server block between servers in a federated system.
Mastodon	The most popular and widely used software on the Fediverse.
Pleroma	A lightweight Fediverse software with an API similar to Mastodon and an alternative feature set.
Server administrator	A person or group who controls the technical infrastructure and configuration of a Fediverse server.
Group sanction	A moderation action against a group, not an individual.

Table 1: Glossary of key terms.



Nathan TeBlunthuis
@groceryheist
Nov 7, 2022

@jeremy I still think the most likely future of sm is corporate, but will be less concentrated as more niches for specialists are explored.

Jeremy Foote
@jeremy@hci.social

@groceryheist But there will always be rich-get-richer effects and some SM platforms that get most of the attention.

Seems like there's an OK chance that Mastodon becomes one of those in the long tail.

Nov 07, 2022, 12:54 · · 1 · 0

Figure 3: A screenshot shows a cross-server interaction on Mastodon. Note that the account name of the replying user indicates their home server (*@user@server*).

peer connections to members of this set using an API feature which discloses server-to-server connections. This approach identified 11,025 servers and 7,445 defederation records. We collected data from these servers by running a script daily. The script uses publicly available and documented REST APIs provided by Mastodon and Pleroma to collect timelines of posts. We collected approximately 104 million posts in total. The post data allows us to identify accounts that engaged in cross-server interaction because usernames indicate the home server.

We identify defederation from the public records servers publish of which other servers they block. We check these records each day and infer the date that defederation occurred by the day a server first appears in the list. We exclude 472 defederation events where we do not know the exact date of defederation because we are missing a daily snapshot of the record. We only analyzed defederation events where the public record persisted at least 92 days after the initial event, excluding 1,347 that did not. Accounts are affected by a defederation event if they had previously sent a reply to an account on the blocked or blocking server. We excluded accounts that may have been affected by defederation events we dropped due to missing data issues.

Our analysis is designed to identify the effects of defederation accounting for trends in activity and toxicity before and after defederation events (defined below). We therefore analyze data from the 91 days prior to and following each defederation event (dropping the day of the defederation itself). The outcomes of interest are quantities and qualities of posting activity, so we exclude inactive accounts by dropping those that posted fewer than 10 times in the 91 days before a defederation event involving their home server. We also include only accounts that posted at least once prior to the 91 day period and at least once in the final 45 days of this period.

Some servers experience multiple defederation events. In order to avoid potential confounding of our estimates from multiple exposures to defederation, we exclude accounts that experienced defederation events on multiple days. After dropping some additional accounts in the matching process (described below), the resulting dataset includes 258 accounts whose 71 home servers were defederated and 150 accounts whose 52 home servers defederated another server. This included a total of 214 defederation events.

Measures

We construct measures within the analytic window for all accounts and servers involved in defederations. We also aggregate our measures over 7 day periods (weeks).

Defederation events: We observe defederation events via the records of federation policies published by Mastodon and Pleroma servers within the Fediverse systems we study. We collected data on 440 blocking servers and 1,136 blocked servers. We record the timestamp of each defederation event, which we use to construct a timeline of defederation events.

Activity: The response variable of our analysis for RQ1 is the posting $activity$ of user accounts affected by defederation events. We measure activity as the number of posts an account makes each week.

Toxic behavior: RQ2 investigates the impact of defederation on toxic posting behavior. We measure post toxicity using the Perspective API model for toxicity, which uses machine learning to predict if a post is a ”rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion” [2]. For each post, the model outputs a score ranging from 0–1 corresponding to the estimated probability that the post contains toxic speech. For each account, we aggregated their toxicity during each week by taking the mean toxicity score for their posts during that window.

[2] https://www.perspectiveapi.com/

Blocked user accounts: Defederation is a directional action from one server to another and may not be reciprocated. In our research questions, we estimate the effects of user accounts being blocked from cross-server interactions in a defederation event. We assign accounts to the *blocked treatment group* if (1) their home server was defederated (blocked) by another server (the blocking server), (2) they replied to an account on the blocking server in the 91 days before defederation, and to ensure the account is active at the time of treatment, (3) they post at least once in the last 45 days prior to defederation . For clarity we refer to the blocked treatment group as U_0 below.

Blocking user accounts: In our research questions, we are also interested in effects of defederation on the blocking servers' accounts. We assign accounts on the blocking server to the blocking treatment group (U_1) if they meet the parallel conditions to those described above for the blocked treatment group.

Time: As noted above, we aggregate our measures over 7 day periods. We therefore measure ($time$) as the number of such weekly periods before/after the defederation event affecting the account in question excluding the day defederation occurred.

Analytic Plan

We pursue an identification strategy to estimate the effects of defederation events. For both sets of outcomes—activity levels and toxic posts—we visualize time series, conduct nonparametric statistical tests, and perform regression analysis using difference-in-differences (DiD) estimators that infer the effects of the treatment on treated accounts from both blocked and blocking servers against a set of matched controls. We use coarsened exact matching to construct this synthetic control group of accounts that are similar to the accounts impacted by defederation (treated accounts).

Constructing Matched Control Groups

All of our analyses depend on comparing user accounts that experienced defederation events with similar accounts that did not. The goal of statistical matching is to reduce bias and account for potential confounders. In this case, we use observable, pre-treatment attributes of accounts that capture how they engage with the wider Fediverse through posting replies to external servers. We match on the following variables: total number of account posts, the number of posts in the 45 days prior to treatment ("post count (45)"), the number of replies sent, the number of other servers engaged with replies, and the number of active accounts on their home server. We applied a log transformation to each of these (highly skewed count) variables. We construct all matching measures over weeks during the 91 days prior to the defederation event experienced by the treated user in question.

We use coarsened exact matching (Iacus, King, and Porro 2012) with Sturges' rule to determine the number of bins for all variables except the number of active accounts on the server, for which we used four bins to be less strict. We used one-to-one matching, selecting the closest match according to Mahalanobis distance and discarded accounts for which there was not a sufficiently good match (139 for the blocked accounts; 63 for the blocking accounts). Accounts that were treated were not eligible to be matched controls.

We denote the resulting matched control groups as C_0, the control group for blocked server users U_0, and C_1, the control group for blocking server users U_1. The blocked accounts had 258 matched units from 397 potential units; the blocking accounts had 150 from 213 potential units. Figure 4 shows the covariate balance using standardized mean differences before and after matching across these groups.

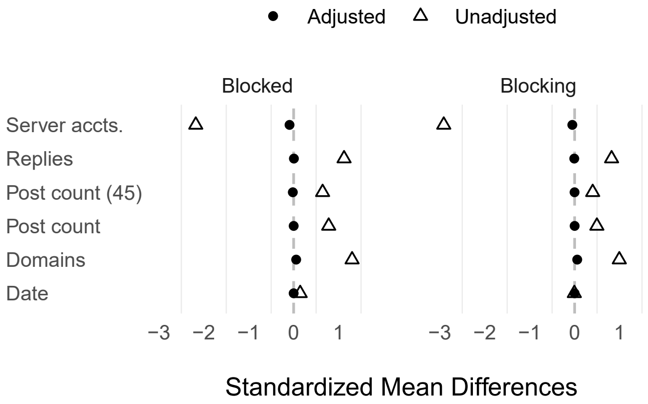


Figure 4: A covariate balance plot shows the standardized mean difference between treatment and control groups for each measure used in our matching procedure before (unadjusted) and after (adjusted) matching. Our procedure effectively found a group of matched controls similar to the treated accounts along these measures.

Analysis

Our analysis proceeds in three parts: descriptive comparison and visualization; non-parametric tests; and difference-in-differences estimation.

Descriptive comparison and visualization: We calculate descriptive and summary statistics for all measures across the treated and control groups. We also visualize posting activity before and after defederation, plotting the number of weekly posts made by the median user account in each of our four study groups (U_0, U_1, C_0, C_1) in the 12 weeks prior to and following defederation with 95% confidence intervals based on order statistics. Both the descriptive analysis and the visualization justify and complement our non-parametric and difference-in-differences analyses.

Non-parametric tests: We first compare changes in activity level and median post toxicity before and after defederation using a non-parametric Wilcoxon signed-rank test, which returns a W test statistic representing the sum of the ranks of the positive differences between paired observation and a p-value which compares to the null hypothesis that the changes between the two groups are zero. This test is robust to outliers, independent of distributional assumptions, and assigns equal weight to each user account irrespective of their activity levels. For toxicity in particular, this non-parametric approach reduces threats of bias due to the opaque machine learning systems that create the values reported from the Perspective API. We conduct this test over all user accounts in both the blocked and blocking groups.

Difference-in-differences (DiD) estimation: Our primary analysis estimates the causal effect of defederation on the *blocked accounts* (U_0) and the *blocking accounts* (U_1) using a difference-in-differences (DiD) framework. DiD systematically quantifies the changes in a treatment group (i.e., blocked accounts; U_0) following an intervention (i.e., a defederation event) relative to changes in a control group (i.e., accounts matched to blocked accounts; C_0). It does this by modeling temporal trends for each group (treatment; U_i and control; U_i) before and after the defederation as well as group-dependent discontinuous "jumps" at the moment of intervention.

Our DiD estimates support causal inferences under several assumptions: The matching procedure should result in control groups of accounts that were not affected by defederation, but were just as likely to have been affected as the accounts that were, the two groups' trends in the outcome variables should be parallel prior to treatment, and the trends should be linear. If so, then the difference between the jumps of the two groups quantifies a "local average treatment effect" (LATE).[3] We include random intercepts in our model to account for variability between subjects. We fit versions of this model for both outcomes and types of server (blocked and blocking) using the `brms` R package.

Model for activity: $Activity$, the number of posts an account makes per week, is an over-dispersed count variable. We therefore use a negative binomial model formally represented as:

$$
\begin{aligned}
y_{i,t} &\sim NegBinomial(\mu_{i,t}, \phi) \\
\mu_{i,t} &= \beta_0 + \zeta_i + \beta_1 I(i \in U_k) \\
&\quad + \beta_2 I(t > T) + \beta_3 t + \beta_4 I(t > T) \cdot t \\
&\quad + \beta_5 I(i \in U_k) \cdot I(t > T) + \beta_6 I(i \in U_k) \cdot t \\
&\quad + \beta_7 I(t > T) \cdot I(i \in U_k) \cdot t
\end{aligned}
$$

Where $y_{i,t}$ is the activity of user account i in week t. T indicates the time of defederation, U_k is the treatment group (U_0 is the set of blocked accounts and U_1 is the set of blocking accounts), I is the indicator function equal to 1 when its parameter is true and 0 if not, and ζ_i is the random intercept for user account i. Our parameter of interest is β_5, the coefficient for post-treatment membership in the treatment group, for testing whether activity increases ($\beta_5 > 0$) or decreases ($\beta_5 < 0$) upon defederation. The model's other terms are intercepts and trends pre-treatment (β_0,β_3) and post-treatment (β_2,β_4,β_6) and membership in the treatment group (β_1).[4]

Model for toxicity: Our measure for toxicity is calculated at the post level, leading to several key differences in how we model this outcome. Instead of aggregating this at the account level, which would obscure variation, we use a multilevel model with random slopes for group to estimate the effects of defederation as an account-level treatment on post-level toxicity. We use beta regression for our analysis of toxicity because the Perspective API scores are on a range between 0 and 1 and are designed to quantify the probability that a comment is toxic. Our model of toxicity is thus:

$$
\begin{aligned}
y_{k,i,t} &\sim Beta(\mu_{i,t}, \phi) \\
\mu_{i,t} &= K_0 + \zeta_i + \eta_0(i \in U_k) + \eta_1(i \notin U_k) \\
\eta_0 &= \beta_0 + \beta_1 I(t > T) + \beta_2 t + \beta_3 I(t > T) \cdot t \\
\eta_1 &= \beta_4 + \beta_5 I(t > T) + \beta_6 t + \beta_7 I(t > T) \cdot t
\end{aligned}
$$

where $y_{k,i,t}$ is the toxicity of a given post k by account i in week t, K_0 is the overall intercept, ζ_i are the random intercepts for each account, η_0 are random effects for the treatment group and η_1 are random effects for the matched control group.

Results

Figure 5 summarizes our dependent variables across the treatment and matched control groups during the pre- and post-treatment periods. Below, we present results for the two outcomes—activity and toxicity—separately.

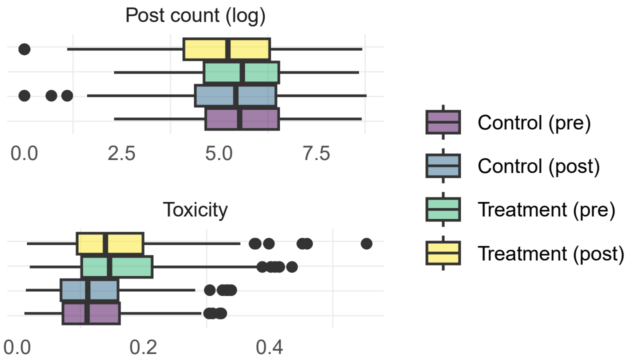


Figure 5: Box and whisker plots visualize the distributions of our dependent variables within the blocked and blocking groups of user accounts and their matched controls before and after defederation. The lines correspond to the median, the boxes to the inter-quartile range (IQR), the whiskers to the range of the data within 1.5 * IQR, and the dots to data points outside the range of the whiskers.

Effects on Activity

Research question 1 asks how defederation affected activity among user accounts whose home server either blocked

[3] We say *local* average treatment effect because the estimate applies in specific times and places that may not generalize to other contexts.

[4] We note that the β_6 and β_7 coefficients model the group-specific trends before and after treatment. If these are both 0, this would provide evidence supporting the parallel trends assumption of our DiD models. As shown in Table 3, the 95% credible intervals for both coefficients contain 0, suggesting the assumption is reasonable.

(blocking group) or was blocked by (blocked group) another server. Our visualization of the time series of affected accounts compared to matched controls, non-parametric tests, and difference-in-differences analysis all indicate that blocked group users decreased their activity following defederation while changes among blocking group users were minor or noisy.

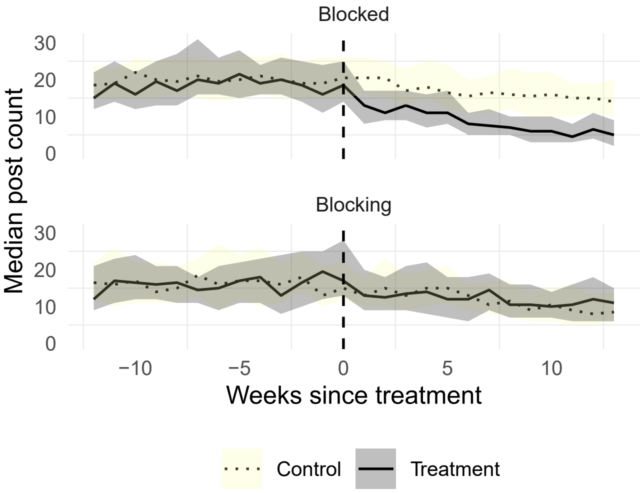


Figure 6: Visualization of activity among blocked and blocking user accounts shows an asymmetric change in activity following defederation. An account with a median post count on the blocked server declines in activity much more rapidly following defederation compared to matched controls while an account with a median post count on the blocking server declines similarly to matched controls.

Before defederation, the trends and weekly median activity levels remained comparable for accounts on both the blocked and blocking servers. Although all groups reduced activity somewhat after defederation, Figure 6 shows the activity levels on the blocked servers quickly diverged from their matched controls after defederation. For instance, while the median post count for treated accounts on blocked servers (U_0) was 18.5 posts one week before defederation, this dropped to 13 posts one week after defederation. Compare this to the analogous median post count among matched controls which was 20.5 posts per week both before and after defederation. In contrast, the median post count among treated accounts on blocking servers (U_1) was 17 in the week before and 13 in the week after defederation compared to 15 and 13 for the matched controls (C_1).

Non-parametric tests: This strong response among the treated accounts on the blocked servers (U_0) also appears in the non-parametric test based on post counts summed for all weeks and shown in Table 2. Prior to defederation, median activity among blocked accounts and matched controls was similar, yet following treatment, activity for the treated blocked accounts decreased by a median of 135.5 posts (21.9%) compared to 18 posts (9.5%) for the matched controls. By contrast, among the user accounts on the blocking servers, we observe very similar post-defederation activity decreases for the treatment and matched control accounts on the blocking server with a decrease in the median of 54.5 posts (10.7%) by treated accounts compared to 53.5 post (11.1%) by the the matched controls. In sum, we observe asymmetric effects of defederation: activity decreases for affected accounts on blocked servers, but not on blocking servers.

Group	median	W	p
U_0	-135.5	41197.5	0.000
C_0	-18.0	35762.0	0.143
U_1	-54.5	12413.0	0.122
C_1	-53.5	12520.0	0.091
Δ_0	-39.0	39927.0	0.000
Δ_1	3.0	10645.5	0.421

Table 2: Non-parametric tests for differences in activity before and after defederation events (summed across all weeks) find a measurable decrease in posting activity for the accounts on blocked servers compared to matched controls but no such change for accounts on blocking servers.

DiD Analysis: Our DiD results corroborate the same pattern: decreased activity among users on the blocked servers, and no change among users on the blocking servers. Table 3 shows the regression results for both models. Figure 7 plots the modeled activity levels of a median account in each group along with the corresponding 95% credible intervals quantifying uncertainty in the model's parameter estimates. User accounts on the the blocked servers (U_0) decrease their activity compared to matched controls ($\beta_5 = -0.24; SE = 0.07$). For a user account with median pre-treatment activity levels on the blocked server, this corresponds to a reduction from 24.9 posts per week immediately before treatment to 19.6 posts per week immediately after treatment (a 21.4% reduction), compared to 22.8 posts per week after treatment for its matched control.

Consistent with our other results for user accounts on the blocking server, the 95% credible interval for the effect of defederation on activity by accounts on the blocking server (U_1) contains 0 ($\beta_5 = -0.08; SE = 0.08$).

Effects on Anti-Social Behavior

We now present results for our second research question on the effects of defederation on the toxicity of posts by user accounts on the blocked and blocking servers. Our three analyses: visualization of the time-series of affected accounts compared to matched controls, non-parametric tests, and difference-in-differences analysis are in agreement that defederation did not result in a statistically detectable change in toxicity for affected accounts on either server.

The median toxicity plotted in Figure 8 remained at consistent levels for accounts which posted on a given week for both treatment and matched control groups on both the blocked and blocking servers. The trends remained sim-

	Blocked				Blocking			
Term	Estimate	Std error	2.5%	97.5%	Estimate	Std error	2.5%	97.5%
β_0 (Intercept)	3.216	0.086	3.049	3.380	2.901	0.118	2.684	3.121
β_1 Group	-0.024	0.116	-0.255	0.198	0.023	0.177	-0.290	0.367
β_2 Treatment	-0.089	0.045	-0.178	0.002	-0.080	0.060	-0.201	0.034
β_3 Time	0.006	0.004	-0.002	0.015	-0.002	0.006	-0.014	0.009
β_4 Treatment : Time	-0.026	0.006	-0.038	-0.014	-0.027	0.008	-0.042	-0.010
β_5 Group : Treatment	-0.241	0.065	-0.367	-0.116	-0.084	0.082	-0.247	0.072
β_6 Group : Time	-0.007	0.006	-0.020	0.004	0.006	0.008	-0.010	0.021
β_7 Group : Treatment : Time	-0.015	0.009	-0.031	0.002	0.009	0.011	-0.013	0.032

Table 3: Difference-in-differences analysis of activity level for user accounts whose server was defederated (blocked group) or whose server defederated another (blocking group). The 95% credible interval negative coefficient for membership in the blocked group post-defederation (β_5) is less than 0, indicating that activity by accounts in this group decreased more than accounts in the matched control group. We do not draw such a conclusion about members of the blocking server because the corresponding credible interval contains 0.

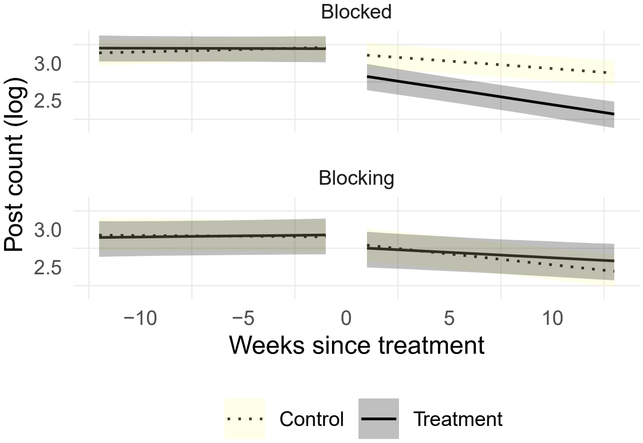


Figure 7: A marginal effects plot visualizes results from our difference-in-differences analysis of account activity by week. We observe a discontinuous activity decrease among accounts on the blocked server that exceeds any decrease in the matched controls, but no corresponding change among accounts on the blocking server. The bands are 95% credible intervals that account for uncertainty in parameter estimates.

ilar before defederation between the treatment and control groups for accounts on both the blocked and blocking servers.

Non-parametric tests Table 4 summarizes aggregate changes in anti-social behavior (toxicity) after defederation for all groups. Anti-social behavior by user accounts on blocked servers decreases slightly. For example, the median toxicity score for posts from these accounts declined by 0.005, but we lack sufficient statistical power to distinguish this from the null hypothesis that there was no difference in changes between treatment and control groups on the blocked server ($p = 0.072$). We do not observe even such a small change in toxicity among posts from accounts on the blocking server.

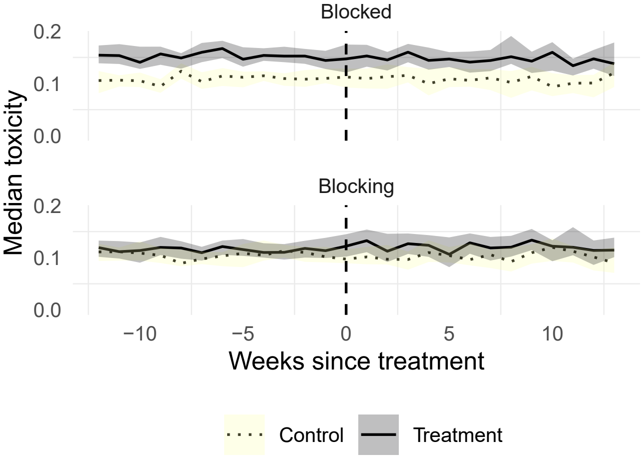


Figure 8: Median toxicity among accounts which posted each week for blocked and blocking user accounts. The median toxicity remained flat for all groups.

DiD Analysis Again, the results of the DiD analysis for toxicity tell a similar story to the data visualization and non-parametric tests. Our models estimate that toxicity levels of posts from affected accounts on both blocked and blocking servers do not change following defederation events. Full regression results are in Table 5. We find no evidence for any significant changes in post toxicity among accounts on either blocked or blocking servers.

Discussion

We investigated defederation, a group-level sanction where administrators of the blocking server disconnect communication channels between users of their server and the blocked server. Although one of few actions by which leaders in such networks can protect their community's servers from trolls, harassers, and objectionable content from other

Group	median	W	p
U_0	-0.006	17746	0.538
C_0	0.004	14000	0.950
U_1	-0.008	6514	0.619
C_1	0.001	5546	0.873
Δ_0	-0.005	17161	0.072
Δ_1	0.000	6414	0.305

Table 4: Non-parametric difference-in-differences for median post toxicity before and after de-federation events. The W test statistic represents the sum of the ranks of the positive differences between paired observations while the p-value compares to the alternative hypothesis that the changes are zero.

servers, defederation's consequences are poorly understood. Therefore, in a quasi-experimental study, we estimate the effects of defederation on the activity and toxicity of users on both the blocking and blocked servers who had sent or received messages across server boundaries. Our results provide the first evidence concerning the impacts of defederation in Fediverse communication networks. This evidence can inform governance decisions among Fediverse server administrators and moderators and the design of future decentralized social media networks.

Research question 1a asked how defederation changed posting activity by affected users on the blocked server. We find that losing communication channels to users on the blocking server causes users on the blocked server who had used such channels to decrease their activity. This helps us understand defederation's potential as an intervention against undesirable behavior. Not only does it protect the blocking server's users from the blocked server, it may decrease the blocked server's value. Defederation may cause a user on the blocked server to lose valued audiences, sources of content, or interpersonal connections and therefore to become less active.

If this positive effect of defederation results from lost connections, defederation might also have a negative consequence: Users of the blocking server may decrease their activity for the same reason. However, our results for research question 1b, which inquired into defederation's consequences on posting activity by affected users on the blocking server, show no measurable effect. These asymmetric effects of defederation on posting activity suggest that administrators may find defederation an effective response to undesirable behavior on other servers without negative impacts on activity by affected user accounts on their own server.

Although sanctions are typically intended to promote normative compliance, group-level sanctions in some prior empirical studies have been met with increases in anti-social behavior or rebellious non-compliance (Mitts, Pisharody, and Shapiro 2022; Ribeiro et al. 2021; Russo et al. 2023). Therefore, we investigated defederation's effects on toxicity in research question 2. If defederation causes an increase in anti-social behavior, we would expect an increase in toxicity on the blocked server. If defederation causes non-compliance, we would expect toxicity on the blocking server to increase. However, we do not find evidence of either of these adverse indirect outcomes. In this sense, defederation appears similar to individual and collective sanctions analyzed in prior studies such as account banning, quarantine, and sub-community removal (Chandrasekharan et al. 2022, 2017; Jhaver et al. 2021). Some of these group-level sanctions have even improved compliance with widely-held norms (Chandrasekharan et al. 2017). If defederation also does so then we would expect affected users on the blocked server to become less toxic. Yet we observe no such effect.

Among the various kinds of content moderation and governance interventions available to social media systems administrators, server defederation may be among the more extreme group sanctions. However, our analysis suggests that it may also provide administrators with an effective response to undesirable behavior that does not undermine the activity of affected user accounts on the server implementing the block. User accounts affected by defederation events can continue to communicate with others on their respective servers as well as other servers where federated interactions remain permitted. The disconnection from the network is not total and can respect the local norms that operate among distinct sub-components of the larger network. We believe this is a design feature of federated systems that distinguishes them from other kinds of content moderation and governance systems in social media environments. Other, more centralized, platforms may wish to experiment with similar features as doing so may advance user trust and safety without undermining activity.

That said, further research is needed to support recommendations about defederation's overall benefits and limitations in decentralized social media systems. Such research should uncover the mechanisms that drive the observed changes in activity levels and toxic behaviors, such as changes in user motivations or migration to alternative servers or communication channels. Future work should also investigate how defederation events are perceived and experienced by the people operating affected accounts on both blocked and blocking servers. In addition, insights into the reasons behind administrator decisions to implement defederation blocks as well as the timing of such changes would enrich the findings reported here. As noted above, the defederation events in our study did not occur at random and were likely undertaken with some awareness of the various actors involved. For these reasons, we encourage caution around the generalizability of our findings beyond the specific context of our study.

Additional limitations of the study relate to the constraints around our data collection, measurement, and analysis strategies. First, our data collection was limited to public status posts, which do not fully capture user interactions within the Fediverse. Private posts and direct messages, which are not accessible through the APIs we used, might exhibit different patterns in response to defederation events. In addition, some servers choose not to publish defederation information. If the effects of defederation emerge anywhere other than public posts, our analysis could not capture them.

Our analysis also relied on the Perspective API to analyze

		Blocked				Blocking			
Group	Term	Estimate	Std error	2.5%	97.5%	Estimate	Std error	2.5%	97.5%
Treatment	β_0 (Intercept)	-0.143	0.438	-1.197	0.763	-0.042	0.284	-0.754	0.518
	β_1 Treatment	0.002	0.006	-0.009	0.018	-0.001	0.010	-0.020	0.017
	β_2 Time	0.001	0.001	0.000	0.002	0.001	0.001	-0.001	0.003
	β_3 Treatment : Time	-0.002	0.001	-0.004	0.000	-0.001	0.002	-0.004	0.003
Control	β_4 (Intercept)	0.136	0.436	-0.907	1.044	0.048	0.283	-0.613	0.649
	β_5 Treatment	0.000	0.006	-0.016	0.011	0.004	0.009	-0.014	0.024
	β_6 Time	0.001	0.001	0.000	0.002	0.003	0.001	0.001	0.005
	β_7 Treatment : Time	-0.005	0.001	-0.007	-0.003	-0.006	0.002	-0.009	-0.002

Table 5: Beta regression coefficients drawn from the posterior of the parametric toxicity DiD model for user accounts whose server was defederated (blocked group) or whose server defederated another (blocking group). For all groups, the 95% credible intervals for a change in toxicity levels after treatment (β_1, β_5) contain 0.

the content of posts by quantifying their toxicity. While this tool provides a useful proxy, its accuracy may vary across languages and contexts and its errors may affect our statistical results (TeBlunthuis, Hase, and Chan 2023). Moreover, it is conceivable that the contexts for which Perspective designed depart from the Fediverse so substantially that Perspective fails to detect forms of misbehavior affected by defederation. Future research could explore alternative methods for assessing content quality and toxicity. Although language models like those utilized in the Perspective API have significant limitations, we believe that these are likely to be time invariant, at least in the short-run. That is, they would happen on both sides of a defederation event. In addition, defederation might affect forms of behavior beyond toxicity that future research may investigate such as topics of post or the extent to which they elicit replies or boosts (analogous to retweets).

In terms of the analysis, our study incorporated several critical assumptions and focused on the effects of defederation events within a relatively narrow timeframe. Matching entails an assumption that selection into treatment or control occurred due to observable variables incorporated into the matching process. A quasi-experimental difference-in-differences analysis similarly assumes parallel trends and comparability across units conditional only on “as-if” random exposure to the treatment in question. For example, possible advanced announcements of defederation events may drive changes in user behavior prior to defederation itself. While we tried to support such assumptions empirically, they may not hold uniformly, and in addition, the effects of defederation on user behavior might change over time as users adapt to new norms and technologies. Additional and longer duration analyses could address these concerns and shed light on the long-term consequences of defederation events.

Conclusion

In this study, we investigated the effects of defederation events on the activity levels and toxic posting behavior of accounts in the Fediverse. The results indicate that such events produce asymmetric effects on activity for affected accounts on blocked servers versus those on blocking servers with no increase in toxicity for any groups. The results also highlight the potential of decentralized social networks and their unique mechanisms, such as defederation, in providing communities with tools to manage content moderation and other aspects of online interactions. Future research could explore the causes or reasons behind defederation events, the long-term consequences of defederation, as well as the mechanisms by which defederation produces (asymmetric) effects. By continuing to study the Fediverse and its affordances, we can better understand how to foster healthy online communities and effective content moderation strategies in a decentralized environment.

Acknowledgements

The authors would like to thank Ceren Budak and the Community Data Science Collective. Support for this work was provided by the U.S. National Science Foundation (IIS-191020, IIS-1908850). Additionally, Dr. TeBlunthuis was partially supported by NSF grant IIS-1815875.

References

Albert, R.; Jeong, H.; and Barabási, A.-L. 2000. Error and Attack Tolerance of Complex Networks. *Nature*, 406(6794): 378–382.

Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Understanding the Effect of Deplatforming on Social Networks. In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci ’21, 187–195. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8330-1.

Anaobi, I. H.; Raman, A.; Castro, I.; Zia, H. B.; Ibosiola, D.; and Tyson, G. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023*, WWW ’23, 3109–3120. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-9416-1.

Buntain, C.; Innes, M.; Mitts, T.; and Shapiro, J. 2023. Cross-Platform Reactions to the Post-January 6 Deplatforming. *Journal of Quantitative Description: Digital Media*, 3.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2022. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. *ACM Transactions on Computer-Human Interaction*, 29(4): 29:1–29:26.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW): 31:1–31:22.

Datta, A.; Buchegger, S.; Vu, L.-H.; Strufe, T.; and Rzadca, K. 2010. Decentralized Online Social Networks. In Furht, B., ed., *Handbook of Social Network Technologies and Applications*, 349–378. New York, NY: Springer US. ISBN 978-1-4419-7142-5.

Frey, S.; and Schneider, N. 2021. Effective Voice: Beyond Exit and Affect in Online Communities. *New Media & Society*, 14614448211044025.

Frey, S.; and Sumner, R. W. 2019. Emergence of Integrated Institutions in a Large Population of Self-Governing Communities. *PLOS ONE*, 14(7): e0216335.

Fulk, J.; Flanagin, A. J.; Kalman, M. E.; Monge, P. R.; and Ryan, T. 1996. Connective and Communal Public Goods in Interactive Communication Systems. *Communication Theory*, 6(1): 60–87.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press. ISBN 978-0-300-17313-0.

Gruzd, A.; Soares, F. B.; and Mai, P. 2023. Trust and Safety on Social Media: Understanding the Impact of Anti-Social Behavior and Misinformation on Content Moderation and Platform Governance. *Social Media + Society*, 9(3): 20563051231196878.

Han, C.; Kumar, D.; and Durumeric, Z. 2022. On the Infrastructure Providers That Support Misinformation Websites. *Proceedings of the International AAAI Conference on Web and Social Media*, 16: 287–298.

Heckathorn, D. D. 1988. Collective Sanctions and the Creation of Prisoner's Dilemma Norms. *American Journal of Sociology*, 94(3): 535–562.

Hill, B. M.; and Shaw, A. 2021. The Hidden Costs of Requiring Accounts: Quasi-experimental Evidence from Peer Production. *Communication Research*, 48(6): 771–795.

Iacus, S. M.; King, G.; and Porro, G. 2012. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 20(1): 1–24.

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 381:1–381:30.

Kaye, D. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports. ISBN 978-0-9997454-8-9.

King, G.; Pan, J.; and Roberts, M. E. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, 107(2).

Mitts, T.; Pisharody, N.; and Shapiro, J. 2022. Removal of Anti-Vaccine Content Impacts Social Media Discourse. In *14th ACM Web Science Conference 2022*, 319–326. Barcelona Spain: ACM. ISBN 978-1-4503-9191-7.

Myers West, S. 2018. Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms. *New Media & Society*, 20(11): 4366–4383.

Ribeiro, M. H.; Hosseinmardi, H.; West, R.; and Watts, D. J. 2023. Deplatforming Did Not Decrease Parler Users' Activity on Fringe Social Media. *PNAS Nexus*, 2(3): pgad035.

Ribeiro, M. H.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_donald and r/Incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 316:1–316:24.

Russo, G.; Verginer, L.; Ribeiro, M. H.; and Casiraghi, G. 2023. Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning. *Proceedings of the International AAAI Conference on Web and Social Media*, 17: 742–753.

Seering, J.; Kraut, R.; and Dabbish, L. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 111–125. New York, NY, USA: ACM. ISBN 978-1-4503-4335-0.

Srinivasan, K. B.; Danescu-Niculescu-Mizil, C.; Lee, L.; and Tan, C. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the Changemyview Community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 163:1–163:21.

Stocking, G.; Mitchell, A.; Matsa, K. E.; Widjaya, R.; Jurkowitz, M.; Ghosh, S.; Smith, A.; Naseer, S.; and Aubin, C. S. 2022. The Role of Alternative Social Media in the News and Information Environment. Technical report, Pew Research Center, Washington D.C.

TeBlunthuis, N.; Hase, V.; and Chan, C.-H. 2023. Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, 18(1).

Tran, C.; Champion, K.; Hill, B. M.; and Greenstadt, R. 2022. The Risks, Benefits, and Consequences of Prepublication Moderation: Evidence from 17 Wikipedia Language Editions. *Proceedings of the ACM on Human-Computer Interaction*, 6: 1–25.

Van Hove, L. 2016. Testing Metcalfe's Law: Pitfalls and Possibilities. *Information Economics and Policy*, 37: 67–76.

Vu, A. V.; Hutchings, A.; and Anderson, R. 2023. No Easy Way Out: The Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment. arxiv:2304.07037.

Zhang, X. M.; and Zhu, F. 2011. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review*, 101(4): 1601–1615.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. We specifically limited the scope of this research to posts on the public internet and additionally analyze this data in aggregate and do not reveal the identity of the accounts or servers included in this study.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, our claims relate to our empirical findings based on our study design within the context of the study.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, our methodological approach is to construct a quasi-experiment to support claims based on both the empirical setting and the study design. We are careful to scope our claims to what can be reasonably supported given our study design.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we present the distributions for the independent variables considered in the study. We limit the scope of our claims to accounts on the Fediverse who were affected by a defederation event.

   (e) Did you describe the limitations of your work? Yes, we discuss several of the limitations to our study design in the discussion and additionally provide suggestions for appropriate interpretations of our findings.

   (f) Did you discuss any potential negative societal impacts of your work? Yes, we address tradeoffs to the effects of the interventions we studied. That said, we do not anticipate much potential for negative social impacts from our work, which considers the effects of an existing content moderation mechanism as implemented by communities. Although the motivations for defederation are diverse, we anticipate that more rigorous empirical analysis of such interventions will primarily benefit content moderation practitioners.

   (g) Did you discuss any potential misuse of your work? We do not directly discuss misuse, though we seek to narrowly contextualize our findings and encourage further investigation to better understand the effects we observe.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? We describe these steps under ”Broader impacts and ethical considerations” below.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, and we believe this work can help contribute to human well-being, avoids active harms, and presents itself honestly.

2. Additionally, if your study involves hypotheses testing...

   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, we motivate our analyses in terms of research questions that are theoretically derived from past work. We clearly state the assumptions of our analytical strategy

   (b) Have you provided justifications for all theoretical results? Yes, we have done our best to justify any theoretical results with appropriate evidence.

   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.

   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

   (e) Did you address potential biases or limitations in your theoretical framework? Yes

   (f) Have you related your theoretical results to the existing literature in social science? Yes

   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes

3. Additionally, if you are including theoretical proofs...

   (a) Did you state the full set of assumptions of all theoretical results? Not applicable. No theoretic proofs.

   (b) Did you include complete proofs of all theoretical results? Not applicable. No theoretic proofs.

4. Additionally, if you ran machine learning experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? While we did apply an existing machine learning model as one of our dependent variables, we did not run any machine learning experiments.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? We did not create or train any machine learning models.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? We did not create or train any machine learning models.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? We did not create or train any machine learning models.

   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? We did not run any machine learning experiments.

   (f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? We did not run any machine learning experiments.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? The primary existing asset we use is the Perspective API, which we cite.

(b) Did you mention the license of the assets? We did not mention the license. The API terms of service are publicly available and far too long to reproduce in a paper.

(c) Did you include any new assets in the supplemental material or as a URL? We are not releasing new assets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? We obtained consent from Perspective, but do not discuss this because documentation on how to access to Perspective is readily available. Regarding data produced by Fediverse users, see ”Broader impacts and ethical considerations” section below.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Perspective doesn't provide PII, so this is not worth mentioning. Regarding data produced by Fediverse users, see ”Broader impacts and ethical considerations” section below.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR)? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? We did not use crowdsourcing. Our data and study was determined not to be human subjects research.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Our study was determined not to be human subjects research.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? This is not applicable to our study.

(d) Did you discuss how data is stored, shared, and deidentified? This is not applicable to our study. Regarding related issues, see ”Broader impacts and ethical considerations” section below.

Broader impacts and ethical considerations

We are sensitive to the potential effects of surveillance on our study population. We made sure to design our data collection to only use public APIs and only collect posts explicitly marked as public[5]. When considering the trade-offs between the potential harms of collecting these data compared to their potential benefits, we believe that the potential for this research to inform the design and effectiveness of decentralized social networks directly benefits the population of interest.

Because Fediverse servers are typically run by volunteers or small organizations, it is important that our data collection procedure not overburden them. Therefore, we complied with rate limits indicated in HTML headers and sent no more than one request per second when the applicable headers were missing. Although the data we collected is publicly available in that it can be obtained via public APIs and HTTPS requests, Fediverse servers are also relatively obscure and users may expect a degree of privacy even regarding technically public posts and metadata. Therefore, we took steps to protect the integrity of the data we collected. We kept data on secure servers on our institutions' academic networks and only collected and stored posts marked as “public”, the highest level of visibility. Similarly, when we used the Perspective API to analyze posts, we ensured that no data were retained on the Perspective API servers.

[5] Mastodon and Pleroma offer other levels of visibility such as “unlisted” and “private”