Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit

Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, Eric Gilbert

General topic:
The authors are studying the effects of social media platform interventions which are "softer alternatives to outright bans." (Abstract)

The specific behavior or activity studied:
"We examine one such intervention called quarantining, that impedes direct access to and promotion of controversial communities. Specifically, we present two case studies of what happened when Reddit quarantined the influential communities r/TheRedPill (TRP) and r/The_Donald (TD)." (Abstract)

The specific research questions:
"RQ1: How were the participation levels within TRP and TD affected by the quarantine? RQ2: To what extent was the influx of new users to TRP and TD affected by the quarantine? RQ3: How was the use of misogynistic and racist language within TRP and TD, respectively, affected by the quarantine?" (1.2 Research Questions)

Challenges:
"Reddit does not provide an official list of all communities that have been quarantined. Therefore, we relied on Reddit search for finding community discussions about quarantining to identify the quarantined communities.". The authors also state: ".our focus on just two communities is a limitation..." (3 Data)

The paradigm that the research question is being asked in: Speech moderation in online communities is a fine line to walk on for social media platforms as moderation can often be construed as suppression. Softer bans, like quarantining by Reddit, are being utilized instead of outright banning subreddits. However, their effect, if any, needs to be studied.

Problem: Examining the effect of Reddit's quarantining policy on the quarantined community's behavior.

Importance: Social media platforms have different intervention methods to counter violations of that platform's policies. This study investigates the effectiveness of one such intervention method.

Claims: Authors claim that the application of quarantine on a subreddit changed its posting activities, reduced recruitment of new members but saw no change in hate speech usage.

State of Knowledge: Quarantining a subreddit, unlike banning, still allows subreddit member participation. The effects of quarantining a subreddit has not been delved into in prior studies.

Evidence: The authors present theoretical evidence by utilizing Reddit posts. From the TRP and TD subreddits, all the submissions and comments during 2018-2019 were collected since the subreddits were quarantined in 2018 and 2019 respectively.

Story Structure: The authors explained the intervention method used by Reddit and outlined the research questions they aim to answer centered around the intervention. Then they outlined
the methods to be employed for investigating each question. Finally, they delineated the causal inference strategies to be applied to the control and treatment subreddits.

Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald
Amaury Trujillo, Stefano Cresci

General topic:
The authors are studying the phenomenon of when "a community suffers a moderation intervention, the behavior of the members of that community might change not only within that community, but also in other communities in which they participate." (Abstract)

The specific behavior or activity studied:
The authors are specifically studying the change in community behavior in the r/The_Donald subreddit with respect to the three interventions it received: quarantine, restriction and ban.

The specific research questions:
"RQ1: What were the effects of the quarantine, in terms of activity and toxicity, within r/The_Donald?
RQ2: What were the effects of the quarantine, in terms of the quality of shared news articles, within r/The_Donald?
RQ3: What were the effects of the sequence of interventions applied to r/The_Donald, on the other communities to which r/The_Donald members participated?"

Challenges:
"...[O]ur analyses are solely focused on the moderation interventions enforced to r/The_Donald and, as such, might suffer from selection bias and lack of generality". In addition, "inherent limitation when estimating causal effects with observational data is the susceptibility of the measured quantities to possible exogenous causes (i.e., confounders)", "our results likely provide only a partial view of the full extent of effects caused by the interventions issued on r/The_Donald", and "a community-level post-intervention effect is the combination of many and potentially diverse user-level effects. Hence, the aggregated community-level effect might be weakly representative of the underlying behavior of individuals or smaller user groups." (6.6 Limitations and Future Work)

The paradigm that the research question is being asked in: Social media platforms have in recent years employed different intervention methods to curb misinformation and misbehavior online. However, some of these interventions have not been well thought out and their application has produced adverse reactions. The authors of this paper study the misbehavior as well as misinformation being spread as a result of the various kinds of subsequent interventions (quarantine, restriction and ban) Reddit imposed on r/The_Donald subreddit.

Problem: The effects of Reddit’s interventions need to be studied on the subreddit itself as well as the ripple effects it could cause in other subreddits. In addition, interventions cannot be studied in isolation only, the sequence of the interventions also need to be taken into account.
**Importance:** Social media interventions are required to be effective but can often have unintended consequences. The results of such interventions need to be studied in order for better design of interventions

**Claims:** The quarantine intervention decreased user activity, increased toxicity, decreased factual reporting, had no effect on political polarization and the core users of the subreddit (r/The_Donald) were affected more than others. Moreover, the policy of using restriction was more effective than quarantine.

**State of Knowledge:** Prior work did not take into account the effects of moderation on political polarization and factual reporting and have mainly focused on user activity and toxicity including hate speech. In addition, prior work has focused on the quarantining and subsequent ban of the r/The_Donald subreddit in isolation and not in a sequential manner.

**Evidence:** Reddit posts are used as evidence from the timeframe centered around the time the quarantine intervention was imposed. The Reddit posts were obtained from the quarantined subreddit as well as other subreddits.

**Story Structure:** The authors first explained the timeline and the types of interventions applied by Reddit on the particular subreddit being examined. Then they explained their choice of core users from the subreddit and the methods used for the causal analysis. Finally, they explained the effects of the interventions on the three aspects targeted in the research questions.