Estimating the Impact of Coordinated Inauthentic Behavior on Content Recommendations in Social Networks

Anonymous Authors[1]

Abstract

Online disinformation is a dynamic and pervasive problem on social networks as evidenced by a spate of public disasters in light of active efforts to combat it. Since the massive amounts of content generated each day on these platforms is impossible to manually curate, ranking and recommendation algorithms are a key apparatus that drive user interactions. However, the vulnerability of ranking and recommendation algorithms to attack from coordinated campaigns spreading misleading information has been established both theoretically and anecdotally. Unfortunately it is unclear how effective countermeasures to disinformation are in practice due to the limited view we have into the operation of such platforms. In such settings, simulations have emerged as a popular technique to study the long-term effects of content ranking and recommendation systems. We develop a multiagent simulation of a popular social network, Reddit, that aligns with the state-action space available to real users based on the platform's affordances. We collect millions of real-world interactions from Reddit to estimate the network for each user in our dataset and utilise Reddit's self-described content ranking strategies to compare the impact of coordinated activity on content spread by each strategy. We expect that this will inform the design of robust content distribution systems that are resilient against targeted attacks by groups of malicious actors.

1. Introduction

The success and scope of coordinated online campaigns to spread misinformation and hate speech poses a new set of challenges to a free and open and Internet (Tucker et al., 2018). A great deal of scientific effort has been devoted to building content classifiers that can accurately identify what is deemed as harmful content (Aldwairi & Alwahedi, 2018; Shu et al., 2017; Zhou & Zafarani, 2020; Zhou et al., 2020) in order to remove it from ranking inventory. This approach suffers from limitations inherent in the contentious definition of what "harm" is (Vraga & Bode, 2020), as well as from resource constraints which lead to the uneven application of content-focused integrity mechanisms, particularly in the Global South (Haque et al., 2020). Notwithstanding the challenges of veracity, there are petabytes of data generated daily on these platforms[1] which necessitates algorithmic content distribution. Ranking and recommendation systems have played a tremendous role with platforms relying heavily on large-scale algorithms to populate the content stream that a user interacts with, called their 'feed' or 'timeline'. However, these systems have been shown to suffer from biases, and even promote hateful content[2] (Jiang et al., 2019; Tomlein et al., 2021; Ribeiro et al., 2020). One of the reasons is that these algorithms often rely on metrics that can be manipulated through coordinated activity including fake profiles and illicit collusion (Giglietto et al., 2019; 2020; Nizzoli et al., 2020). This makes such algorithms a popular attack vector for coordinated campaigns attempting to promote misleading content on social networks.

In fact, a May 2022 post published by a leading cybersecurity firm Nisos reveals the existence of professional botnet tool designed to simulate online personas replete with profile pictures, behaviors specific to the social media platforms, all with the goal of manufacturing fake 'trending social media events en masse'[3]. The software, clearly targeting mechanisms that drive content virality on social media, existed as early as 2019. In a similar report published in 2020, The Wire (India) conducted an investigation confirming the existence of a tool that artificially promoted hashtag-based trends using coordinated activity across different fake and

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]https://research.facebook.com/blog/2014/10/facebook-s-top-open-data-problems/

[2]https://about.fb.com/news/2018/11/myanmar-hria/

[3]https://www.nisos.com/blog/fronton-botnet-report/

real social media accounts[4].

The publication of quarterly transparency reports highlighting networks of such coordinated accounts successfully identified by platforms like Facebook and Twitter sheds light on the issue[5]. These data confirm that similar strategies are being adopted across multiple geographies in order to share misleading content that could influence susceptible users. Some of these are similar in principle to the BEND framework published in (Beskow & Carley, 2019b) that attempts to formalize disinformation manoeuvres observed on social media.

2. Contributions

We provide a multiagent simulation that can serve as a virtual test-bed for platforms and the public to evaluate the impact of such activity and prototype countermeasures to these patterns laying emphasis on a decomposable, transparent, and realistic content ranking mechanism that drives content virality on social networks. Using this simulator, we compare the harms arising from coordinated inauthentic behavior on online communities of users. Our model of agent activity is based on the popular social network Reddit, mapping the state-action space of agents in our simulation to the primary set of actions a user might take on the platform, whilst engaging with content online. Reddit has publicly accessible posts and comments dating back to 2007 and importantly, relies on decomposable content ranking mechanisms that are documented in public-facing posts published by the platform.

Substantively, we investigate the following research questions:

1. Can we quantify the effects of coordinated campaigns that target content amplification by ranking and recommendation systems?

2. If yes, can this inform the design of algorithms that are less susceptible to harms arising from such behavior?

3. Related Literature

Early work on detecting influence operations focuses on graph-theoretic approaches studying misinformation mitigation in social network graphs (Nguyen et al., 2012; Zhang et al., 2015; Amoruso et al., 2017; Saxena et al., 2020). These contributions assume traditional means of information propagation over static friend-follower networks and employ influence models (Kempe et al., 2015; Chen et al., 2013) to measure responses to interventions such as the debunking of information in social networks.

Agent-based models were also employed to simulate the spread of misinformation in a manner analogous to the spread of an infection in a population, drawing on epidemiological dynamics theories to determine the bounds on detection methods under specific assumptions on information propagation (Dong et al., 2013; Wang et al., 2014). Common applied methods from epidemiology utilise basic compartmental models like the *Susceptible-Infected-Recovered* (SIR) model (El-Sayed et al., 2012; Tambuscio et al., 2015; Wang et al., 2014; Shelke & Attar, 2019) or the *Susceptible-Exposed-Infected-Recovered* (SEIR) model (Zhou et al., 2019) to simulate spread with some work framing interventions on it in the form of 'vaccinated' agents that debunk misinformation claims in the network (Serrano & Iglesias, 2016). On the theoretical end of agent-based modeling, the authors of (Beskow & Carley, 2019a) simulate bot disinformation manoeuvres that target susceptible online communities. While useful as a prototyping tool for which they provide the BEND framework to categorize malicious patterns of behavior (Beskow & Carley, 2019b; Carley, 2020), there is no mention of any platform-specific mechanisms driving agent interactions with content, assuming that content distribution occurs via friend-follower networks. The existing literature on agent-based models involving social networks has largely ignored algorithmic confounding (Chaney et al., 2018) and other micro-mechanisms in their aim to explain macro-level phenomena, in the process marginalising away individual uniqueness in behavior.

3.1. Agent-based Models of Social Networks

The novelty of employing agent-based models has been in trivial introductions of micro-mechanisms that collectively cause complex macro-trends, as evidenced in Schelling's model of segregation (Schelling, 1969) over five decades ago and Conway's Game of Life Simulations thereafter (Conway et al., 1970). Now-abandoned work towards agent-based social simulations typically focused on high-fidelity user interactions (Ryczko et al., 2017) and their network evolution over time (Stadtfeld, 2015) without accounting for the platform-level mechanisms that underpin most of these interactions. Recent work continues to accord due importance to such micro-mechanisms and accounts for some of the platform-level nuances including recommendation algorithms such as (Lucherini et al., 2021) permitting the introduction of user and item-level attributes as part of the system. Similarly, the authors of (Mladenov et al., 2021) emphasize the need for simulated recommender systems to operate in an ecosystem involving complex, *multi-turn interactions* similar to modern-day recommendation systems operating on dynamic social networks. While these are use-

[4] https://thewire.in/tekfog/en/1.html

[5] https://about.fb.com/news/tag/coordinated-inauthentic-behavior/, https://transparency.twitter.com/en/reports/information-operations.html

ful frameworks to examine socially-meaningful outcomes and design complex recommendation algorithms respectively, they neither utilise real-world data to inform their simulation choices, nor examine coordinated activity in the context of ranking and recommendation systems on a social network which forms the primary contribution of our work. That said, we hope to provide a compatible implementation of our simulation in their respective frameworks, to stimulate collaborative research in this direction. The notion of coordinated inauthentic behavior (CIB) was introduced by Meta (formerly Facebook) quite recently, in 2018. Most existing work in the area of coordinated inauthentic behavior focuses on its detection or mitigation with little emphasis on the mechanism through which the accounts involved in these activities attempt to game the ranking and recommendation algorithms. As the first step towards addressing this challenge, we contribute to the measurement of harms arising from coordinated disinformation campaigns on social networks. Operationally, we combine multiagent simulations with algorithmic content distribution via ranking and recommendation systems and track metrics that are relevant to a healthy ecosystem. We rely on priors informed by a large-scale dataset comprising millions of interactions between users of Reddit, a pseudonymous social network. The emphasis on priors informed by real-world data reduces our reliance on an arbitrary choice of simulation parameters increasing our confidence in the results. We adopt information-theoretic definitions similar to (Lucherini et al., 2021) to quantify the harms arising from coordinated inauthentic behavior and conduct a comparative study of risk-quantification conditional upon choice of recommendation algorithm. Lastly, we generalize the components of our simulator to different social media platforms to describe how our software serves as a framework for analyzing the impact of coordinated attacks on end-users of any of these platforms.

4. Simulating Coordinated Inauthentic Behavior

"Coordinated inauthentic behavior" (CIB) is a term coined by Facebook in late 2018, describing the promotion of content via coordinated networks of accounts on its platform, with the 'intent to mislead people about who they are and what they are doing'[6]. The operationalization of CIB is often through different attack vectors for the exploitation of ranking and recommendation algorithms to target content visibility. The harms arising from CIB are difficult to quantify due to a lack of the apparatus to track the impact of coordinated inauthentic behavior across evolving networks in an open and transparent manner supported by platforms. Furthermore, due to the inherent lag caused by internal investigations, most data about CIB is published weeks or even months after attacks. This delay further limits researchers' ability to quantify CIB harms. Due to the nature of disinformation campaigns involving unsuspecting users (Starbird et al., 2019), it is critical to verify authentic users are not part of the takedowns by site integrity teams. Simulations of social media provide a means to study CIB by modeling the ecosystem in which it takes place. In this paper we propose to model two popular patterns of coordinated activity and present preliminary results from the first:

1. Brigading, in the Reddit context describing inter-community conflicts during which antagonistic members of one subreddit actively downvote comments in order to deprioritise them for content recommendation systems and effectively censor them[7] (Datta & Adar, 2019).

2. Influence Operations, wherein a set of 'puppet' accounts are used to push a certain narrative through repetitive posts and comments reiterating the same, falsifying the appearance of credibility and popular sentiment.

4.1. Simulating Social Network Activity for Reddit

Reddit is a pseudonymous online social network that primarily comprises a network of communities, as stated on their website, `reddit.com`. Our interest in Reddit as a content-sharing platform arises from the fact that it has largely passed under the radar due to the high-visibility consequences of disinformation on social networks like Twitter and Facebook despite having nearly the same number of monthly active users as Twitter[8]. That has not been for want of controversy; Reddit also dealt with the controversial removal of thousands of communities sharing misinformation, particularly United States politics (Chandrasekharan et al., 2022) and misinformation relating to the COVID-19 pandemic. It is possible that Reddit's community structure and human-moderated posting may have reduced the effectiveness of disinformation. It seems equally plausible that there is simply more research being conducted into the impact of disinformation on Twitter and Facebook than into Reddit. In any case, given the open nature of Reddit data and open-source tools to sift through these troves of user activity provided by Pushshift (Baumgartner et al., 2020), it is a useful platform to model activity on. Our conclusion was confirmed by the authors of a large-scale survey (Proferes et al.,

[6] https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[7] https://institute.global/policy/social-media-futures-what-brigading

[8] https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users

2021) who state that Reddit is only recently turning into a increasingly popular avenue for research, having examined several hundred publications employing Reddit datasets in their analyses. Our design choices are informed by a careful examination of past simulation tools (Lucherini et al., 2021; Mladenov et al., 2021; Ryczko et al., 2017; Stadtfeld, 2015) and a forward-looking perspective, as stated earlier, to provide a compatible API to interface with other tools as well as run Monte Carlo-based inference algorithms in a tractable manner. We employ the probabilistic programming language Pyro (Bingham et al., 2019) to develop a forward simulator Among other advantages it extends the popular Pytorch framework (Paszke et al., 2019) and is packaged to include Monte Carlo and variational inference methods that will enable us to condition the simulation on user activity collected from various social networks. This implies that while making design choices, we will rely on simpler control flow logic to enable efficient inference of agent-level latent variables that will be defined in the simulator. As a general principle, we strike a balance between the fidelity of the state-action space available to the user and the tractability of statistical inference to obtain posterior estimates over the parameter values given access to user data from social networks. In practice, this means making choices to simulate only the voting, posting, and commenting behavior, with little emphasis laid on modeling the shifts in the network (or subreddit-membership) of the users. Here is a technical overview of our framework following which we provide a description of the agent activity model for Reddit.

1. Our simulation comprises of 'Content' and 'User' which are the two atomic entities defined in an object-oriented manner–each comprising a set of properties that are used to simulate interactions in the ecosystem.

   For the 'Content' object, these properties are its unique identifier, author, time of creation, topic attributes (what it is about), scoring statistics (number of comments, shares, and votes), and recency (last active);

   For the 'User' object, these properties are its unique identifier, interests, session activity (determines number of posts per session), influenceability, response (determines tendency to comment or vote), and posting probabilities.

2. Specializing our model to Reddit, the subreddit membership is determined based on the user's interests and the content's topic attributes.

3. We parametrize behavioral models describing agent-level activity shown in 5 by placing data-driven priors on parameters driving their behavior.

   These refer to the 'interaction' and 'response' properties of the 'User' based on their frequency of posting and commenting in different subreddits

   These also include the static membership of subreddits for each user that is a part of the simulation.

4. We use counterfactual simulations to estimate the effects of accounts engaging in CIB by targeting specific content to amplify, lying beyond a reasonable agreement with their existing set of interests inferred from community membership.

5. For agents that do not engage in CIB, activities (vote, comment, post, do nothing) are sampled while browsing through the content feed curated by a particular content distribution algorithm.

6. In the simulation, some agents target specific communities based on a predetermined agenda, in a coordinated fashion.

7. Such activity affects the inputs that go into a recommendation algorithm which filters and ranks content on a user's feed.

8. To determine the level of impact, we track metrics of interest and compare them across multiple runs of the simulation for a CIB setting versus a counterfactual setting without the presence of CIB to determine how much damage is done as a result of CIB. We examine:

   The content diversity or the variance of all recommended content at each timestep to measure the degree of homogenization of visible content similar to (Chaney et al., 2018; Lucherini et al., 2021)

   The share of views allocated to misleading content promoted via coordinated activity relative to the counterfactual scenario.

   The share of engagement with such content in comparison to the counterfactual scenario.

5. Running the Simulation

The algorithm in 5 describes a single run of the simulator called an 'execution trace'. Each execution trace provides a set of metrics over time that we collect over multiple runs in order to provide a confidence interval over the metrics of interest. We conduct such an experiment for each choice of ranking and recommendation algorithm as well as behavior type in order to provide a complete picture of algorithmic susceptibility towards that kind of coordinated inauthentic behavior.

6. Dataset

For data-driven simulations, we collect a dataset of Reddit user actvity that contains millions of posts from 2011 - 2021. We start by scraping all of the posts on the 'r/politics' subreddit and a sample of upto 10,000 comments per post, with

the motivation that it often provides different, disjoint perspectives on policy issues and would comprise interesting behaviors and a variety of tones of discourse. We filter the most active 5,000 users based on recent activity and limit the time period for modeling to 2016 - 2021 for computational reasons. For each user in our dataset we scrape a history of upto 6,000 past Reddit posts and comments that we collectively term as 'interactions'. This results in a total of nearly 32,000 different subreddits. We present an analysis involving 2,500 of these users as a pedagogical example of conducting research into the impact of coordinated behavior on social networks. The challenge in simulating a realistic subreddit network is the sparsity of subreddit membership and long-tail of behaviors corresponding to these communities (Krishnan et al., 2018). however, in order to utilise all available information about users without creating an intractable computational simulation, we develop a strategy that we call 'subreddit categorization' which uses a supervised labeling approach to cluster subreddits into a five-level hierarchy of subreddit 'categories'.

6.1. Subreddit Categorization

We manually collect data from 'Wikis' that are user-generated labels for describing what content category is associated for a given subreddit. For example the subreddit 'r/cats' would fall under the category 'Animal Kingdom' and subcategory 'Animals'. We find 26 top-level categories, 151 sub-categories, and 304 sub-sub-categories. While there is technically no limit to the number of levels associated with such a hierarchy, we examine the sparsity of labeled subcategories and accordingly limit the number of levels to a maximum of 5 for the data collection In practice we utilise the top-level category for the current model. We evaluate our subreddit categorization model to find that it has a top-k precision of 0.485 @ $k = 1$. This strategy gives us a massive computational advantage by reducing our modeling granularity to only 26 subreddit categories for the simulation instead of the previous 32,000. We provide a visual description of a few top-level content categories in figure 6 and the real-world interactions collected for 5,000 of the most active users on 'r/politics' divided into these categories shown in figure 4.

6.2. Ranking and Recommendation on Reddit

Reddit has historically been a platform with lightly personalized content; as a result community dynamics are the key drivers of content visibility. Content ranking in their user 'feeds' is primarily done by considering various properties of the content such as its score, total upvotes and downvotes, age, recent activity, and number of comments, with the personalization focused on the time spent by a user on a subreddit. They offer a neat breakdown[9] of their content recommendation algorithms that are visually accessible on the platform as in figure 1 to generate ranked content 'feeds' for users to browse. Our simulation makes the simplifying assumption that a single algorithm drives all the interactions for users in a single execution trace in order to address the concerns of interaction effects as we study algorithmic integrity in this setting. This is a trivial assumption to drop as we could simulate interactions with potentially a different choice of recommendation algorithm at each timestep.

7. Results

We parametrize our simulator with informed priors drawn from historical agent-level behavior. This includes accounts that have a history of engagement with misleading posts and continue to engage in coordinated behavior on the platform. The simulator is then run in a 'forward' execution mode in order to generate interaction data between users and content on the social network. In order to examine the robustness of the recommendation algorithm to the type of coordinated activity, we track the metrics across multiple execution traces of the simulation. We consider four recommendation algorithms on Reddit:

1. Controversial: Promotes content with a large number of upvotes as well as a large number of downvotes
2. Rising: Promotes recent content with a large number of comments and upvotes
3. Top: Promotes content from a fixed time period with a large number of upvotes
4. New: Promotes content based on lower age

The preliminary results in 3 indicate interesting dynamics are emerging in the system such as the controversial recommendation algorithm resulting in much higher views of disinformation over time while the rising algorithm almost entirely diminishes the disinformation views. We continue to examine interesting patterns in our ongoing experiments.

8. Future Work

8.1. Measuring Interventional Impact on Misinformation Spread

It is well-documented that recommender systems, underpinning most engagement on social networks, have egregious design flaws resulting in disturbing consequences. Yet, a causal understanding of the impact of recommender systems

[9] https://www.reddit.com/r/blog/comments/o5tjcn/evolving_the_best_sort_for_reddits_home_feed/

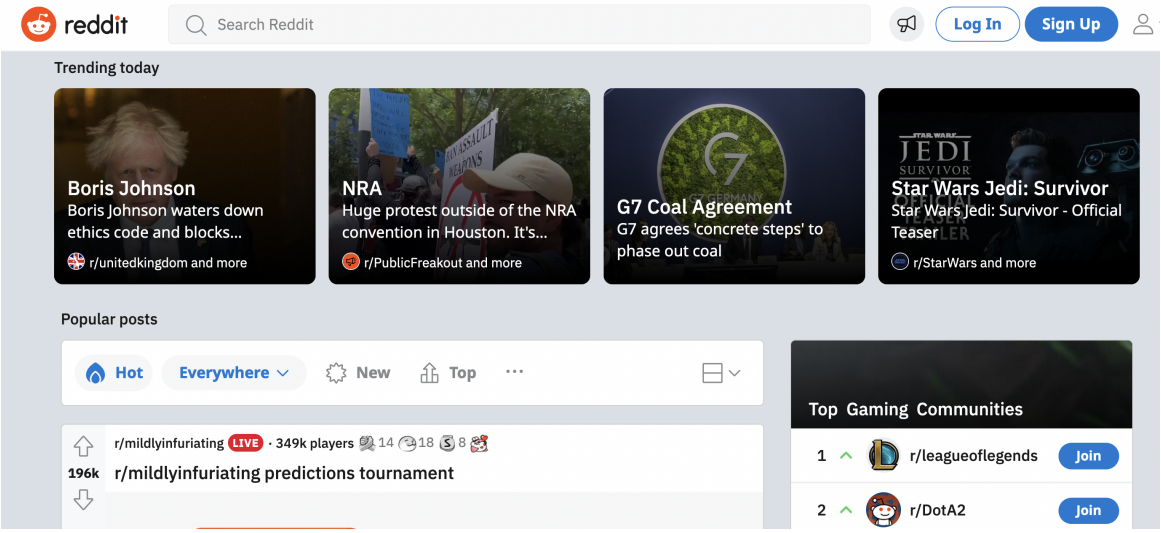

Figure 1. Reddit's Feed permits users to choose their content ranking and recommendation algorithm

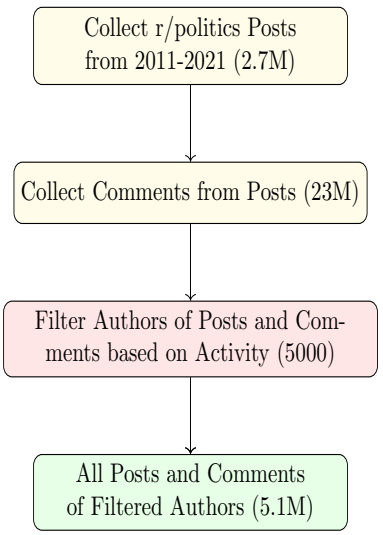

Figure 2. Collecting a large-scale dataset of user interactions from Reddit via Pushshift

has remained tenuous in practice, partly owing to a lack of opportunities for long-term studies involving empirical data. We provide a rich test-bed for estimating the longitudinal impact of interventions on social networks, in particular on content ranking and recommendation algorithms. This extends the means to study the societal impact of recommendations (**?**) including the side-effects of interventions on user behavior. Our work also makes it possible to study the adversarial nature of combating online disinformation by employing reinforcement-learning based policy-learning mechanisms to check what strategies are learnt by agents trained to promote disinformation or gain influence over longer time horizons and how they could be countered.

8.2. Community-specific Behavioral Modeling

The ecosystem we present can be used to study novel agent behaviors specific to a community with specialized simulations to detect problematic user-level patterns particular to a subreddit. One can model user activity on subreddits that are known to contain a history of misleading posts. For further evaluation, we can collect verified misleading posts relating to the COVID-19 pandemic from the subreddit 'r/coronavirus'. One can then divide these posts into a training and test set and collect historical user activity for those accounts that engage with the misleading posts. The detection strategy could employ simulation-based inference to identify instances of coordinated behavior among these accounts which would be evaluated using the test dataset. In such a setup, one can also track the false positive rate of a detection strategy using the fraction of detected CIB occurrences involving the account that share a verified post containing misinformation. This also provides insight into the long-tail of agent behaviors particularly in communities

where there is a skewed record of agent-behaviorsin the collected dataset (Krishnan et al., 2018). By encoding theories of social influence (Hsu et al., 2021) in combination with community dynamics in an intuitive simulation, one can communicate the learnings to a broader audience than was previously possible . This is a part of ongoing work within the simulation framework we provided.

References

Aldwairi, M. and Alwahedi, A. Detecting fake news in social media networks. *Procedia Computer Science*, 141: 215–222, 2018.

Amoruso, M., Anello, D., Auletta, V., and Ferraioli, D. Contrasting the spread of misinformation in online social networks. In *AAMAS*, 2017.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pp. 830–839, 2020.

Beskow, D. M. and Carley, K. M. Agent based simulation of bot disinformation maneuvers in twitter. In *2019 Winter simulation conference (WSC)*, pp. 750–761. IEEE, 2019a.

Beskow, D. M. and Carley, K. M. Social cybersecurity: an emerging national security requirement. Technical report, Carnegie Mellon University Pittsburgh United States, 2019b.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

Carley, K. M. Social cybersecurity: an emerging science. *Computational and mathematical organization theory*, 26 (4):365–381, 2020.

Chandrasekharan, E., Jhaver, S., Bruckman, A., and Gilbert, E. Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4): 1–26, 2022.

Chaney, A. J., Stewart, B. M., and Engelhardt, B. E. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 224–232, 2018.

Chen, W., Lakshmanan, L. V., and Castillo, C. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

Conway, J. et al. The game of life. *Scientific American*, 223 (4):4, 1970.

Datta, S. and Adar, E. Extracting inter-community conflicts in reddit. In *Proceedings of the international AAAI conference on Web and Social Media*, volume 13, pp. 146–157, 2019.

Dong, W., Zhang, W., and Tan, C. W. Rooting out the rumor culprit from suspects. In *2013 IEEE International Symposium on Information Theory*, pp. 2671–2675. IEEE, 2013.

El-Sayed, A. M., Scarborough, P., Seemann, L., and Galea, S. Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic Perspectives & Innovations*, 9(1):1–9, 2012.

Giglietto, F., Righetti, N., and Marino, G. Understanding coordinated and inauthentic link sharing behavior on facebook in the run-up to 2018 general election and 2019 european election in italy. 2019.

Giglietto, F., Righetti, N., Rossi, L., and Marino, G. Coordinated link sharing behavior as a signal to surface sources of problematic information on facebook. In *International Conference on Social Media and Society*, pp. 85–91, 2020.

Haque, M. M., Yousuf, M., Alam, A. S., Saha, P., Ahmed, S. I., and Hassan, N. Combating misinformation in bangladesh: roles and responsibilities as perceived by journalists, fact-checkers, and users. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–32, 2020.

Hsu, C.-C., Ajorlou, A., and Jadbabaie, A. Persuasion, news sharing, and cascades on social networks. *News Sharing, and Cascades on Social Networks (September 30, 2021)*, 2021.

Jiang, R., Chiappa, S., Lattimore, T., György, A., and Kohli, P. Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 383–390, 2019.

Kempe, D., Kleinberg, J. M., and Tardos, É. Maximizing the spread of influence through a social network. *Theory Comput.*, 11:105–147, 2015.

Krishnan, A., Sharma, A., and Sundaram, H. Insights from the long-tail: Learning latent representations of online user behavior in the presence of skew and sparsity. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 297–306, 2018.

Lucherini, E., Sun, M., Winecoff, A., and Narayanan, A. T-recs: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959*, 2021.

Mladenov, M., Hsu, C.-W., Jain, V., Ie, E., Colby, C., Mayoraz, N., Pham, H., Tran, D., Vendrov, I., and Boutilier, C. Recsim ng: Toward principled uncertainty modeling for recommender ecosystems. *arXiv preprint arXiv:2103.08057*, 2021.

Nguyen, D. T., Nguyen, N. P., and Thai, M. T. Sources of misinformation in online social networks: Who to suspect? In *MILCOM 2012-2012 IEEE Military Communications Conference*, pp. 1–6. IEEE, 2012.

Nizzoli, L., Tardelli, S., Avvenuti, M., Cresci, S., and Tesconi, M. Coordinated behavior on social media in 2019 uk general election. *arXiv preprint arXiv:2008.08370*, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., and Zimmer, M. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004, 2021.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., and Meira Jr, W. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 131–141, 2020.

Ryczko, K., Domurad, A., Buhagiar, N., and Tamblyn, I. Hashkat: large-scale simulations of online social networks. *Social Network Analysis and Mining*, 7(1):1–13, 2017.

Saxena, A., Hsu, W., Lee, M. L., Chieu, H. L., Ng, L., and Teow, L.-N. Mitigating misinformation in online social network with top-k debunkers and evolving user opinions. *Companion Proceedings of the Web Conference 2020*, 2020.

Schelling, T. C. Models of segregation. *The American economic review*, 59(2):488–493, 1969.

Serrano, E. and Iglesias, C. A. Validating viral marketing strategies in twitter via agent-based social simulation. *Expert Systems with Applications*, 50:140–150, 2016.

Shelke, S. and Attar, V. Source detection of rumor in social network–a review. *Online Social Networks and Media*, 9: 30–42, 2019.

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

Stadtfeld, C. Netsim: A social networks simulation tool in r. *R package vignette http://www. social-networks. ethz. ch/research/research-projects. html*, 2015.

Starbird, K., Arif, A., and Wilson, T. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

Tambuscio, M., Ruffo, G., Flammini, A., and Menczer, F. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th international conference on World Wide Web*, pp. 977–982, 2015.

Tomlein, M., Pecher, B., Simko, J., Srba, I., Moro, R., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., and Bielikova, M. An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes. In *Fifteenth ACM Conference on Recommender Systems*, pp. 1–11, 2021.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

Vraga, E. K. and Bode, L. Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1):136–144, 2020.

Wang, Z., Dong, W., Zhang, W., and Tan, C. W. Rumor source detection with multiple observations: Fundamental limits and algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 42(1):1–13, 2014.

Zhang, H., Zhang, H., Li, X., and Thai, M. T. Limiting the spread of misinformation while effectively raising awareness in social networks. In *CSoNet*, 2015.

Zhou, X. and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.

Zhou, Y., Wu, C., Zhu, Q., Xiang, Y., and Loke, S. W. Rumor source detection in networks based on the seir model. *IEEE access*, 7:45240–45258, 2019.

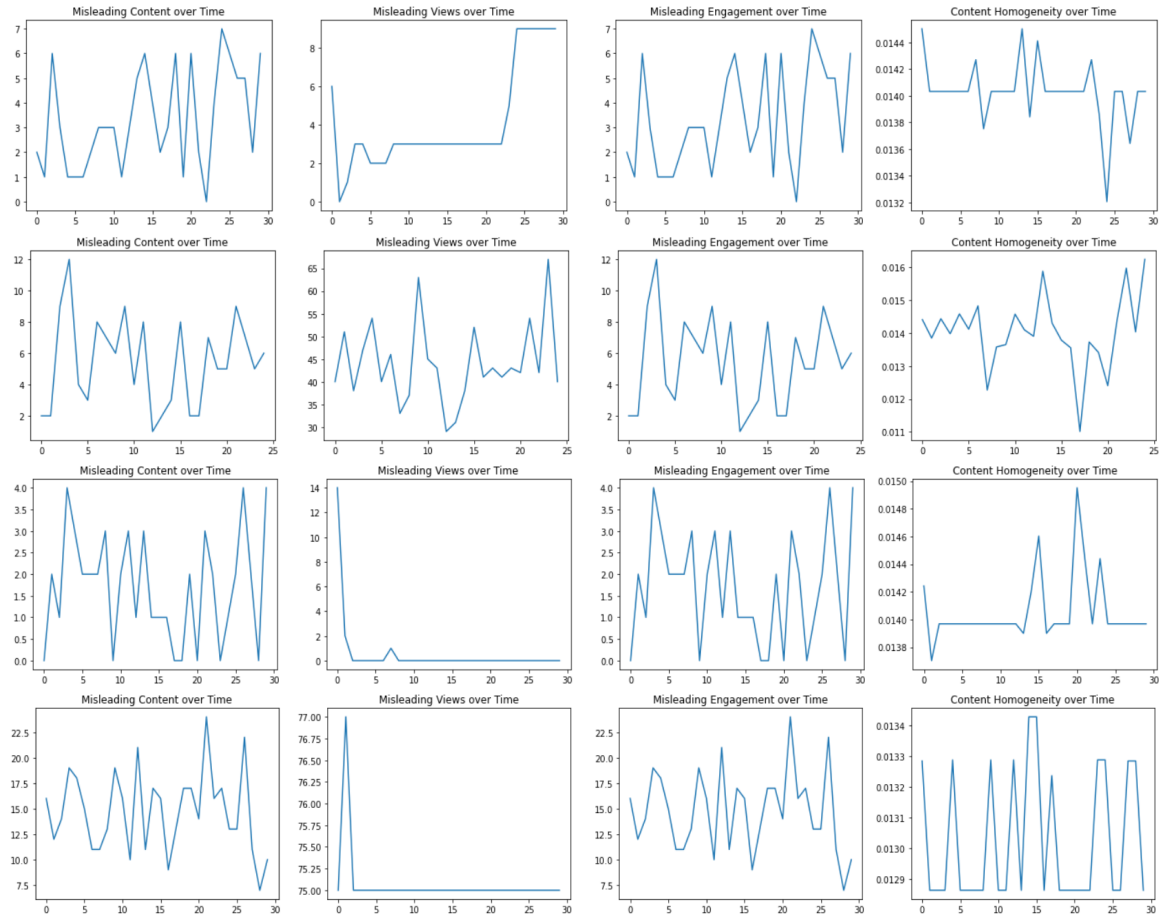

Figure 3. Preliminary Results of Simulating the impact of Brigading on Reddit for (1) Controversial (2) New (3) Rising (4) Top Ranking Algorithms

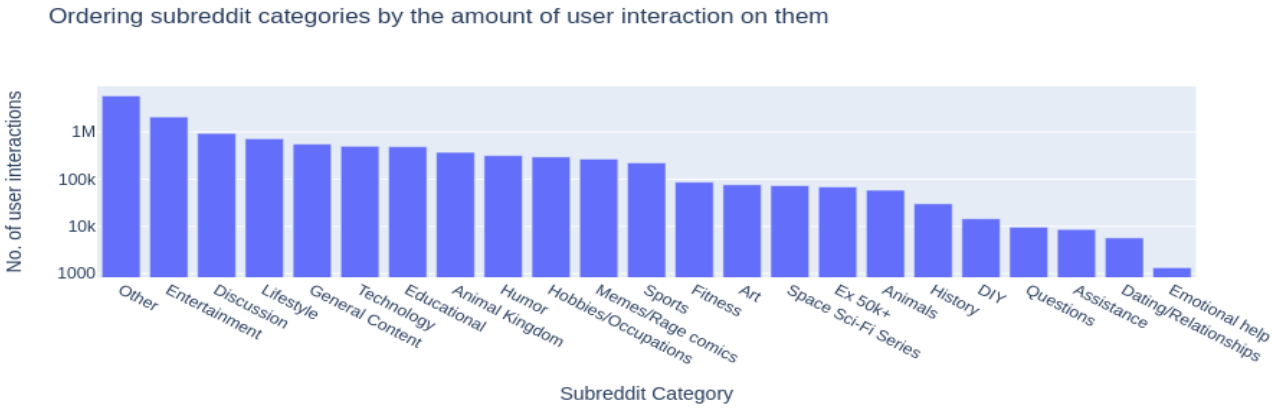

Figure 4. Estimated subreddit categories for 5,000 users' Reddit interactions

REDDIT MODEL

Simulation components

- Content
  - attributes (vector), -1, …, 1 for each content category
  - parent: another content. if present, this content is a comment on another content. if not present, this content is a post
  - score (scalar) integer of accumulated up/down votes
  - subreddit membership (vector) same thing with user's subreddit membership, but it is a one-hot vector
  - time of creation
- User
  - attributes (vector), -1, ,.., 1, for each content category
  - curiosity (scalar), 0, …, 1
  - influenceability (scalar), 0, …, 1
  - memory (scalar), 0, …, 1
  - subreddit membership (vector) of size number of subreddits. binary components indicating membership in each subreddit.

Simulation pseudocode

- N: number of users
- T: number of time steps
- S: number of subreddits / topics
- F: number of posts in each user's feed
- A: similarity threshold for a user to engage with a content

- Initialize users:
  - For each user
    - Sample (based on a profile)
      - (f, c, i, m) ~ priors (can be based on data)
      - s = f(t) or s ~ f(t)

- Initialize the contents
  - Add seed content based on experimental goal
    - One predominant topic
    - Evenly distributed Topics

- Run
  - For each time step
    - Shuffle the user order (in order to avoid bias due to sequential execution of users)
    - For each user
      - Generate a user feed for the user based on a recommendation system (one of the following), picking a fixed number of posts from the global feed
        - New(global feed, for each content, consider: recency)
        - Top(global feed, for each content, consider: recency, score)
        - Personalized(global feed, for each content, consider: recency, similarity of post and user attributes)
        - Controversial(global feed, for each content, consider: upvotes, downvotes)
        - Rising(global feed, for each content, consider: latest upvoted/commented, score)
      - Activity (one of the following)
        - Sample number of posts to interact with (level of activity in the session)
        - For each post (in the sequence given by the recommendation algorithm)
          - Sample action for a categorical (respond=(comment or vote), post, do nothing)
            - Flip a coin for a (comment, vote), and do one of the following
              - Comment
              - Vote
                - Compute similarity
                - if similarity over threshold: Upvote
                - Else: Downvote
          - Post
          - Do Nothing
      - Update user topics (world view)
        - As a function of user's influenceability and the mean topic of their user feed (topic += influenceability * mean_topic_of_user_feed). influenceability is something like a small step size.
    - Compute metrics as a function of the global feed, the user's topics etc.
      - Information entropy
      - Others

Figure 5. The Pseudocode for simulating Activity on Reddit

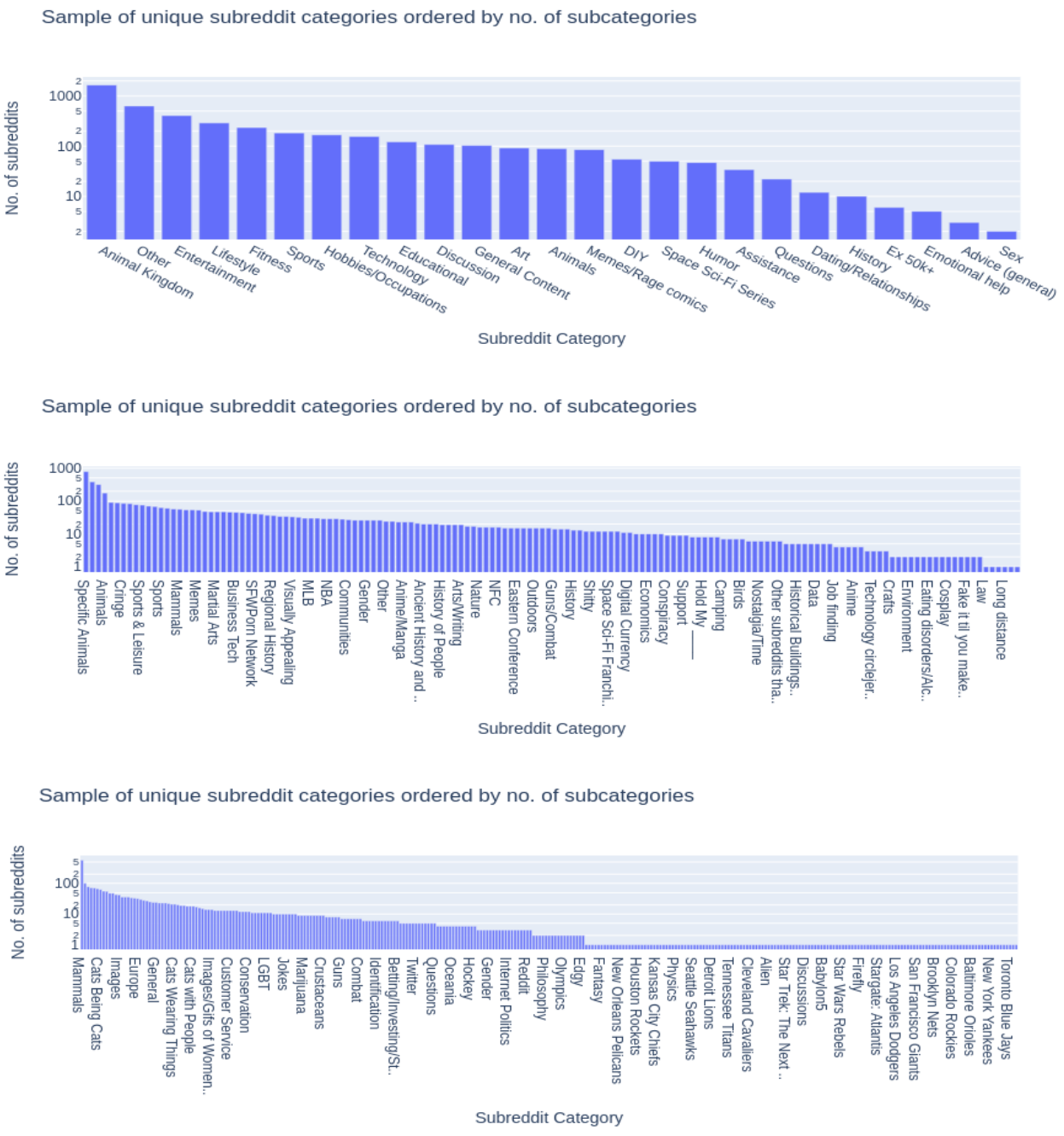

Figure 6. Subreddit Categorization