Network interventions to reduce hate speech on Nigerian Twitter

Keywords: Randomized Control, Experiment Design, Twitter, Interference, Networks

Extended Abstract

Online social networks allow unprecedented communication between communities and give a voice to marginalized groups. At the same time, social media platforms have created mechanisms that amplify extreme and potentially hateful content in ways that would not be possible with traditional media [1]. To counter these negative consequences, social media platforms invest in content moderation, which involves deleting hateful posts or suppressing accounts generating hateful content. There is mixed evidence on whether these punitive measures are effective, as they can backfire by incentivizing users to migrate to less moderated platforms. Recent work suggests that community-driven interventions, such as empathy-based counter-speech and celebrity messaging might be more effective in curbing hate speech by changing the norms around posting hateful language [2]. However, the effectiveness of these approaches at scale has not been tested. This study aims to evaluate the effectiveness of a community-driven ad-based intervention aimed at reducing hate speech on Nigerian Twitter at scale.

Experimental design: Our experiment aims to reduce overall engagement with hate content by targeting either the producers on the supply side or the consumers on the demand side. We will target producers or consumers with celebrity pro-social messages using Twitter ads featuring video clips and measure the number of hate tweets produced and consumed (liked or retweeted). Our goal is to compare the estimated treatment effects under each treatment arm over short and long term. The treatments will focus on the three major ethnic groups in Nigeria, namely Hausa-Fulani, Igbo, and Yoruba. Our dataset includes full timelines of approximately 2 million Twitter users from Nigeria with user ethnicity inferred using a combination of name-matching and label propagation on the followership network. We also created a hate speech detection algorithm using a pre-trained language model.

In randomized experiments in social networks, a unit's outcome depends on the treatment status of other units, a phenomenon often referred to as interference [3]. Thus, following [4], we assign treatment to clusters of units to minimize interference using a modified 3-net graph clustering algorithm modified that accounts for highly-skewed degree distributions. By constructing clusters of users that minimally interact with users outside their cluster and assigning the whole cluster to a treatment arm, we minimize exposure to different treatments outside the cluster, significantly reducing bias in our estimate for the average treatment effect (ATE) [5].

The ATE compares how much hate content is produced (or consumed) between the scenarios in which either every hate producer, or every consumer or no user is treated. We assume that the potential outcome of a unit depends on its treatment and the treatment of its immediate producer or consumer neighbors. In other words, we assume *interference only occurs through direct links of distance 1.* We use Hajek and Horvitz-Thompson with inverse propensity weighting and propensity-score matching estimators to measure treatment effect (including spill-overs) under two exposure models [4]: full exposure, where all neighbors must be treated, or q-fraction , where only $q = 70\%$ of neighbors in the ego-network should be treated for the ego to have the same potential outcome as one under a treatment vector where all producers (or consumers) are treated.

Results: The interaction networks in online social media, such as Twitter, tend to have highly heterogeneous degree distributions. This results in 3-net clustering outputs with disproportionate size distributions and large estimator variance. To overcome this, we tested several modifications to the 3-net clustering algorithm. For instance, we ensure the highest-degree users always fall inside a cluster rather than its periphery to avoid large weights that inflate variance. Nevertheless, unbiased estimators such as Horvitz-Thompson will have extremely large variances due to a few large degree users connected to many clusters. As such, the goal of our results here is to compare the distribution of the estimated ATE using different estimators through simulation.

For these simulations, we only present results on the producer treatment arm; results for consumer treatment are qualitatively similar. In these simulations, we assume spillovers occur through the *retweet network* of active users during 2022. Of the nearly 2 million users in our dataset, 60594 are hate-content producers. Our modified 3-net clustering algorithm leads to 1191 clusters of hate content producers, with a median size of 812 users per cluster, ranging from a minimum of 153 to a maximum of 28459 users (42% of all producers). In comparison, the largest cluster under the original 3-net algorithm leads to a more disparate cluster size distribution with the largest cluster of 46142 users (76% of all hate-content producers).

To conduct a meaningful power analysis before running the experiment, we conducted a pretest to select the most effective prosocial message for our ad campaign. Nigerian Twitter users were recruited to select the message (from among six candidate messages) that they thought would be most effective at reducing inter-ethnic hate speech. The pretest also helped us generate a rough estimate of a 20% reduction in the number of hate tweets produced per treated user. This initial estimate formed the basis of the outcome generation model. For this model, we assumed half of the 20% reduction in hate content production per user originates from the user's treatment, and the remaining 10% reduction occurs when all neighbors of the user are treated. Figure 1 compares the design-based distribution of the ATE using Hajek, Horvitz–Thompson, and propensity score matching estimators over 2000 simulations, along with a naive estimator which uses independent treatment assignment to individual users while ignoring interference. As expected, the naive estimator is heavily biased. While unbiased, the Horvitz-Thompson estimator suffers from a much larger variance than the Hajek and matching estimators, which do not seem to introduce excessive bias. The Hajek estimator and, particularly, the propensity score matching estimator offer a great improvement on the variance over the Horvitz-Thompson with a minimal penalty on the bias. Therefore, they will form the basis of our power analysis and estimation strategy in the field experiment.

References

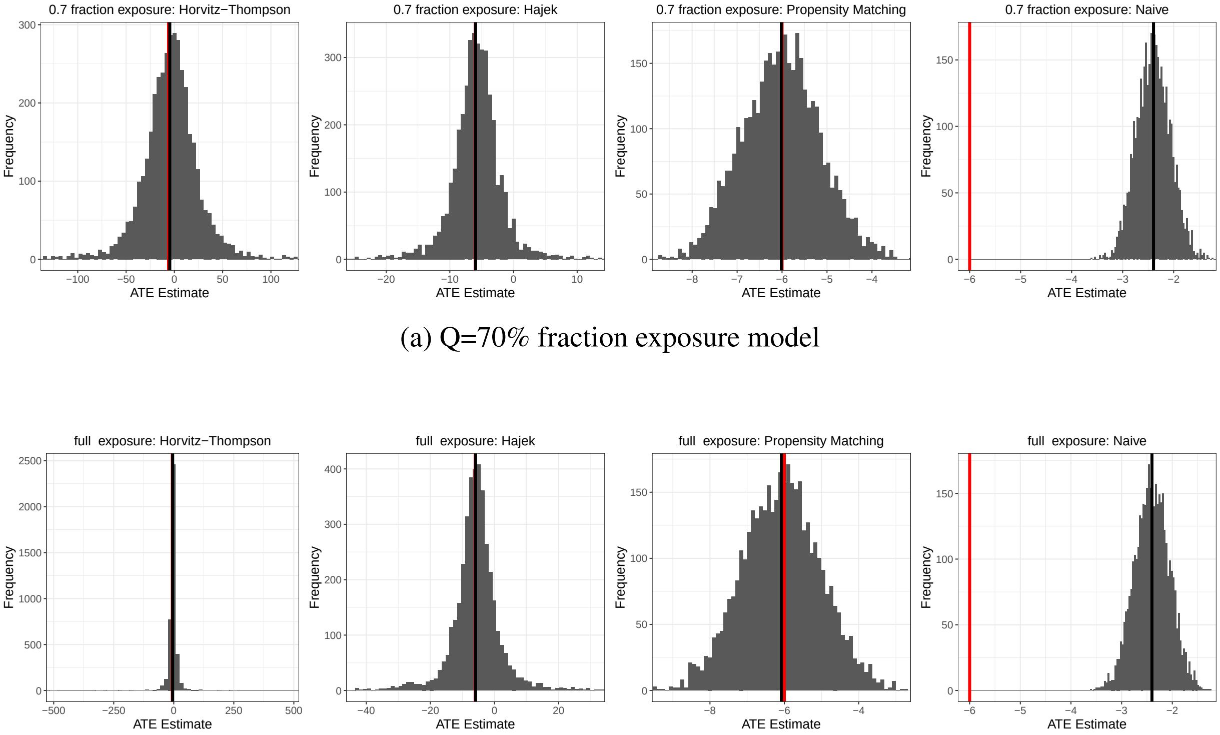
[1] Chris Bail. "Breaking the social media prism". In: *Breaking the Social Media Prism*. Princeton University Press, 2021.

[2] Deen Freelon. "Discourse architecture, ideology, and democratic norms in online political discussion". In: *New media & society* 17.5 (2015), pp. 772–791.

[3] Charles F Manski. "Identification of treatment response with social interactions". In: *The Econometrics Journal* 16.1 (2013), S1–S23.

[4] Johan Ugander et al. "Graph cluster randomization: Network exposure to multiple universes". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 329–337.

[5] Peter M Aronow and Cyrus Samii. "Estimating average causal effects under general interference, with application to a social network experiment". In: *The Annals of Applied Statistics* 11.4 (2017), pp. 1912–1947.



(a) Q=70% fraction exposure model

(b) Full exposure model

Figure 1: The comparison of four estimation methods under q-fraction (top row) and full exposure (bottom row) models. Columns correspond to different estimators with graph cluster randomization: Horvitz-Thompson (first column from the left), Hajek (second column), propensity score matching (third column) and the naive scheme with individual random assignment ignoring interference (fourth column). The red line corresponds to the true effect size based on the outcome generation model, and the black line corresponds to the average (expected value) of the estimator distribution.