"They're with the Banned": Networked Genealogy for Understanding Toxicity on Reddit

Benjamin Freixas Emery, Department of Information Science, University of Colorado Boulder

May 5, 2024

Abstract

In modern society, hate groups and conspiratorial movements have decentralized and become prominent within online platforms. This has extended their reach and impact, but also allowed for the ability to study them and possible adapt society to become more robust to their behavior. One platform where these spaces have flourished is Reddit. Here we examine six communities on Reddit that have been banned by the platform for toxic behavior. We construct a community genealogy network, capturing migration to these communities from others immediately preceding their respective bans. We also generate trajectory graphs for the individual users during these periods of migration, investigating them using graph embedding and matrix decomposition. We note unexpected features of the genealogy network and potential utility in trajectory graphs. We identify clear next steps for progress toward more interpretable and significant results.

I Introduction

A Background

The popular forum network Reddit has been historically known for maintaining staunchly standing by free-speech absolutism in their policies around moderation. This has added extra weight to their choices to ban certain topical subforums (commonly referred to as subReddits) in recent years. Prior to 2017, Reddit had almost exclusively limited the banning of communities to cases of explicitly orchestrating illegal processes, such as the sale of illicit firearms. Beginning in 2017, the platform began banning communities for inciting violence, in the wake of a vehicular mass-murder by a user of the r/Incels forum. Despite this establishment of a link between hate-speech on Reddit and real-world violence, the banning of subReddits remains rare.

Reddit's compartmental nature, with its activity explicitly categorized by topic, lends itself well to study of discourse and activity around certain topics and narratives. In the present study, we leverage this organizational property of Reddit as well as the relative rarity of moderation to study community genealogy and migration leading up to behavior deemed damaging enough to expel a subReddit.

B Existing work

In the past two decades, Reddit has been a subject of study for qualitative social scientists and quantitative data scientists alike. Among many other subtopics, toxicity in forms such as misogyny, racism, fatphobia, transphobia, Islamophobia, and antisemitism have been investigated [1–5].

We draw on methods established by Chenhao Tan for constructing genealogy graphs for communities [6]. While Tan's work characterized the construction of communities, we characterize the

movement of the users who ultimately led to the subReddit's demise. We expand on this process in Section IIB. We also adapt user trajectory specifications from [7], which we expand on in Section IIC.

C Focus of study

We characterize the activity of users on the following subReddits, all of which were banned for spreading disinformation or promoting violence against marginalized groups: "Incels", "frenworld", "TumblrInAction", "CringeAnarchy", "GenderCritical", "NoNewNormal". We define both of these criteria for banishment from Reddit as "toxicity", broadly speaking. We identify anomalously active users shortly before the banning of each subReddit, analyze and characterize their movement in aggregate and individually.

II Methods

In the following subsections, we describe the methods for retrieving our data and conducting analysis. Network specification and all network analysis, apart from graph embedding, is conducted using Graph-Tool [8].

A Data retrieval

Prior to April of 2023, the Pushshift API [9] provided large-scale access and granular filtering of Reddit data for research. In April 2023, in the light of the use of their data for large language model training, Reddit placed extreme restrictions on access to its data, making no exceptions for non-commercial academic research. To satisfy our data requirements for this research, we constructed a PostgreSQL database containing all submissions and comments on Reddit from 2005 through 2022. Included with each entry is the date, author username, subReddit, a unique identifier for the activity, and − for comments only − the parent activity identifier to which the comment is a reply.

For each of the studied "child" subReddits, we collect all comments on the subReddit from the two years before the date of its ban, and isolate the 100 most active users on the subReddit during this period. We generate a time-series vector for each of these users by volume of activity, using an eight-week-wide sliding window moving in steps of one week, normalized by dividing by peak activity. We cluster these time-series vectors using $k = 6$ means and do the remainder of the analysis using only the cluster skewing the latest. We visualize an example of this time-series clustering in Figure 1. In cases where the latest-skewed cluster contains fewer than ten users, we use the two latest clusters. We refer to this cluster or pair of clusters as the "final cohort". We collect all the comments made by these users in a thirty-day period preceding the first decile of their activity in the child subReddit.

B Genealogy of the fall

Using the identified users paired with the time periods prior to 90% of their activity on the associated child subReddit, we generate a genealogy network: a directed graph where edges from a parent subReddit to a child subReddit is defined by one of these users commenting in the parent before becoming majorly active in the child subReddit. These edges are weighted as

$$w_{ij} = \frac{\#\text{ users from final cohort who posted in } i \text{ then } j}{|\text{final cohort of } j|},$$

essentially capturing the fraction of users coming from the parent subReddit. We use weighted out-degree and weighted undirected betweenness centrality to investigate the prominence of parent subReddit nodes.

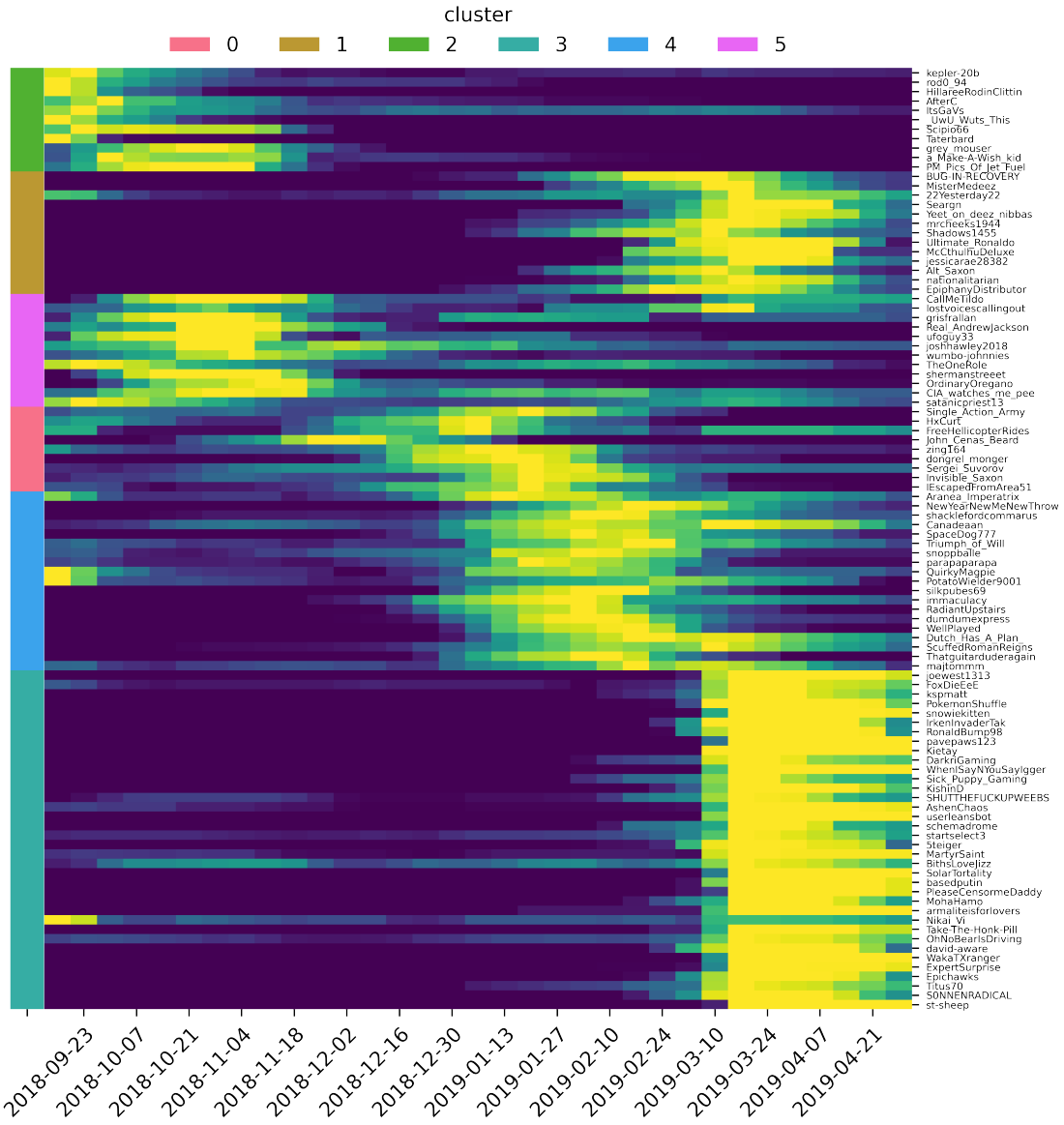

Figure 1: Time-series clustering of activity on the subReddit "frenworld". In this case, cluster 3 would be considered the final cohort.

C Individual user trajectories

For each user in the final cohort of the studied child subReddits, we generate a trajectory graph by treating each subsequent activity in the 30-day studied period as an edge from the subReddit of the previous comment to the subReddit of the following comment. These edges include the total duration between comments as an attribute, which may be used in later studies.

We generate graph embeddings for each of these using Graph2Vec [10]. For analysis, we decompose the embedding matrix $\mathbf{E}$ with singular value decomposition, defined by the equation

$$\mathbf{E} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \mathbf{W}\mathbf{V}^{\mathrm{T}}.$$

III Results

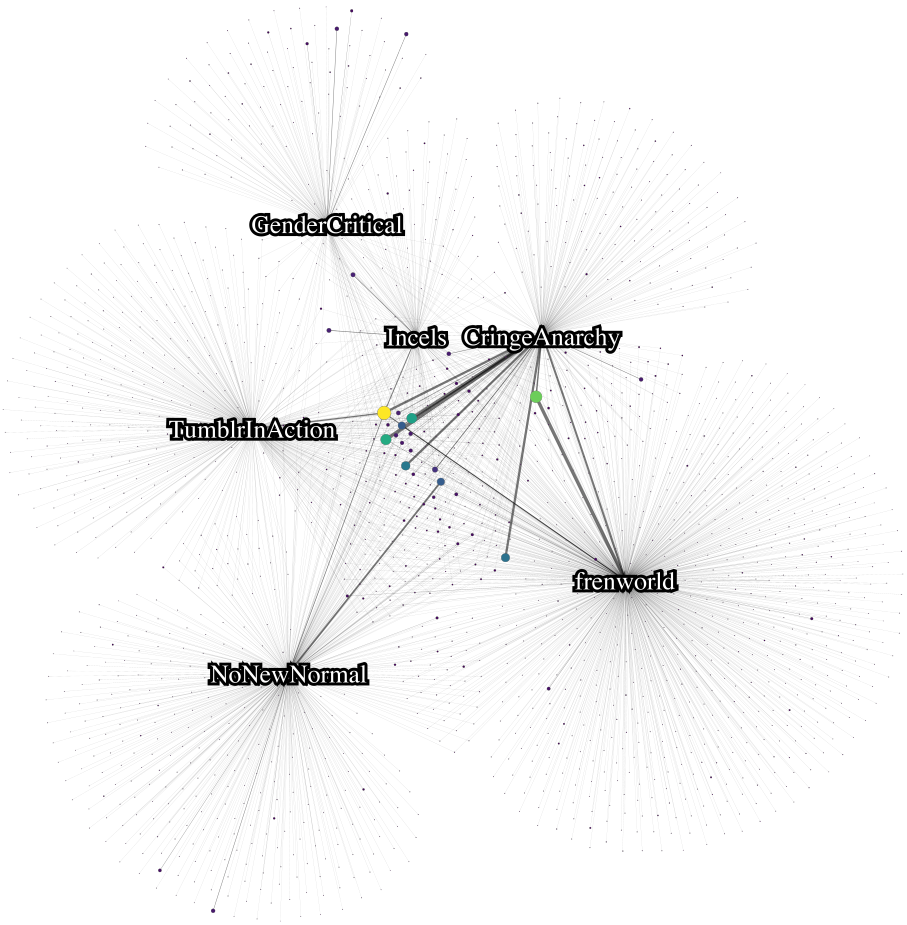

Figure 2: The full genealogy graph produced by the methods described in IIB. Nodes are sized and colored by their out-degree, and the six child subReddit nodes are labeled.

A Community genealogy

We show a spring-force layout of the full genealogy graph in Figure 2. Each edge terminates at one of the six studied child subReddit nodes, and a child subReddit may in some cases be a parent of another. Table 1 shows the ten highest ranked nodes, excluding the six child subReddits, by weighted out-degree and weighted undirected betweenness centrality. In this context, out-degree captures from which subReddits people are migrating to at least one of the child subReddits, and betweenness captures the subReddits feeding the widest breadth of the child subReddits. The top nodes by out-degree are largely very popular mainstream subReddits, many of which were included in the default subReddits prior to the discontinuation of defaults in 2017. The top nodes by betweenness, by contrast, are more niche and have less widespread appeal, making it especially curious that they are sources for users joining multiple different toxic subReddits.

subReddit	out-degree	betweenness	subReddit	out-degree	betweenness
AskReddit	5.630890	0.000000	gatekeeping	0.052081	0.012078
The_Donald	3.243314	0.000000	AskMen	0.069002	0.010831
worldnews	2.732391	0.000000	antiMLM	0.067231	0.010059
news	2.528267	0.000000	creepyPMs	0.057053	0.009408
politics	2.398932	0.000000	NatureIsFuckingLit	0.041903	0.008152
todayilearned	2.223714	0.000000	Frugal	0.044311	0.008136
PoliticalHumor	1.751044	0.000000	jobs	0.044311	0.008136
conspiracy	1.511596	0.001786	IdiotsInCars	0.052081	0.007927
videos	1.260235	0.000000	socialskills	0.042540	0.007306
pics	1.143178	0.000000	space	0.058824	0.007291

Table 1: Highest-ranking nodes by weighted out-degree and weighted undirected betweenness centrality, with the six studied child subReddits removed.

B User trajectories

Trajectories of individual users, as described in Section IIC, are characterized by an abundance of self loops and short cycles. Two examples of these are shown in Figure 3.

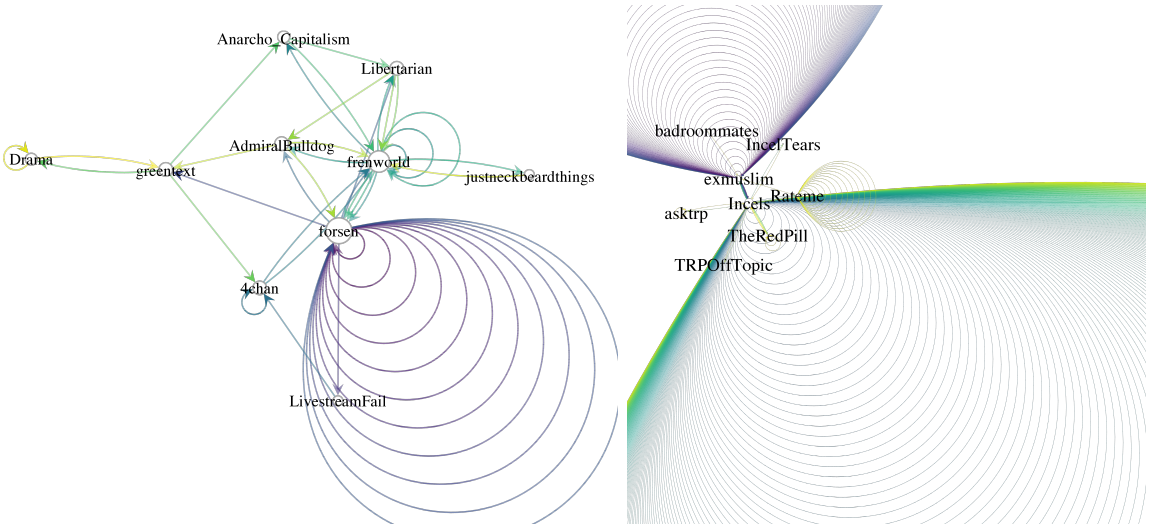

Figure 3: Examples of user trajectories from r/frenworld (left) and r/Incels (right). Edges are colored from dark to light, indicating the order in which they occurred.

We include kernel-density plots of the distributions of the first two singular vectors of the embedding matrix in Figure 4, separated by their associated child subReddits. While there is overwhelming overlap between the different distributions, we note the considerable narrowness of r/NoNewNormal and the rightward shift of r/TumblrInAction in the first dimension, and the rightward shift of r/GenderCritical and r/Incels in the second dimension.

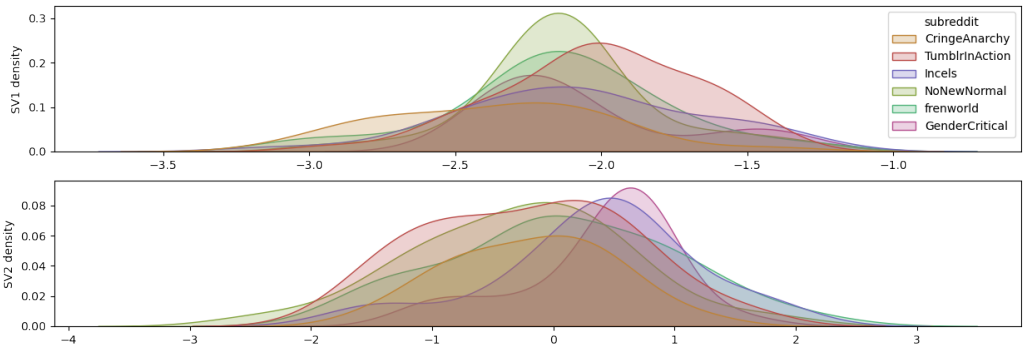

Figure 4: Distributions of the first two dimensions, or singular vectors, of the matrix **W** from the singular value decomposition of the embeddings.

IV Discussion

A Implications

The goal when researching the behavior of these populations on this medium is to ultimately develop more robust strategies for curbing the deleterious impacts of misinformation, hate speech, and conspiracy theories online. To that end, this work presents only an early first step toward these outcomes. Online spaces that host this sort of content not only desensitizes users in those spaces to further vitriol, they also are a hotbed for promoting and inciting actual violence in the forms of doxxing − revealing personal information of a vulnerable person to people who are likely to try and hurt them, swatting − making bogus calls to police to provoke a heavily armed response, or otherwise spread misinformation so vile that readers believe they have no choice but to attack members of the target population.

In addition to threats of physical violence that online spaces such as these pose, they present a political threat. For instance, disinformation about the trans community has prompted states to enact laws banning trans children from participating in sports, using a bathroom consistent with their identity, or consuming literature that represents their community in a positive light. Lies perpetuated on the internet have forced people to essentially remove themselves from society or face severe legal ramifications.

This work has demonstrated the potential for characterizing these spaces such that the emergence of new toxic online spaces maybe noticed before significant damage has been done. Our examination of the genealogy network revealed bridge nodes (those with a high betweenness centrality) whose prominence would be far from obvious. These nodes were overwhelmingly "second tier" subReddits – mainstream communities that have around one million subscribers. These stand in stark contrast to the high-out-degree nodes, which were largely the biggest communities on Reddit, with tens or hundreds of millions of subscribers. We've also identified continuing potential for identifying users migrating into these types of spaces with graph embedding, as we saw some difference between groups, although the sample sizes are too small to say anything conclusive about larger trends.

B Limitations and Future Work

We've identified some small-scale trends and promising techniques for investigating toxicity on Reddit and elsewhere, however, much more work is needed for firm and actionable conclusions to be drawn. Much of the limitation here comes from the narrow scope of this case study, which

only covered six of many toxic subReddits. To expand this, we'd need a procedure for determining toxicity since many such subReddits have never been banned. SubReddits that have been banned are easier to categorize as toxic because that determination has been made by the platform, but determining which subReddits are toxic and the point in time at which they became toxic could prove substantially nontrivial.

It's also impossible to make claims that any of the properties discovered here are exclusive to toxic communities, since we have not compared to any other groups. In the near future, we plan to construct a reasonable methodology for assembling a set of baseline communities to study, and conduct similar analyses using their genealogy and trajectories. At this point we cannot say whether migration patterns can reveal a signal unique to these communities, and we suspect linguistic analysis would need to be integrated into this study to arrive at a robust methodology for characterizing these spaces.

V Conclusion

This work sits at the early end of what we hope to be a an extensive body of literature on toxic online spaces, the threat they pose to a healthy society and democracy, how to identify and predict their behavior, and how to impede the damage they may seek to cause. In our pursuit of this understanding, we remind ourselves of those who have been hurt, killed, or repressed as a result of the rhetoric in these spaces. We must continue to recognize the reason for this research to be greater than an academic or intellectual pursuit, but as a means toward a safe and dignified life for all human beings.

References

1. Gothard, K. *et al. The incel lexicon: Deciphering the emergent cryptolect of a global misogynistic community* Available online at https://arxiv.org/abs/2105.12006. 2021.
2. Dubois, E. & Reepschlager, A. How harassment and hate speech policies have changed over time: Comparing Facebook, Twitter and Reddit (2005–2020). *Policy & Internet* (2024).
3. Mohan, S. *et al. The impact of toxic language on the health of reddit communities* in *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings 30* (2017), 51–56.
4. Chandrasekharan, E. *et al.* You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction* **1,** 1–22 (2017).
5. Rieger, D., Kümpel, A. S., Wich, M., Kiening, T. & Groh, G. Assessing the extent and types of hate speech in fringe communities: A case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society* **7,** 20563051211052906 (2021).
6. Tan, C. *Tracing Community Genealogy: How New Communities Emerge from the Old* in *Proceedings of ICWSM* (2018).
7. Keegan, Brian and Gergle, Darren and Contractor, Noshir. *Staying in the loop: structure and dynamics of Wikipedia's breaking news collaborations* in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (2012), 1–10.
8. Peixoto, T. P. The graph-tool python library. *figshare.* `http://figshare.com/articles/graph_tool/1164194` (2014) (2014).

9. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. *The Pushshift Reddit dataset* in *Proceedings of the international AAAI conference on web and social media* **14** (2020), 830–839.

10. Narayanan, A. *et al.* graph2vec: Learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017).