Federating the Social Web

Carl Colglazier

2025-09-01

Decentralized online social networks (DOSNs) represent a new way of organizing online communities and present both challenges and opportunities. The decentralized design of spaces like the Fediverse provide more indepdence, but this increased autonomy comes at the expense of collective action problems which must be solved in new ways. This work considers how the collective actions problems are addressed in practice and designs potential new ways to solve them.

Table of contents

1 Introduction **5**

2 Background **7**

2.1 The Emergence of the Fediverse . . . 7
2.1.1 Interoperability . . . 7
2.1.2 Early Examples of Federated Social Websites . . . 7
2.1.3 ActivityPub . . . 8
2.1.4 Mastodon . . . 8
2.2 Challenges for Online Communities . . . 9
2.3 Collective Action Problems . . . 9
2.4 Content Moderation . . . 10
2.5 Discovery . . . 10

3 Newcomers **12**

3.1 Introduction . . . 12
3.2 Background . . . 13
3.2.1 Empirical Setting . . . 13
3.2.2 The Mastodon Migrations . . . 13
3.2.3 Newcomers in Online Communities . . . 14
3.2.4 Recommendation Systems and Collaborative Filtering . . . 14
3.2.5 Evaluation of Recommendation Systems . . . 16
3.3 Data . . . 16
3.3.1 Initial Findings . . . 16

4 Rules **20**

4.1 Introduction . . . 20
4.2 Background . . . 20
4.2.1 Institutional Isomorphism . . . 20
4.2.2 Empirical Setting . . . 21
4.2.3 Content Moderation . . . 21
4.2.4 Rules in Online Communities . . . 21
4.3 Data . . . 22

5 Defederation Study **23**

5.1 Introduction . . . 23

5.2 Background 24
5.2.1 Social Media Content Moderation and Governance 24
5.2.2 Governance and Moderation in Federated Communication Networks . . . 26
5.2.3 The Effects of Community-Level Defederation as a Group Sanction 27
5.3 Data, Measures, and Methods 28
5.3.1 Empirical Setting 29
5.3.2 Data Collection 29
5.3.3 Measures 30
5.4 Analytic Plan 31
5.4.1 Constructing Matched Control Groups 31
5.5 Analysis 32
5.6 Results 34
5.6.1 Effects on Activity 35
5.6.2 Effects on Anti-Social Behavior 36
5.7 Discussion 39
5.8 Conclusion 42
5.9 Acknowledgements 42

6 References **43**

1 Introduction

The World Wide Web has reshaped our communication landscape with pervasive economic and cultural effects (Litan and Rivlin 2001; Crandall, Lehr, and Litan 2007; Najarzadeh, Rahimzadeh, and Reed 2014; P. DiMaggio et al. 2001). The Internet, which forms the technological underpinning for the Web, solved the problem of connecting disparate networks, computers, and operators around the world by creating a flexible system based around interoperability: protocols like TCP/IP and HTTP allowed enough standardization for meaningful communication while allowing individual systems to still experiment and retain control over their data, access, and code. One key to its success is that the largely decentralized nature of the TCP/IP protocol allowed new nodes to join the network with limited friction (Campbell-Kelly and Garcia-Swartz 2013, 28).

This approach has facilitated a strong network of person-to-person communication, developing various systems to facilitate communication between their stakeholders. Early Internet developers used email to coordinate their efforts (Leiner et al. 2009, 24, 25). Later, early online communities developed on Bulletin Board Systems, USENET, and discussion forums (Rafaeli 1984; Hauben and Hauben 1997). The *blogosphere* of the 2000s became a significant media power in its own right and many alumni from this era successfully transitioned into roles in traditional mass media, which itself adopted many of the norms and practices from the bloggers (Drezner and Farrell 2008). Further, the rise of social networking sites brought online communication to a larger audience (see Figure 1.1) and marked a shift in the organization of communities from being around topics to being around people (boyd and Ellison 2007, 219).

The history of the design and use of these systems each reflect the needs and values of the people who built and use them. Early email, for instance, was closed off to academics and government employees who quite literally met in person to plan the early email programs (Partridge 2008). As operators could easily be mapped to their real name and identities, there was less need for security and privacy protections when the protocols were first developed.

Online communities have not come without problems; connecting so much together creates a number of benefits, but it also comes with challenges. Significant resources are spent by the companies who run them to try to keep them free from spam, harassment, and illegal content (Gillespie 2018). Differences in international law and norms can make this a challenge and norms can vary by culture and location (Kaye 2019). The largest platforms have becoming increasingly closed off and less transparent to researchers (Freelon 2018). Concerns persist over the dominant economic model for the commercialized Web, which relies heavily on advertising and attention (Davenport and Beck 2001).

```
Error in element_text(hjust = 1): could not find function "element_text"
```

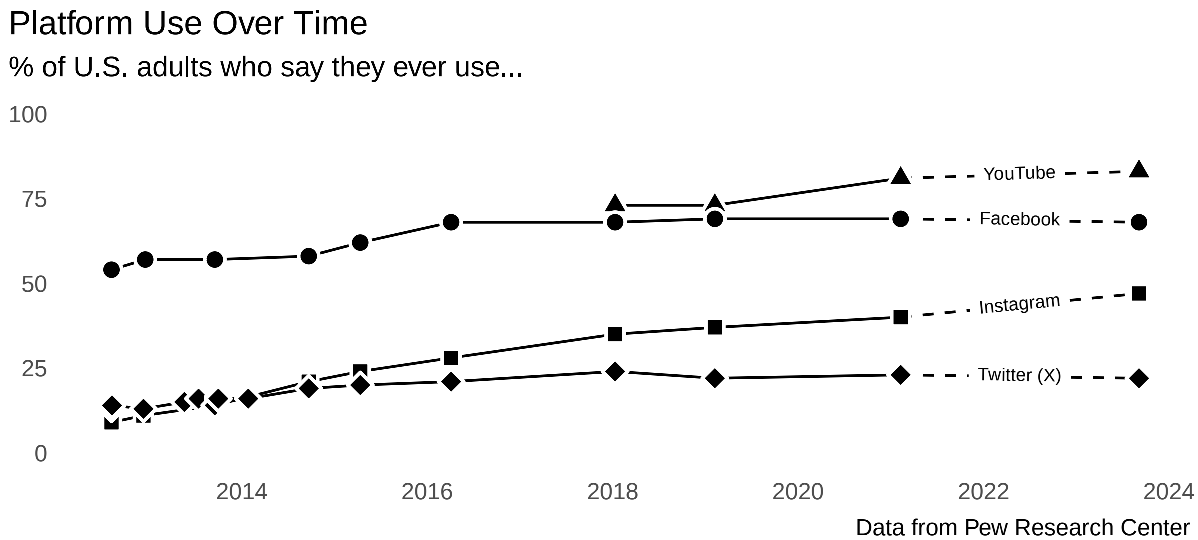

Figure 1.1: **Reported use of a selection of major social media platforms by adults in the U.S. over time according the surveys conducted by the Pew Research Center (PRC).** Note that in 2021 and before, PRC administered the survey over the phone, while after 2021 they administered the survey online and via mail.

While the commercial Web has received significant media and lawmaker focus, it does not represent the totality of Web. Much of the early online communities were run by hobbyists and non-profit organizations (Driscoll 2022). Projects like IndieWeb have sought to create a more decentralized Web, where people own their own data and can interact with others on their own terms (Jamieson, Yamashita, and McEwen 2022). Today's Fediverse is an extension of the original spirit of the Web: powered by the ActivityPub standard, the Fediverse is a collection of interoperable and interconnected websites that retain their independence and autonomy. With millions of active accounts, the Fediverse has established itself as a viable alternative to the commercial Web.

While significant work has gone into technical interoperability on the Fediverse (e.g. how do we pass messages between servers?), work on social interoperability is still emerging (e.g. how do we determine which servers we want to get messages from?). Fediverse servers must handle many of the challenges faced by the commercial Web in addition to some of the new challenges imposed by their decentralized design. This prospectus outlines a research agenda to understand how the Fediverse handles these challenges.

2 Background

2.1 The Emergence of the Fediverse

2.1.1 Interoperability

A system supports interoperability when information can be exchanged between parts of the system. Interoperability can have tremendous benefits because it guarantees parts of the system can work together while at the same time supporting components that are developed or operated independently. For example, the Internet is a highly interoperable system because it allows computers and networks to communicate with each other using shared protocols. Similarly, Email allows servers and accounts to pass messages to each other using shared protocols, but run their own software of choice[1].

What makes federated systems different from simple linking in practice is how they handle data. Individual nodes in a federated system do not simply link to data from other nodes in practice, but often store and replicate a copy of the data. In practice, this means such systems are less vulnerable to censorship, but it also introduces new complications regarding privacy. Deleting data from such a system is non-trivial.

2.1.2 Early Examples of Federated Social Websites

The Fediverse's cultural roots are in the free software movement, which emphasizes permissive licensing and open source software. Early Fediverse projects largely attempted to create *libre* alternatives to corporate social media platforms (Mansoux and Abbing 2020, 125). For example, the GNU Social project (formerly known as StatusNet) was created as a free software approach to microblogging. Similarly, the Diaspora project launched in 2010, inspired by its founders' shared concerns over the consolidation of information on the cloud (Nussbaum 2010). The network claimed over two hundred thousand users by November 2011 (Bielenberg et al. 2012) and was designed to be decentralized, with data stored and managed on independent servers known as "pods".

[1] 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 [tab:did.activity.coefs], the 95% credible intervals for both coefficients contain 0, suggesting the assumption is reasonable.

More recently, the IndieWeb created a method for interlinking websites using shared standards such as Microformats 2 to mark semantic data (Jamieson, Yamashita, and McEwen 2022). Culturally, the IndieWeb often encouraged people post on their own websites and then syndicate to other websites using a system called POSSE (Post on your Own Site, Syndicate Elsewhere). Protocols like WebMentions allow IndieWeb sites to interact with each other.

2.1.3 ActivityPub

The ActivityPub (AP) protocol's story starts from a place of fragmentation. While projects like GNU Social and StatusNet had a small but dedicated following, it was difficult to pass messages between servers with incompatible protocols like OStatus and Pump.io (Göndör and Küpper 2017). The ActivityPub project sought to bridge the collective of early Fediverse projects under a unifying standard and was recommended by the World Wide Web Consortium in 2018.

Although the protocols underlying the internet remains invisible for most of its users, a tremendous amount of time and effort go into their development. The development of a protocol can be challenging because it is impossible to anticipate all possible use-cases and there are trade-offs to most design decisions. ActivityPub, for instance, has been criticized for its lack of optimization and vulnerability to accidental distributed-denial-of-service attacks (Das 2024). It is important to remember that the ActivityPub standard represents the specific needs of their stakeholders at the time it was designed and adopted: namely, bridging the prior Fediverse protocols under a single, unified standard.

Fragmentation remains a challenge for DOSNs. It is hard to balance keeping wide compatibility with the creation of new and innovative features. While ActivityPub has been largely successful at bridging the fledgling communities it was designed for, further development on DOSNs may well leave AP behind. For instance, Bluesky has opted to produce its own protocol instead of adopting AP citing issues with data portability and scalability ("AT Protocol FAQ" n.d.).

2.1.4 Mastodon

Mastodon represents the most important Fediverse project to date. But despite its millions of registered users and coverage in major media publications, the project had humble beginnings. Eugen Rochko released the first public edition of Mastodon in October 2016. In a comment on the Hacker News thread on launch, Rochko wrote about the project's ethos: "This isn't a startup, it's an open-source project. Most likely the Twitters and Facebooks will win, but people should have a viable choice… Plus this is an incredibly fun project to be working on, to be quite honest" (Rochko 2016). The software soon found an audience and eclipsed the user base of other Fediverse software.

Early reporting on Mastodon often described it an alternative to other platforms like Twitter, a framing which Zulli, Liu, and Gehl (2020) criticized.

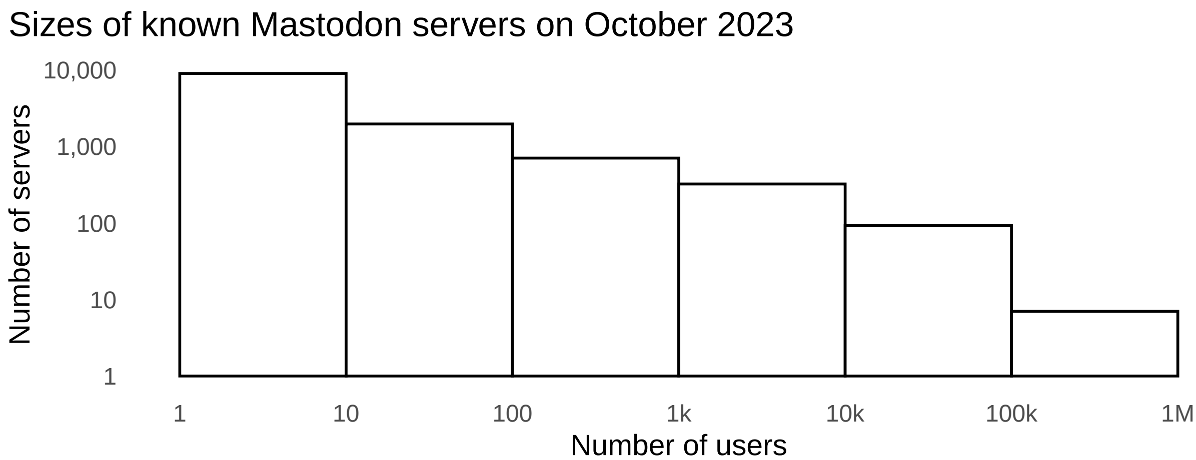

Figure 2.1: **Histogram of Mastodon servers by number of users on October 2023.** Note that the bins and y-axis both use a log 10 scale. The majority of Mastodon servers are small.

Most Mastodon servers are small. The vast majority of Mastodon servers have fewer than 10 accounts. Many of these have only a single account. The distribution of accounts on servers, however, is highly skewed: the median server has 3 accounts, while the mean has 595 accounts.

2.2 Challenges for Online Communities

2.3 Collective Action Problems

Kollock (1998, 183) defines social dilemmas as situations where individually rational behavior leaves the collective worse off. These social dilemmas have a deficient equilibrium where there is an outcome that leaves everyone better off, but no individual incentive to move toward that outcome (Kollock 1998, 185).

Collective action problems are social dilemmas that occur when self-interested individuals have no incentive to work toward a public good (Olson 1965, 2). Hardin (1968) described one such collective action problem, where people individually overexploit shared resources, as the *tragedy of the commons* in a Malthusian argument against population growth. More recent scholarship, however, has shown that the tragedy of the commons is not an inevitable outcome from shared resources (Ostrom 1990). Instead, people can and do work together to manage shared resources.

Kollock (1999) argues that online communities produce public goods, often in the form of knowledge with a nearly limitless potential audience. Access to free and accurate information leaves everyone better off. This idea has been the foundation of knowledge production projects like Wikipedia and open source software projects like the Linux kernel, both of which function as public goods (non-excludable and non-rivalrous).

Due to the differences in the means of production within the Fediverse, many of the challenges faced by commercial websites are trasformed into collective action problems. These can create challenegs for the sytem as a whole. For instance, several major Fediverse servers have shut down over the years due to hosting costs. The vast majority of people with accounts on the Fediverse do not directly financially contribute to their servers—though many do.

2.4 Content Moderation

All websites which rely on third-party, user-generated content must perform some form of content moderation to remain viable (Gillespie 2018). Without it, online spaces would become dominated by spam, pornography, or other unwanted content (Gillespie 2020, 330–31). Much of this work remains largely invisible by design (Roberts 2019, 14).

While all social websites must perform content moderation, approaches vary. Large, well-resourced websites like Facebook hire teams of contractors which do the bulk of cleaning up their website. Other websites built around named subcommunities like Reddit hand off moderation duties to unpaid volunteer community members.

The small size of the average Mastodon server affects content moderation. Despite the small size of most Mastodon servers, the average Mastodon account is on a large server. Raman et al. (2019) found the top 5% of Mastodon servers host 90.6% of Mastodon accounts and send 94.8% of the posts. This means while the bulk of the moderation work concentrates on a few large servers, vulnerabilities and problems can come from a large set of smaller, less resourced servers.

Nicholson, Keegan, and Fiesler (2023) characterized the written rules on a number of Mastodon servers.

2.5 Discovery

Recommender systems help people filter information to find resources relevant to some need (Ricci, Roa, and Shapira 2022). The development of these systems as an area of formal study harkens back to information retrieval (e.g. Salton and McGill (1987)) and foundational works imagining the role of computing in human decision-making (e.g. Bush (1945)). Early work on these systems produced more effective ways of filtering and sorting documents in searches such as the probabilistic models that motivated the creation of the okapi (BM25) relevance function (Robertson and Zaragoza 2009). Many contemporary recommendation systems use collaborative filtering, a technique which produces new recommendations for items based on the preferences of a collection of similar users (Koren, Rendle, and Bell 2022).

Collaborative filtering systems build on top of a user-item-rating ($U - I - r$) model where there is a set of users who each provide ratings for a set of items. The system then uses the ratings from other users to predict the ratings of a user for an item they have not yet rated and uses these predictions to create a ordered list of the best recommendations for the user's needs (Ekstrand, Riedl, and Konstan 2011, 86–87). Collaborative filtering recommender systems typically produce better results as the number of users and items in the system increases; however, they must also deal with the "cold start" problem, where limited data makes recommendations unviable (Lam et al. 2008). The cold start problem has three possible facets: boostrapping new communities, dealing with new items, and handling new users (Schafer et al. 2007, 311–12). In each case, limited data on the entity makes it impossible to find similar entities without some way of building a profile. Further, uncorrected collaborative filtering techniques often also produce a bias where more broadly popular items receive more recommendations than more obscure but possibly more relevant items (Zhu et al. 2021). Research on collaborative filtering has also shown that the quality of recommendations can be improved by using a combination of user-based and item-based collaborative filtering (Sarwar et al. 2001).

Although all forms of collaborative filtering use some combination of users and items, there are two main approaches to collaborative filtering: memory-based and model-based. Memory-based approaches use the entire user-item matrix to make recommendations, while model-based approaches use a reduced form of the matrix to make recommendations. This is particularly useful because the matrix of items and users tends to be extremely sparse, e.g. in a movie recommendor system, most people have not seen most of the movies in the database. Singular value decomposition (SVD) is one such dimension reduction technique which transforms a $m \times n$ matrix M into the form $M = U\Sigma V^T$ (Paterek 2007). SVD is particularly useful for recommendation systems because it can be used to find the latent factors which underlie the user-item matrix and use these factors to make recommendations.

While researchers in the recommendation system space often focus on ways to design the system to produce good results mathematically, human-computer interaction researchers also consider various human factors which contribute to the overall system. Crucially, McNee et al. argued "being accurate is not enough": user-centric evaluations, which consider multiple aspects of the user experience, are necessary to evaluate the full system. HCI researchers have also contributed pioneering recommender systems in practice. For example, GroupLens researchers Resnick et al. (1994) created a collaborative filtering systems for Usenet and later produced advancements toward system evaluation and explaination of movie recommendations (Herlocker et al. 2004; Herlocker, Konstan, and Riedl 2000). Cosley et al. (2007) created a system to match people with tasks on Wikipedia to encourage more editing. This prior work shows that recommender systems can be used to help users find relevant information in a variety of contexts.

Mastodon and other decentralized online social networks are particularily vulnerable to discovery problems. As information and accounts are spread out across many different servers, location matters in a way that is not relevant on centralized social networks. At the same time, any recommendation system run on a particular server is limited to the information on that server unless some system is in place to spread recommendations across servers, e.g. using federated machine learning.

3 Newcomers

3.1 Introduction

Following Twitter's 2022 acquisition, Mastodon—an open-source, decentralized social network and microblogging community—saw an increase in activity and attention as a potential Twitter alternative (He et al. 2023; La Cava, Aiello, and Tagarelli 2023). While millions of people set up new accounts and significantly increased the size of the network, many newcomers found the process confusing and many accounts did not remain active. Unlike centralized social media platforms, Mastodon is a network of independent servers with their own rules and norms (Nicholson, Keegan, and Fiesler 2023). Each server can communicate with each other using the shared ActivityPub protocols and accounts can move between Mastodon servers, but the local experience can vary widely from server to server.

Although attracting and retaining newcomers is a key challenge for online communities (Kraut, Resnick, and Kiesler 2011, 182), Mastodon's onboarding process has not always been straightforward. Variation among servers can also present a challenge for newcomers who may not even be aware of the specific rules, norms, or general topics of interest on the server they are joining (Diaz 2022). Various guides and resources for people trying to join Mastodon offered mixed advice on choosing a server. Some suggest that the most important thing is to simply join any server and work from there (Krasnoff 2022; Silberling 2023), while others have created tools and guides to help people find potential servers of interest by size and location(**thekinrarMastodonInstances2017?**; J.-M. King 2024).

Mastodon's decentralized design has long been in tension with the disproportionate popularity of a small set of large, general-topic servers within the system (Raman et al. 2019). Analysing the activity of new accounts that join the network, we find that users who sign up on such servers are less likely to remain active after 91 days. We also find that many users who move accounts tend to gravitate toward smaller, more niche servers over time, suggesting that established users may also find additional utility from such servers.

In response to these findings, we propose a potential way to create server and tag recommendations on Mastodon. This recommendation system could both help newcomers find servers that match their interests and help established accounts discover "neighborhoods" of related servers to enable further discovery.

3.2 Background

3.2.1 Empirical Setting

The Fediverse is a set of decentralized online social networks which interoperate using shared protocols like ActivityPub. Mastodon is a software program used by many Fediverse servers and offers a user experience similar to the Tweetdeck client for Twitter. It was first created in late 2016 and saw a surge in interest in 2022 during and after Elon Musk's Twitter acquisition.

Mastodon features three kinds of timelines. The primary timeline is a "home" timeline which shows all posts from accounts followed by the user. Mastodon also supports a "local" timeline which shows all public posts from the local server and a "federated" timeline which includes all posts from users followed by other users on their server. The local timeline is unique to each server and can be used to discover new accounts and posts from the local community. On larger servers, this timeline can be unwieldy; however, on smaller servers, this presents the opportunity to discover new posts and users of potential interest.

Discovery has been challenging on Mastodon. Text search, for instance, was impossible on most servers until support for this feature was added on an optional, opt-in basis using Elasticsearch in late 2023 (Rochko 2023b). Recommendation systems are currently a somewhat novel problem in the context of decentralized online social networks. Trienes, Cano, and Hiemstra (2018) developed a recommendation system for finding new accounts to follow on the Fediverse which used collaborative filtering based on BM25 in an early example of a content discovery system on Mastodon.

Individual Mastodon servers can have an effect on the end experience of users. For example, some servers may choose to federate with some servers but not others, altering the topology of the Fediverse network for their users. At the same time, accounts can only map to one specific server. Because of Mastodon's data portability, users can move their accounts freely between servers while retaining their followers, though their post history remains with their original account.

3.2.2 The Mastodon Migrations

Mastodon saw a surge in interest in 2022 and 2023, particularly after Elon Musk's Twitter acquisition. In particular, four events of interests drove measurable increases in new users to the network: the announcement of the acquisition (April 14, 2022), the closing of the acquisition (October 27, 2022), a day when Twitter suspended a number of prominent journalists (December 15, 2022), and a day when Twitter experienced an outage and started rate limiting accounts (July 1, 2023). Many Twitter accounts announced they were setting up Mastodon accounts and linked their new accounts to their followers, often using tags like `#TwitterMigration`(He et al. 2023) and driving interest in Mastodon in a process La Cava, Aiello, and Tagarelli (2023) found consistent with social influence theory.

Some media outlets have framed reports on Mastodon (Hoover 2023) through what Zulli, Liu, and Gehl (2020) calls the "Killer Hype Cycle", whereby the media finds a new alternative social media platform, declares it a potential killer of some established platform, and later calls it a failure if it does not displace the existing platform. Such framing fails to take systems like the Fediverse seriously for their own merits: completely replacing existing commercial systems is not the only way to measure success, nor does it account for the real value the Fediverse provides for its millions of active users.

Mastodon's approach to onboarding has also changed over time. In much of 2020 and early 2021, the Mastodon developers closed sign-ups to their flagship server and linked to an alternative server, which saw increased sign-ups during this period. They also linked to a list of servers on the "Join Mastodon" webpage (Mastodon gGmbH n.d.), where all servers are pre-approved and follow the Mastodon Server Covenant which guarantees certain content moderation standards and data protections. Starting in 2023, the Mastodon developers shifted toward making the flagship server the default when people sign up on the official Mastodon Android and iOS apps (Rochko 2023a; Roth 2023).

3.2.3 Newcomers in Online Communities

Onboarding newcomers is an important part of the life cycle of online communities. Any community can expect a certain amount of turnover, and so it is important for the long-term health and longevity of the community to be able to bring in new members (Kraut, Resnick, and Kiesler 2011, 182). However, the process of onboarding newcomers is not always straightforward.

The series of migrations of new users into Mastodon in many ways reflect folk stories of "Eternal Septembers" on previous communication networks, where a large influx of newcomers challenged the existing norms (Driscoll 2023; Kiene, Monroy-Hernández, and Hill 2016). Many Mastodon servers do have specific norms which people coming from Twitter may find confusing, such as local norms around content warnings (Nicholson, Keegan, and Fiesler 2023). Variation among servers can also present a challenge for newcomers who may not even be aware of the specific rules, norms, or general topics of interest on the server they are joining (Diaz 2022). Mastodon servers open to new accounts must thus be both accommodating to newcomers while at the same ensuring the propagation of their norms and culture, either through social norms or through technical means.

3.2.4 Recommendation Systems and Collaborative Filtering

Recommender systems help people filter information to find resources relevant to some need (Ricci, Roa, and Shapira 2022). The development of these systems as an area of formal study harkens back to information retrieval (e.g. Salton and McGill (1987)) and foundational works imagining the role of computing in human decision-making (e.g. Bush (1945)). Early work on these systems produced more effective ways of filtering and sorting documents in searches such as the probabilistic models that motivated the creation of the okapi (BM25) relevance function (Robertson and

Zaragoza 2009). Many contemporary recommendation systems use collaborative filtering, a technique which produces new recommendations for items based on the preferences of a collection of similar users (Koren, Rendle, and Bell 2022).

Collaborative filtering systems build on top of a user-item-rating ($U-I-r$) model where there is a set of users who each provide ratings for a set of items. The system then uses the ratings from other users to predict the ratings of a user for an item they have not yet rated and uses these predictions to create a ordered list of the best recommendations for the user's needs (Ekstrand, Riedl, and Konstan 2011, 86–87). Collaborative filtering recommender systems typically produce better results as the number of users and items in the system increases; however, they must also deal with the "cold start" problem, where limited data makes recommendations unviable (Lam et al. 2008). The cold start problem has three possible facets: boostrapping new communities, dealing with new items, and handling new users (Schafer et al. 2007, 311–12). In each case, limited data on the entity makes it impossible to find similar entities without some way of building a profile. Further, uncorrected collaborative filtering techniques often also produce a bias where more broadly popular items receive more recommendations than more obscure but possibly more relevant items (Zhu et al. 2021). Research on collaborative filtering has also shown that the quality of recommendations can be improved by using a combination of user-based and item-based collaborative filtering (Sarwar et al. 2001).

Although all forms of collaborative filtering use some combination of users and items, there are two main approaches to collaborative filtering: memory-based and model-based. Memory-based approaches use the entire user-item matrix to make recommendations, while model-based approaches use a reduced form of the matrix to make recommendations. This is particularly useful because the matrix of items and users tends to be extremely sparse, e.g. in a movie recommendor system, most people have not seen most of the movies in the database. Singular value decomposition (SVD) is one such dimension reduction technique which transforms a $m \times n$ matrix M into the form $M = U\Sigma V^T$ (Paterek 2007). SVD is particularly useful for recommendation systems because it can be used to find the latent factors which underlie the user-item matrix and use these factors to make recommendations.

While researchers in the recommendation system space often focus on ways to design the system to produce good results mathematically, human-computer interaction researchers also consider various human factors which contribute to the overall system. Crucially, McNee et al. argued "being accurate is not enough": user-centric evaluations, which consider multiple aspects of the user experience, are necessary to evaluate the full system. HCI researchers have also contributed pioneering recommender systems in practice. For example, GroupLens researchers Resnick et al. (1994) created a collaborative filtering systems for Usenet and later produced advancements toward system evaluation and explaination of movie recommendations (Herlocker et al. 2004; Herlocker, Konstan, and Riedl 2000). Cosley et al. (2007) created a system to match people with tasks on Wikipedia to encourage more editing. This prior work shows that recommender systems can be used to help users find relevant information in a variety of contexts.

3.2.5 Evaluation of Recommendation Systems

Evaluating recommender systems can be tricky because a measure of good performance must take into account various dimensions(Zangerle and Bauer 2022). A measure of accuracy must be paired with a question of "accuracy toward what?" Explainability requires a transparent means of showing the user why a certain item was recommended.

It is often important to both start with an end goal in mind and to keep evaluation integrated throughout the entire process of creating a recommender systems, from conceptualization to optimization. There are several considerations to keep in mind such as the trade-off between optimizing suggestions and the risks of over-fitting. For example, a system designed to create suggestions with the highest propensity that the user will like the recommendations may struggle with a reduced diversity of its suggestions.

Recommender systems can be evaluated using three board categories of techniques: offline, online, and user studies. Offline evaluation uses pre-collected data and a measure to describe the performance of the system, assuming there is insufficient relevance to the difference in time between when the data was collected and the present moment. Online evaluation uses a deployed, live system, e.g. A/B testing. In this case, the user is often unaware of the experiment. In contrast, user studies involve subjects which are aware they are being studied.

3.3 Data

Mastodon has an extensive API which allows for the collection of public posts and account information. We collected data from the public timelines of Mastodon servers using the Mastodon API with a crawler which runs once per day. We also collected account information from the opt-in public profile directories on these servers.

3.3.1 Initial Findings

3.3.1.1 Account Survival

Our initial findings suggest that servers do matter for newcomers on Mastodon. Figure 3.2 uses a Kaplan–Meier estimator to show that accounts on the largest Mastodon servers featured on the Join Mastodon website are less likely to remain active compared to accounts on smaller servesr featured on Join Mastodon. Further, Table 3.1 uses a Cox Proportional Hazard Model with mixed effects to suggest that small servers are significantly better at retaining new accounts and that general servers are less likely to retain new accounts; being featured on the Join Mastodon website appears to have no significant effect.

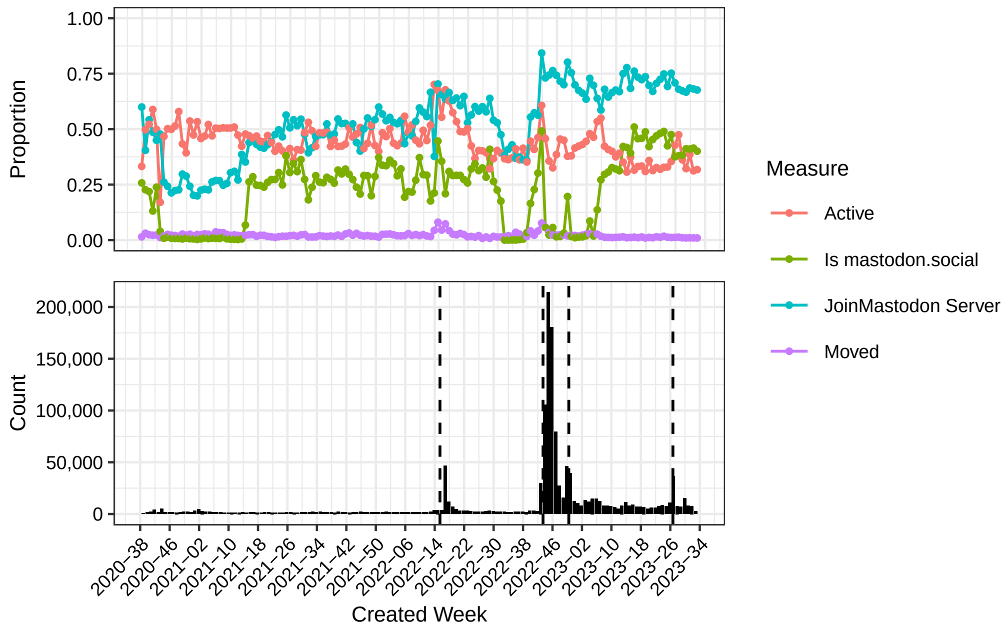

Figure 3.1: Accounts in the dataset created between January 2022 and March 2023. The top panels shows the proportion of accounts still active 45 days after creation, the proportion of accounts that have moved, and the proportion of accounts that have been suspended. The bottom panel shows the count of accounts created each week. The dashed vertical lines in the bottom panel represent the annoucement day of the Elon Musk Twitter acquisition, the acquisition closing day, a day where Twitter suspended a number of prominent journalist, and a day when Twitter experienced an outage and started rate limiting accounts.

Table 3.1: **Coefficients for the Cox Proportional Hazard Model with Mixed Effects.** The model includes a random effect for the server.

Term	Estimate	Low	High	p-value
Join Mastodon	0.115	0.972	1.296	0.117
General Servers	0.385	1.071	2.015	0.017
Small Server	−0.245	0.664	0.922	0.003

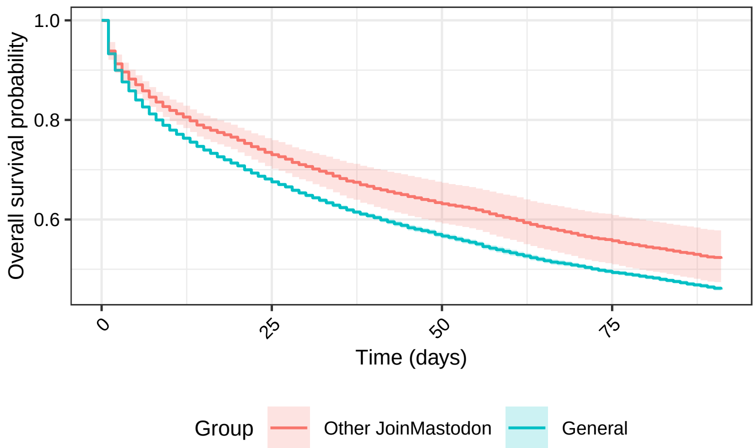

Figure 3.2: **Survival probabilities for accounts created during May 2023.** Using N accounts created from May 1 to June 30, 2023, we create a Kaplan–Meier estimator for the probability that an account will remain active based on whether the account is on one of the largest general instances featured at the top of the Join Mastodon webpage or otherwise if it is on a server in the Join Mastodon list. Accounts are considered active if they have made at least one post after the censorship period M days after account creation.

3.3.1.2 Moved Accounts

To corroborate these findings, we also looked at accounts that moved from one server to another. We find that accounts are more likely to move from larger servers to smaller servers.

Table 3.2: **Exponential family random graph models for account movement between Mastodon servers.** Accounts in Model A were created in May 2022 and moved to another account at some later point. Accounts in Model B were created at some earlier point and moved after October 2023.

	Model A		Model B	
	Coef.	Std.Error	Coef.	Std.Error
(Sum)	−9.529	***0.188	−10.268	***0.718
nonzero	−3.577	***0.083	−2.861	***0.254
Smaller server	0.709	***0.032	0.629	***0.082
Server size (outgoing)	0.686	***0.013	0.655	***0.042
Open registrations (incoming)	0.168	***0.046	−0.250	0.186
Languages match	0.044	0.065	0.589	0.392

4 Rules

4.1 Introduction

While the Fediverse promises more autonomy in social media operations, most servers on the Fediverse have actually adopted very similar rules, even directly copying them in many cases. Why, despite the opportunity to innovate, do Fediverse servers tend to adopt the same rules?

We consider this puzzle through the persepctive of institutional isomorphism. Through interviews with 17 server moderators and administrators combined with longitudinal records from thousands of servers, we find evidence of isomorphism between Fediverse servers and describe some of the mechanisms through which this occurs.

Our study suggests that rules on the Fediverse perform an isometric role in addition to a practical role: as a signal to other servers and to people on the server. Rules encode not only what behaviors are permissible, but also community values.

4.2 Background

4.2.1 Institutional Isomorphism

Meyer and Rowan (1977) consider rules as myths which insitutions incorporate to bolster their own legitimatacy, sometimes at the expense of internal efficiency. Under this neo-institutionalist view, individual actions are constrained by societal expectations, which in turn lead to the development of formal and informal rules within insitutions.

P. J. DiMaggio and Powell (1983) consider the problem of why orginizations tend to look similar to each other and theorize that orginizations tend to adopt similar norms through a process of insitutional isomorphism, which has three types: normative, coersive, and mimetic. Normative isomorphic change is a byproduct of professionalization. Coersive isomporpohic change comes from pressures from other organizations. Mimetic isomorphism describes how orgainizations tend to imitate each other.

4.2.2 Empirical Setting

The Fediverse is a set of independently operated servers which send data with the ActivityPub protocol. The most popular servers on the Fediverse use Mastodon, though alternatives like Pleroma and Misskey are also common. servers can pass messages to each other regardless of the actual software being used.

As the servers are autonomous, content moderation is handled differently by administrators and moderators on each server.

4.2.3 Content Moderation

Significant scholarship has investigated content moderation in the context of commercial, centralized social media platforms (e.g. Gillespie 2018; Roberts 2019). The function of content moderation in this context is shaped by economic factors: platforms often use contract labor and the people making case-by-case decisions are often divorced from the context of the communities they moderate (Gray and Suri 2019). However, a significant amount of the Web also relies on volunteer moderation: service work from unpaid people who are usually active members of their own communities. These volunteers might have different motivations for their actions.

4.2.3.1 Content Moderation on the Fediverse

Due to its technical design, the Fediverse faces unique content moderation challenges. Administrators may be responsible not only for posts on their own servers, but they may also have to deal with posts and conflicts from external servers.

Like the servers themselves, the costs of content moderation is distributed among many servers.

Servers must also consider their relationship to other servers from which they may receive posts. Hassan et al. (2021) explored the use of server-to-server federation policies and suggested that some server-level blocks may have collatoral damage on some users. Colglazier, TeBlunthuis, and Shaw (2024) measured the effects of server-to-server de-federation events and found asymmetric effects on activity for affected accounts on the blocked servers.

4.2.4 Rules in Online Communities

Some commercial websites like Reddit allow volunteer moderators for sub-communities to create their own rules; however, moderators must also consider the site-wide rules (Fiesler et al. 2018). The Fediverse, by contrast, requires no base layer of rules for each community and account: there are no site-wide rules (Nicholson, Keegan, and Fiesler 2023). Rules on the Fediverse therefore do more than on sites like Reddit.

4.3 Data

My analysis of the use for and function of rules on the Fediverse uses data from interviews with N people on the Fediverse and trace data from the metadata on M Fediverse servers.

5 Defederation Study

5.1 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; Ribeiro et al. 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.

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.

5.2 Background

5.2.1 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 (G. 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 trade-off. 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; G. 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; Ribeiro et al. 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.

5.2.2 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.

5.2.3 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.

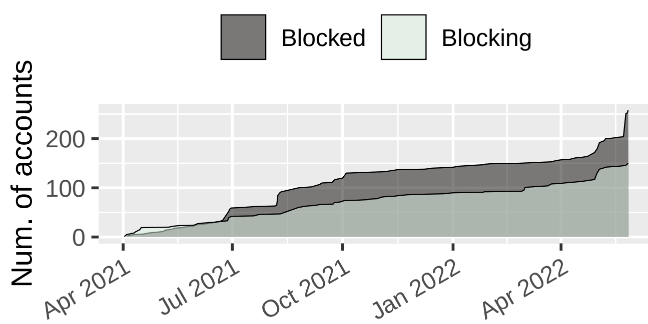

Figure 5.1: The y-axis shows the cumulative number of blocked and blocking accounts included in our analysis over our study period.

5.3 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 7445 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.

5.3.1 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.

5.3.2 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 peer connections to members of this set using an API feature which discloses server-to-server connections. This approach identified 11025 servers and 7445 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

[1] https://mastodon.social/

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.

5.3.3 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 1136 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" (Jigsaw 2021). 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.

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.

5.4 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 non-parametric 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).

5.4.1 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 Figure 5.2 shows the covariate balance using standardized mean differences before and after matching across these groups.

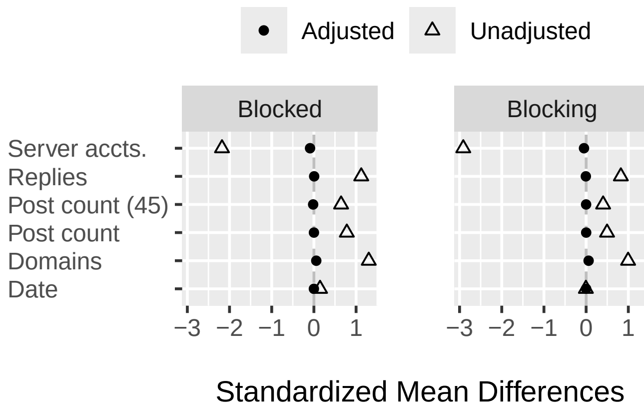

Figure 5.2: A covariate balance plot shows the standardized mean difference between treatment and control groups for each measure used in our matching procedure before (unadjusted) and after (adjusted) matching. Our procedure effectively found a group of matched controls similar to the treated accounts along these measures.

5.5 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).[2] 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).[3]

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

[3] 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 [tab:did.activity.coefs], the 95% credible intervals for both coefficients contain 0, suggesting the assumption is reasonable.

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 multi-level 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.

5.6 Results

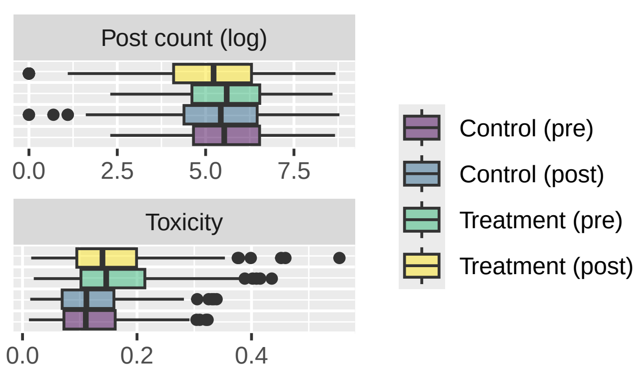

Figure 5.3: Box and whisker plots visualize the distributions of our dependent variables within the blocked and blocking groups of user accounts and their matched controls before and after defederation. The lines correspond to the median, the boxes to the inter-quartile range (IQR), the whiskers to the range of the data within 1.5 * IQR, and the dots to data points outside the range of the whiskers.

Figure Figure 5.3 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.

5.6.1 Effects on Activity

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

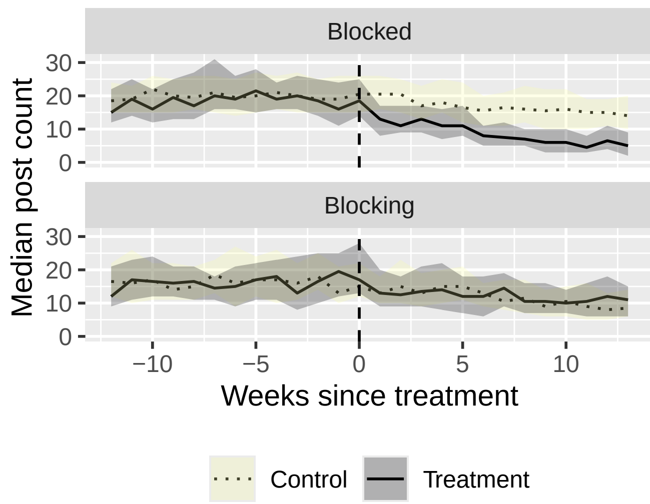

Figure 5.4: Visualization of activity among blocked and blocking user accounts shows an asymmetric change in activity following defederation. An account with a median post count on the blocked server declines in activity much more rapidly following defederation compared to matched controls while an account 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 Figure 5.4 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).

Table 5.1: 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.

Group	median	W	p
U_0	−135.5	41 197.5	0.000
C_0	−18.0	35 762.0	0.143
U_1	−54.5	12 413.0	0.122
C_1	−53.5	12 520.0	0.091
Δ_0	−39.0	39 927.0	0.000
Δ_1	3.0	10 645.5	0.421

5.6.1.1 Non-parametric tests

5.6.1.2 DiD Analysis

5.6.2 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 5.6 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 similar before defederation between the treatment and control groups for accounts on both the blocked and blocking servers.

5.6.2.1 Non-parametric tests

Table 5.3 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.

Table 5.2: 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.

	Blocked				Blocking			
Term	Estimate	Std. err.	Low	High	Estimate	Std. err.	Low	High
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

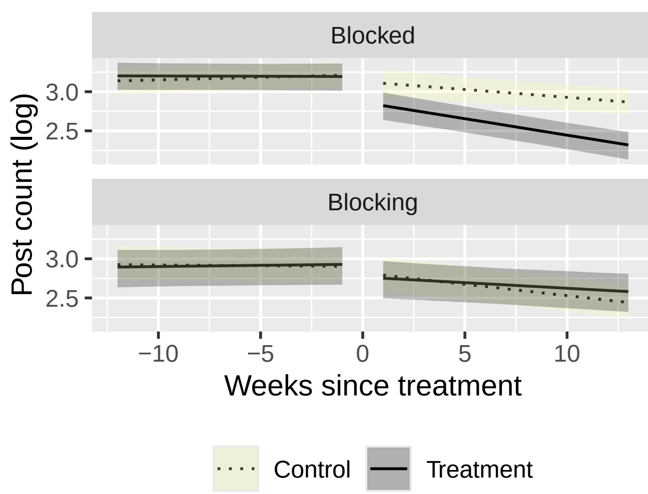

Figure 5.5: A marginal effects plot visualizes results from our difference-in-differences analysis of account activity by week. We observe a discontinuous activity decrease among accounts on the blocked server that exceeds any decrease in the matched controls, but no corresponding change among accounts on the blocking server. The bands are 95% credible intervals that account for uncertainty in parameter estimates.

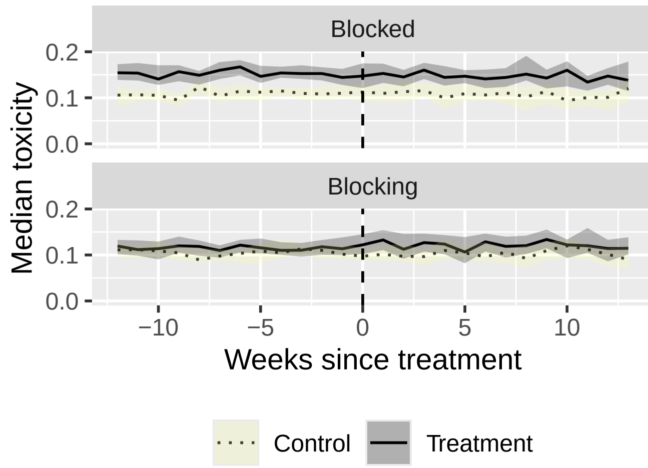

Figure 5.6: Median toxicity among accounts which posted each week for blocked and blocking user accounts. The median toxicity remained flat for all groups.

Table 5.3: Non-parametric difference-in-differences for median post toxicity before and after defederation 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.

Group	median	W	p
U_0	−0.006	17 746	0.538
C_0	0.004	14 000	0.950
U_1	−0.008	6514	0.619
C_1	0.001	5546	0.873
Δ_0	−0.005	17 161	0.072
Δ_1	0.000	6414	0.305

Table 5.4: 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.

```
# A tibble: 8 x 10
  selected  term     value.x    sd.x .lower.x .upper.x  value.y    sd.y .lower.y
  <chr>     <chr>      <dbl>   <dbl>    <dbl>    <dbl>    <dbl>   <dbl>    <dbl>
1 Treatment "$\\be~ -1.43e-1 4.38e-1 -1.20e+0  7.63e-1 -4.23e-2 2.84e-1 -0.754
2 Treatment "$\\be~  1.60e-3 6.35e-3 -9.11e-3  1.75e-2 -6.97e-4 9.99e-3 -0.0202
3 Treatment "$\\be~  8.89e-4 6.78e-4 -2.45e-4  2.28e-3  7.48e-4 9.97e-4 -0.00117
4 Treatment "$\\be~ -1.73e-3 1.13e-3 -4.00e-3  3.79e-4 -6.65e-4 1.64e-3 -0.00382
5 Control   "$\\be~  1.36e-1 4.36e-1 -9.07e-1  1.04e+0  4.76e-2 2.83e-1 -0.613
6 Control   "$\\be~ -4.55e-4 6.47e-3 -1.62e-2  1.12e-2  3.63e-3 9.03e-3 -0.0140
7 Control   "$\\be~  8.09e-4 6.94e-4 -3.72e-4  2.28e-3  3.29e-3 1.03e-3  0.00121
8 Control   "$\\be~ -4.97e-3 1.21e-3 -7.41e-3 -2.72e-3 -5.65e-3 1.67e-3 -0.00880
# i 1 more variable: .upper.y <dbl>
```

5.6.2.2 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 5.4. We find no evidence for any significant changes in post toxicity among accounts on either blocked or blocking servers.

5.7 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 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 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.

5.8 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.

5.9 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.

6 References

Albert, Réka, Hawoong Jeong, and Albert-László Barabási. 2000. "Error and Attack Tolerance of Complex Networks." *Nature* 406 (6794): 378–82. https://doi.org/10.1038/35019019.

Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. "Understanding the Effect of Deplatforming on Social Networks." In *Proceedings of the 13th ACM Web Science Conference 2021*, 187–95. WebSci '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3447535.3462637.

Anaobi, Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Damilola Ibosiola, and Gareth Tyson. 2023. "Will Admins Cope? Decentralized Moderation in the Fediverse." In *Proceedings of the ACM Web Conference 2023*, 3109–20. WWW '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3543507.3583487.

"AT Protocol FAQ." n.d. *AT Protocol*. https://atproto.com/guides/faq. Accessed December 4, 2024.

Bielenberg, Ames, Lara Helm, Anthony Gentilucci, Dan Stefanescu, and Honggang Zhang. 2012. "The Growth of Diaspora - A Decentralized Online Social Network in the Wild." In *2012 Proceedings IEEE INFOCOM Workshops*, 13–18. https://doi.org/10.1109/INFCOMW.2012.6193476.

boyd, danah m., and Nicole B. Ellison. 2007. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13 (1): 210–30. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Buntain, Cody, Martin Innes, Tamar Mitts, and Jacob Shapiro. 2023. "Cross-Platform Reactions to the Post-January 6 Deplatforming." *Journal of Quantitative Description: Digital Media* 3 (March). https://doi.org/10.51685/jqd.2023.004.

Bush, Vannevar. 1945. "As We May Think." *The Atlantic* 176 (1): 101–8.

Campbell-Kelly, Martin, and Daniel D Garcia-Swartz. 2013. "The History of the Internet: The Missing Narratives." *Journal of Information Technology* 28 (1): 18–33. https://doi.org/10.1057/jit.2013.4.

Chandrasekharan, Eshwar, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2022. "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit." *ACM Transactions on Computer-Human Interaction* 29 (4): 29:1–26. https://doi.org/10.1145/3490499.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech." *Proc. ACM Hum.-Comput. Interact.* 1 (CSCW): 31:1–22. https://doi.org/10.1145/3134666.

Colglazier, Carl, Nathan TeBlunthuis, and Aaron Shaw. 2024. "The Effects of Group Sanctions on Participation and Toxicity: Quasi-experimental Evidence from the Fediverse." *Proceedings* *of the International AAAI Conference on Web and Social Media* 18 (May): 315–28. https://doi.org/10.1609/icwsm.v18i1.31316.

Cosley, Dan, Dan Frankowski, Loren Terveen, and John Riedl. 2007. "SuggestBot: Using Intelligent Task Routing to Help People Find Work in Wikipedia." In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, 32–41. IUI '07. New York, NY, USA: ACM. https://doi.org/10.1145/1216295.1216309.

Crandall, Robert, William Lehr, and Robert Litan. 2007. "The Effects of Broadband Deployment on Output and Employment: A Cross-sectional Analysis of U.S. Data." 6. Brookings Institution.

Das, Ankush. 2024. "Please Don't Share Our Links on Mastodon: Here's Why!" *It's FOSS News*. https://news.itsfoss.com/mastodon-link-problem/.

Datta, Anwitaman, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe, and Krzysztof Rzadca. 2010. "Decentralized Online Social Networks." In *Handbook of Social Network Technologies and Applications*, edited by Borko Furht, 349–78. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4419-7142-5_17.

Davenport, Thomas H., and John C. Beck. 2001. *The Attention Economy: Understanding the New Currency of Business*. Boston, Mass: Harvard Business School Press.

Diaz, Jesus. 2022. "Using Mastodon Is Way Too Complicated to Ever Topple Twitter." *Fast Company*. https://www.fastcompany.com/90808984/using-mastodon-is-way-too-complicated-to-ever-topple-twitter.

DiMaggio, Paul J., and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48 (2): 147–60. https://doi.org/10.2307/2095101.

DiMaggio, Paul, Eszter Hargittai, W. Russell Neuman, and John P. Robinson. 2001. "Social Implications of the Internet." *Annual Review of Sociology* 27 (1): 307–36. https://doi.org/10.1146/annurev.soc.27.1.307.

Drezner, Daniel W., and Henry Farrell. 2008. "Introduction: Blogs, Politics and Power: A Special Issue of Public Choice." *Public Choice* 134 (1/2): 1–13. https://www.jstor.org/stable/27698207.

Driscoll, Kevin. 2022. *The Modem World: A Prehistory of Social Media*. Yale University Press.

———. 2023. "Do We Misremember Eternal September?" *FLOW*.

Ekstrand, Michael D., John T. Riedl, and Joseph A. Konstan. 2011. "Collaborative Filtering Recommender Systems." *Foundations and Trends® in Human–Computer Interaction* 4 (2): 81–173. https://doi.org/10.1561/1100000009.

Fiesler, Casey, Jialun" Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R. Brubaker. 2018. "Reddit Rules! Characterizing an Ecosystem of Governance." In *Proceedings of the International AAAI Conference on Web and Social Media*, 72–81. Stanford, CA: AAAI.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–68. https://doi.org/10.1080/10584609.2018.1477506.

Frey, Seth, and Nathan Schneider. 2021. "Effective Voice: Beyond Exit and Affect in Online Communities." *New Media & Society*, September, 14614448211044025. https://doi.org/10.1177/14614448211044025.

Frey, Seth, and Robert W. Sumner. 2019. "Emergence of Integrated Institutions in a Large Population of Self-Governing Communities." *PLOS ONE* 14 (7): e0216335. https://doi.org/10.

1371/journal.pone.0216335.

Fulk, Janet, Andrew J. Flanagin, Michael E. Kalman, Peter R. Monge, and Timothy Ryan. 1996. "Connective and Communal Public Goods in Interactive Communication Systems." *Communication Theory* 6 (1): 60–87. https://doi.org/10.1111/j.1468-2885.1996.tb00120.x.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. New Haven: Yale University Press.

———. 2020. "Platforms Throw Content Moderation at Every Problem." In *Fake News: Understanding Media and Misinformation in the Digital Age*. The MIT Press. https://doi.org/10.7551/mitpress/11807.001.0001.

Göndör, Sebastian, and Axel Küpper. 2017. "The Current State of Interoperability in Decentralized Online Social Networking Services." In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, 852–57. https://doi.org/10.1109/CSCI.2017.148.

Gray, Mary, and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston, MA: Houghton Mifflin Harcourt.

Gruzd, Anatoliy, Felipe Bonow Soares, and Philip Mai. 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. https://doi.org/10.1177/20563051231196878.

Han, Catherine, Deepak Kumar, and Zakir Durumeric. 2022. "On the Infrastructure Providers That Support Misinformation Websites." *Proceedings of the International AAAI Conference on Web and Social Media* 16 (May): 287–98. https://doi.org/10.1609/icwsm.v16i1.19292.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162 (3859): 1243–48. https://doi.org/10.1126/science.162.3859.1243.

Hassan, Anaobi Ishaku, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2021. "Exploring Content Moderation in the Decentralised Web: The Pleroma Case." In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, 328–35. CoNEXT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3485983.3494838.

Hauben, Michael, and Ronda Hauben. 1997. *Netizens: On the History and Impact of Usenet and the Internet*. Washington: IEEE computer.

He, Jiahui, Haris Bin Zia, Ignacio Castro, Aravindh Raman, Nishanth Sastry, and Gareth Tyson. 2023. "Flocking to Mastodon: Tracking the Great Twitter Migration." In *Proceedings of the 2023 ACM on Internet Measurement Conference*, 111–23. IMC '23. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3618257.3624819.

Heckathorn, Douglas D. 1988. "Collective Sanctions and the Creation of Prisoner's Dilemma Norms." *American Journal of Sociology* 94 (3): 535–62. https://doi.org/10.1086/229029.

Herlocker, Jonathan L., Joseph A. Konstan, and John Riedl. 2000. "Explaining Collaborative Filtering Recommendations." In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, 241–50. CSCW '00. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/358916.358995.

Herlocker, Jonathan L., Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. "Evaluating Collaborative Filtering Recommender Systems." *ACM Transactions on Information Systems* 22 (1): 5–53. https://doi.org/10.1145/963770.963772.

Hill, Benjamin Mako, and Aaron Shaw. 2021. "The Hidden Costs of Requiring Accounts: Quasi-experimental Evidence from Peer Production." *Communication Research* 48 (6): 771–95. https://doi.org/10.1177/0093650220910345.
Hoover, Amanda. 2023. "The Mastodon Bump Is Now a Slump." *Wired*, February.
Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20 (1): 1–24. https://doi.org/10.1093/pan/mpr013.
Jamieson, Jack, Naomi Yamashita, and Rhonda McEwen. 2022. "Bridging the Open Web and APIs: Alternative Social Media Alongside the Corporate Web." *Social Media + Society* 8 (1): 20563051221077032. https://doi.org/10.1177/20563051221077032.
Jhaver, Shagun, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 381:1–30. https://doi.org/10.1145/3479525.
Jigsaw. 2021. "Perspective API." https://www.perspectiveapi.com/.
Kaye, David. 2019. *Speech Police: The Global Struggle to Govern the Internet*. New York: Columbia Global Reports.
Kiene, Charles, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. "Surviving an 'Eternal September': How an Online Community Managed a Surge of Newcomers." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1152–56. New York, NY: ACM. https://doi.org/10.1145/2858036.2858356.
King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107 (2). https://doi.org/10.1017/S0003055413000014.
King, Jaz-Michael. 2024. "Mastodon Near Me." *Jaz-Michael King*.
Kollock, Peter. 1998. "Social Dilemmas: The Anatomy of Cooperation." *Annual Review of Sociology* 24 (January): 183–214. https://www.jstor.org/stable/223479.
———. 1999. "The Economies of Online Cooperation: Gifts and Public Goods in Cyberspace." In *Communities in Cyberspace*, edited by Marc Smith and Peter Kollock, 220–39. London, UK: Routledge.
Koren, Yehuda, Steffen Rendle, and Robert Bell. 2022. "Advances in Collaborative Filtering." In *Recommender Systems Handbook*, edited by Francesco Ricci, Lior Roa, and Bracha Shapira, Third edition, 91–142. New York, NY: Springer.
Krasnoff, Barbara. 2022. "Mastodon 101: How to Follow (and Unfollow) Other Accounts." *The Verge*. https://www.theverge.com/23519279/mastodon-instance-follow-friend.
Kraut, Robert E., Paul Resnick, and Sara Kiesler. 2011. *Building Successful Online Communities: Evidence-Based Social Design*. Cambridge, Mass: MIT Press.
La Cava, Lucio, Luca Maria Aiello, and Andrea Tagarelli. 2023. "Drivers of Social Influence in the Twitter Migration to Mastodon." *Scientific Reports* 13 (1): 21626. https://doi.org/10.1038/s41598-023-48200-7.
Lam, Xuan Nhat, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. "Addressing Cold-Start Problem in Recommendation Systems." In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication*, 208–11. ICUIMC '08. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1352793.1352837.

Leiner, Barry M., Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, and Stephen Wolff. 2009. "A Brief History of the Internet." *ACM SIGCOMM Computer Communication Review* 39 (5): 22–31. https://doi.org/10.1145/1629607.1629613.

Litan, Robert E., and Alice M. Rivlin. 2001. "Projecting the Economic Impact of the Internet." *The American Economic Review* 91 (2): 313–17. https://www.jstor.org/stable/2677780.

Mansoux, Aymeric, and Roel Rocsam Abbing. 2020. "Seven Theses on the Fediverse and the Becoming of Floss." In *The Eternal Network: The Ends and Becomings of Network Culture*, edited by Kristoffer Gansing and Inga Luchs. Institute of Network Cultures and transmediale e.V.

Mastodon gGmbH. n.d. "Servers." *Join Mastodon*. https://joinmastodon.org/servers. Accessed March 4, 2024.

Meyer, John W., and Brian Rowan. 1977. "Institutionalized Organizations: Formal Structure as Myth and Ceremony." *American Journal of Sociology* 83 (2): 340–63. https://www.jstor.org/stable/2778293.

Mitts, Tamar, Nilima Pisharody, and Jacob Shapiro. 2022. "Removal of Anti-Vaccine Content Impacts Social Media Discourse." In *14th ACM Web Science Conference 2022*, 319–26. Barcelona Spain: ACM. https://doi.org/10.1145/3501247.3531548.

Myers West, Sarah. 2018. "Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms." *New Media & Society* 20 (11): 4366–83. https://doi.org/10.1177/1461444818773059.

Najarzadeh, Reza, Farzad Rahimzadeh, and Michael Reed. 2014. "Does the Internet Increase Labor Productivity? Evidence from a Cross-Country Dynamic Panel." *Journal of Policy Modeling* 36 (6): 986–93. https://doi.org/10.1016/j.jpolmod.2014.10.003.

Nicholson, Matthew N., Brian C Keegan, and Casey Fiesler. 2023. "Mastodon Rules: Characterizing Formal Rules on Popular Mastodon Instances." In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, 86–90. CSCW '23 Companion. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3584931.3606970.

Nussbaum, Emily. 2010. "How Diaspora Is a Very Different Kind of Social Network." *New York Magazine*, September.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. The Political Economy of Institutions and Decisions. Cambridge ; New York: Cambridge University Press.

Partridge, Craig. 2008. "The Technical Development of Internet Email." *IEEE Annals of the History of Computing* 30 (2): 3–29. https://doi.org/10.1109/MAHC.2008.32.

Paterek, Arkadiusz. 2007. "Improving Regularized Singular Value Decomposition for Collaborative Filtering." *Proceedings of KDD Cup and Workshop*, August.

Rafaeli, Sheizaf. 1984. "The Electronic Bulletin Board: A Computer-Driven Mass Medium." *Social Science Micro Review* 2 (3): 123–36. https://doi.org/10.1177/089443938600200302.

Raman, Aravindh, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. "Challenges in the Decentralised Web: The Mastodon Case." In *Proceedings of the Internet*

Measurement Conference, 217–29. IMC '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3355369.3355572.
Resnick, Paul, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. "Grouplens: An Open Architecture for Collaborative Filtering of Netnews." In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175–86. CSCW '94. New York, NY, USA: ACM. https://doi.org/10.1145/192844.192905.
Ribeiro, Manoel Horta, Homa Hosseinmardi, Robert West, and Duncan J Watts. 2023. "Deplatforming Did Not Decrease Parler Users' Activity on Fringe Social Media." *PNAS Nexus* 2 (3): pgad035. https://doi.org/10.1093/pnasnexus/pgad035.
Ribeiro, Manoel Horta, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. "Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW2): 316:1–24. https://doi.org/10.1145/3476057.
Ricci, Francesco, Lior Roa, and Bracha Shapira, eds. 2022. *Recommender Systems Handbook*. Third edition. New York, NY: Springer.
Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven: Yale University Press.
Robertson, Stephen, and Hugo Zaragoza. 2009. "The Probabilistic Relevance Framework: BM25 and Beyond." *Foundations and Trends® in Information Retrieval* 3 (4): 333–89. https://doi.org/10.1561/1500000019.
Rochko, Eugen. 2016. "Show HN: A New Decentralized Microblogging Platform."
———. 2023a. "A New Onboarding Experience on Mastodon." *Mastodon Blog*. https://blog.joinmastodon.org/2023/0 new-onboarding-experience-on-mastodon/.
———. 2023b. "Mastodon 4.2." *Mastodon Blog*. https://blog.joinmastodon.org/2023/09/mastodon-4.2/.
Roth, Emma. 2023. "It's Getting Easier to Make an Account on Mastodon." *The Verge*. https://www.theverge.com/2023/5/1/23707019/mastodon-account-creation-twitter-alternative.
Russo, Giuseppe, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 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 (June): 742–53. https://doi.org/10.1609/icwsm.v17i1.22184.
Salton, Gerard, and Michael J. McGill. 1987. *Introduction to Modern Information Retrieval*. 3. pr. McGraw-Hill International Editions. New York: McGraw-Hill Book Comp.
Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl. 2001. "Item-Based Collaborative Filtering Recommendation Algorithms." In *Proceedings of the 10th International Conference on World Wide Web*, 285–95. WWW '01. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/371920.372071.
Schafer, J. Ben, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. "Collaborative Filtering Recommender Systems." In *The Adaptive Web: Methods and Strategies of Web Personalization*, edited by Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, 291–324. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-72079-9_9.
Seering, Joseph, Robert Kraut, and Laura Dabbish. 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*, 111–25. CSCW '17. New York, NY, USA: ACM. https://doi.org/10.1145/2998181.2998277.

Silberling, Amanda. 2023. "A Beginner's Guide to Mastodon, the Open Source Twitter Alternative | TechCrunch." *TechCrunch*. https://techcrunch.com/2023/07/24/what-is-mastodon/.

Srinivasan, Kumar Bhargav, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 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): 163:1–21. https://doi.org/10.1145/3359265.

Stocking, Galen, Amy Mitchell, Katerina Eva Matsa, Regina Widjaya, Mark Jurkowitz, Shreenita Ghosh, Aaron Smith, Sarah Naseer, and Christopher St Aubin. 2022. "The Role of Alternative Social Media in the News and Information Environment." Washington D.C.: Pew Research Center.

TeBlunthuis, Nathan, Valerie Hase, and Chung-Hong Chan. 2023. "Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can!" *Communication Methods and Measures* 18 (1). https://doi.org/10.1080/19312458.2023.2293713.

Tran, Chau, Kaylea Champion, Benjamin Mako Hill, and Rachel Greenstadt. 2022. "The Risks, Benefits, and Consequences of Prepublication Moderation: Evidence from 17 Wikipedia Language Editions." *Proceedings of the ACM on Human-Computer Interaction* 6 (November): 1–25. https://doi.org/10.1145/3555225.

Trienes, Jan, Andrés Torres Cano, and Djoerd Hiemstra. 2018. "Recommending Users: Whom to Follow on Federated Social Networks." arXiv. https://doi.org/10.48550/arXiv.1811.09292.

Van Hove, Leo. 2016. "Testing Metcalfe's Law: Pitfalls and Possibilities." *Information Economics and Policy* 37 (December): 67–76. https://doi.org/10.1016/j.infoecopol.2016.09.001.

Vu, Anh V., Alice Hutchings, and Ross Anderson. 2023. "No Easy Way Out: The Effectiveness of Deplatforming an Extremist Forum to Suppress Hate and Harassment." arXiv. https://doi.org/10.48550/arXiv.2304.07037.

Zangerle, Eva, and Christine Bauer. 2022. "Evaluating Recommender Systems: Survey and Framework." *ACM Computing Surveys* 55 (8): 170:1–38. https://doi.org/10.1145/3556536.

Zhang, Xiaoquan Michael, and Feng Zhu. 2011. "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia." *American Economic Review* 101 (4): 1601–15. https://doi.org/10.1257/aer.101.4.1601.

Zhu, Ziwei, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. 2021. "Popularity-Opportunity Bias in Collaborative Filtering." In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 85–93. WSDM '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3437963.3441820.

Zulli, Diana, Miao Liu, and Robert Gehl. 2020. "Rethinking the 'Social' in 'Social Media': Insights into Topology, Abstraction, and Scale on the Mastodon Social Network." *New Media & Society* 22 (7): 1188–1205. https://doi.org/10.1177/1461444820912533.