Effects of Banning Content on Online Hate Community Social Structure

Vincent Wong, Yong-Yeol Ahn

Indiana University
vmwong@indiana.edu, yyahn@iu.edu

February 12, 2018

Abstract

There has been increasing concern about the presence and moderation of hate content on social media. As these platforms ban hate groups, those communities exhibit backlashing. This raises the question about how banning these users affects the structure of the community they are in. To address this, I will build and analyze interaction networks from before and after the ban of two major subreddits (r/fatpeoplehate and r/CoonTown) from Reddit, the discussion board platform. This ban was notable because a number of subsequent related hate subreddits popped up shortly after the ban. I hypothesize that vocal members of these groups will become more densely connected with each other and more isolated from other groups that stay on Reddit. This work may lead to better understanding of how administrative actions affect communities online.

I. Introduction

There is a common notion that the internet is renowned as a cesspool of hate [1], but there are concerns about how the internet provides a platform that proliferates hate content in society at large [2, 3, 4]. Microblogging sites like Twitter [5], discussion boards like Reddit [6], and VoIP applications like Discord [7] have been removing users and communities deemed hateful. There are major discussions about the banning of hate speech.

For this project, I am interested in studying how administrative removal of hateful communities impacts their social network. There are concerns about whether removing hateful content is effective at reducing the presence or accessibility of this content. In some cases, there is immediate backlash from these communities when their spaces for distributing content are removed, resulting in greater dissemination of the content in other spaces [8].

To study this, I will look at the events on Reddit, the discussion board, when they banned major hate communities in 2015. At the time, Reddit banned a number of subreddits (subforums) it deemed hateful, including "r/fatpeoplehate", a subreddit dedicated to deriding obese people, and "r/CoonTown", a subreddit dedicated to racism against Black people. The resulting backlash ostensibly spilled hate speech across the website [8].

Using publically available historical Reddit post and comment data organized by subreddit, I will built an interaction network from before and after the 2015 subreddit banning to assess the impact this banning had on the social structure of reddit. By looking at the local social structure of users who posted on the banned subreddits, I can determine how user behaviors re-integrated or changed.

There is very little research about the effects of administrative action on social media, or about their effects on online social networks. Although a previous study found that the observed amount of hatespeech across Reddit

decreased [9], it is not clear how the banning translated to the social dynamics that resulted in the backlash seen.

This research will uncover how the administrative policies of social media platforms affect their communities. It may inform methods to handle the moderation of this content in a way that ensures minimal subsequent damage and unintended consequence. Handling these issues requires understanding how the community structure will react in response to these major changes. Outside of the specific context of Reddit, there are broader concerns about how Internet communications may foster spaces for hate groups. The results of this project may inform how platformsâĂŹ actions affect the space of communities on the Internet as a whole.

II. Data

Reddit has a public API which can be used to request and download data. This includes information on the content of a post, meta-information about the post (such as subreddit, time of post, edits, and number of upvotes), the user who posted it, and meta-information about the user. Banned content may not be available through the API, but a number of people stream live Reddit data and upload monthly repositories of their collections. In particular, I will likely use the data provided by user Stuck_In_the_Matrix, who uploads monthly Reddit data on "r/datasets", the dataset sharing/request subreddit [10].

III. Planned Methods

To construct the networks, I will consider both strong and weak connections relative to the level of engagement in a subreddit. Subreddits are user-created subforums dedicated to a particular topic, and users can subscribe to subreddits to see associated content in their individual, aggregated feed. Users can be said to share a weak connection if they both subscribe to a particular subreddit, as their subscription indicates their continual interest in participating in that content as a viewer. I expect many more users subscribe to subreddits than actually participate in discussions on posts in those subreddits.

Users will share a strong connection if they comment on the same post, or if one comments on another's post. This indicates that they both participated in the same discussion. An even stronger definition could involve looking at only direct replies to comment threads. However, there may not be enough data to construct a meaningfully full network using that definition.

By constructing networks of users who participated in the banned subreddits, we can look at the kinds of interactions they shared with users inside and outside of those subreddits. It will be possible to compare their structures of communication before and after the ban. Although there may be high attrition in the accounts involved in the banned subreddits, I want to focus on the users who remained and continued to disseminate hate content in backlash.

New subreddits arose after the original set were banned, which raises questions about the social cohesion of users participating. As part of the post-ban interaction networks, these new subreddits will likely be observable with their own network structures that we can compare to the pre-ban networks. I will determine differences in the connectedness of these networks before and after the ban.

With this data, it will also be possible to investigate the alternative routes that users in those subreddits used to continue expressing their demand for the banned content. For example, we can confirm popular ideas about whether those users created a number of new subreddits for the content (by looking at the number of unique subreddits), whether they switched accounts (through the number of new unique users), or if they simply fled Reddit.

IV. Conclusion

I propose the above project to investigate the effects of banning hate group subreddits on Reddit in 2015. By investigating the interaction networks of users from these subreddits before and after the ban, I will determine how the ban affected these communities. This work aims to address the effects of administrative action on social media platforms.

References

[1] Wolchover, Natalie. "Why Is Everyone on the Internet So Angry?" Scientific American. July 15, 2012. Accessed February 2018. https://www.scientificamerican.com/article/why-is-everyone-on-the-internet-so-angry/

[2] Manzar, Osama. "Hate speech and the role of social media." Livemint. February 9, 2018. Accessed February 12, 2018. http://www.livemint.com/Opinion/ZAHBp4YDLp1BcCnlIuwFON/Hate-speech-and-the-role-of-social-media.html

[3] Sobkowicz, P., Sobkowicz, A. "Dynamics of hate based Internet user networks." Eur. Phys. J. B (2010) 73: 633. https://doi.org/10.1140/epjb/e2010-00039-0

[4] Massanari, Adrienne. "# Gamergate and The Fappening: How RedditâĂŹs algorithm, governance, and culture support toxic technocultures." New Media & Society. Vol 19, Issue 3, pp. 329 - 346. https://doi.org/10.1177/1461444815608807

[5] Neidig, Harper. "Twitter launches hate speech crackdown." The Hill. December 18, 2017. Accessed February 2018. http://thehill.com/policy/technology/365424-twitter-to-begin-enforcing-new-hate-speech-rules

[6] Robertson, Ali. "Reddit bans 'Fat People Hate' and other subreddits under new harassment rules". The Verge. June 10, 2015. Accessed February 2018. https://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit

[7] Newton, Casey. "Discord bans servers that promote Nazi ideology." The Verge. August 14, 2017. Accessed February 2018. https://www.theverge.com/2017/8/14/16145432/discord-nazi-ban-white-supremacist-altright

[8] Koebler, Jason. "This Is the Site Redditors Are Migrating to Now That r/FatPeopleHate Is Banned". VICE: Motherboard. June 10, 2015. Accessed February 2018. https://motherboard.vice.com/en_us/article/9akjme/this-is-the-site-redditors-are-migrating-to-now-that-rfatpeoplehate-is-banned

[9] Chandrasekharan, E, Pavalanathan, U, Srinivasan, A, Glynn, A,Eisenstein, J, Gilbert, E. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech." Journal Proceedings of the ACM on Human-Computer Interaction archive. Vol. 1 Issue CSCW, November 2017. No. 31. ACM New York, NY, USA. doi>10.1145/3134666

[10] User: Stuck_ In_ the_ Matrix. Title: "I have every publicly available Reddit comment for research. 1.7 billion comments @ 250 GB compressed. Any interest in this?" Source: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/