Decentralizing Moderation on Social Media

Andrew Bahsoun

April 2024

1 Introduction

Private ownership of social media platforms such as Twitter and Reddit has come under increasing scrutiny. Considerable effort has been invested in decentralized alternatives such as Steemit [5], Lens Protocol [3], Mastodon, and more, but none of them has gained a following comparable to the major platforms. Even Mastodon, with 1.8 million monthly active users [11], has been a mixed success. Moreover, Mastodon is not entirely decentralized but instead made up of many centralized instances.

After decades of efforts, the fact that decentralized social media has not been a mainstream success for any obvious technological reason raises the question of what is missing. Of course, lack of investment and convenience for consumers is part of the explanation. However, there may also be deeper reasons for the lack of success of decentralized social media.

We contend that decentralized moderation is an unsolved problem. Social media platforms like Twitter and Reddit developed various algorithms and guidelines for moderating user-generated content. These governance mechanisms are approved and guaranteed by the legal owners of the platforms. While there has been a significant amount of research on moderation in social media, much less is known about how to make moderation work in a decentralized setting where all users are peers and no central authority exists.

This project will build a social media platform in order to test various designs for decentralized moderation in real-world scenarios.

2 Goals of Research

On the theory side, we will propose rules for participants on a social media platform to self-moderate user-created content. On the software engineering side, we will create an open-source Slack bot to implement these rules. We will make this bot available as an app so that users can report back to us, and we can adapt both the theory and the implementation in an effort to discover which rules for moderation work best in practice.

3 Project Narrative

In the following, we briefly review the state of the art of moderation on centralized systems and then describe our approach to decentralized moderation.

State of the Art The moderation on Reddit and Twitter has both centralized and decentralized aspects. Most of the moderation is done by algorithms and volunteer moderators, but ultimately, decision responsibility rests with the owners and central admins.

For example, Reddit's most recent transparency report [13], shows that central admins removed 15.5% of the content, the remaining percentage being removed by moderators, mostly with the help

of Reddit's AutoMod algorithm, which accounts for 71.6% of content removal. AutoMod is available for free in all subreddits and configured by the moderators in that subreddit. It can be programmed to automatically recognize and remove certain content, such as specific languages, words, phrases, posts not containing a link or image, and more [1].

For our purposes, the most interesting decentralized aspect of Twitter is Twitter Notes [4], which is open source. Notes are used to add small messages to Twitter content on the main feed, and are moderated by its own users. Users on Notes can mark other messages as helpful or not helpful, and algorithms are used to rate messages and users in the community.

Methodology The aim of our project is to investigate governance mechanisms for *decentralized moderation*. For this purpose, it is not necessary to build on top of social media that are themselves decentralized. Instead, we will create an open-source Slack bot that will implement decentralized moderation in the sense that moderation decisions will be taken cooperatively by the users and not be made by a central administrator.

The two main ingredients are inspired by Reddit's *automoderation* and the *community moderation* of Twitter Notes.

For decentralized automoderation, we will introduce a voting mechanism that allows users to agree among themselves on the parameters of the automoderation algorithm. AutoMod itself is not open-source; however, Reddit allows for custom bots to be run on subreddits, and some of these custom bots are open-source [9].

For decentralizing Twitter Notes the decision of who can become a moderator (contributor) needs to be decentralized. Instead of anonymous moderators, moderators will be pseudonymous and rated according to reputation. The selection of moderators will be voluntary, randomized and reputation based.

Planning Our research will be split into the following parts, planned over 11 weeks.

- Literature review "Moderation on Social Media" (20%). Deliverable (week 3): Paper summarizing the state of the art in content moderation on Twitter and Reddit.
- Designing a decentralized community-based system for self-moderation of user-created content (10%). Deliverable (week 3): Document outlining architecture, APIs, algorithms.
- Implementation (50%). Deliverable (week 7 / week 11): GitHub repository with design document, code, installation instructions and technical documentation.
- Testing (20%). Deliverable: GitHub Issues and Pull Requests.

Literature review and design will be conducted concurrently. The implementation will be broken down into three milestones (prototype, moderation1, moderation2). Testing will start once the prototype is completed and run through to the end of the project.

4 Conclusion

This project has been designed to strike the right balance between novelty and feasibility. On the one hand, decentralized moderation is an open problem. On the other hand, building a Slack bot for user cooperation is a tried and tested approach [8]. Applying the latter to the former gives us a firm starting point.

In the interest of feasibility, we also do not tackle the issues that come with replacing Slack with a decentralized communication protocol. Social media based on decentralized protocols such as Nostr [6], ActivityHub [15], and Solid [2] are under active development. We are planning to combine the research described here with those efforts in the future.

Finally, we believe that the insights this project will generate on decentralized moderation will have ramifications on decentralized governance more generally, which, in turn, plays an increasing role in a range of areas such as cryptocurrencies [10], decentralized finance [7], health care [14] and the energy grid [12].

References

[1] ashtena7. AutoModerator. https://www.reddit.com/wiki/automoderator/, 2024. Information on Reddit's AutoMod.

[2] Tim Berners-Lee. Solid: Social linked data. https://solid.mit.edu. Accessed: 2024-04-15.

[3] Mayank Jain et. al. Revamping Social Networking using Blockchain: Conceptual case-study of Lens Protocol. https://ssrn.com/abstract=4367051, 2022.

[4] Stefan Wojcik et al. Research behind Twitter Notes. https://arxiv.org/abs/2210.15723, 2024. Twitter Notes.

[5] Zhiyong Liu et. al. User incentive mechanism in blockchain-based online community: An empirical study of steemit. https://doi.org/10.1016/j.im.2022.103596., 2022.

[6] Fiatjaf and contributors. NOSTR - notes and other stuff transmitted by relays. https://github.com/fiatjaf/nostr, 2023. Accessed: 2024-04-15.

[7] World Economic Forum. Decentralized finance (defi) policy-maker toolkit. http://www3.weforum.org/docs/WEF_DeFi_Policy_Maker_Toolkit_2021.pdf, 2021. Accessed: 2024-04-15.

[8] Daniel Kronovet. Chore wheel. https://github.com/zaratanDotWorld/choreWheel, 2024. GitHub Repository.

[9] kungming2. Ada-BOT. https://github.com/kungming2/Ada-BOT/, 2024. open-source Reddit Bot.

[10] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf, 2008. Accessed: 2024-04-15.

[11] Meera Navlakha. Turns out Mastodon has way more active users than it thought. https://mashable.com/article/mastodon-monthly-users, 2023. Mastodon monthly users.

[12] Yael Parag and Benjamin K Sovacool. Electricity market design for the prosumer era. *Nature Energy*, 1:16032, 2016. https://www.nature.com/articles/nenergy201632.

[13] Reddit. Transparency report: January to june 2023. https://www.redditinc.com/policies/2023-h1-transparency-report, 2023. Reddit's Transparency Report.

[14] John Shaw and Frank Rudzicz. Blockchain and decentralized autonomous organizations: Applications in health care. *Blockchain in Healthcare Today*, 1, 2017. https://blockchainhealthcaretoday.com/index.php/journal/article/view/8.

[15] Christopher Lemmer Webber and Jessica Tallon. Activitypub. https://www.w3.org/TR/activitypub/, 2018. W3C Recommendation.