9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

The Cross-platform Consequences of Deplatforming on Social Media

Keywords: Social media, deplatforming, cross-platform, toxicity, polarization

In an era where social media plays a crucial role in shaping our democratic life, the regulation of online content has become a hotly debated topic. Deplatforming, or the removal of malicious accounts from mainstream social media, has been seen as a way to counter misinformation and protect users. However, little consideration has been given to how such moderation policies impact the wider social media ecosystem. In part, this is because few studies have been able to track social media users after being suspended from mainstream platforms.

This paper takes a step towards addressing this gap, presenting a comprehensive study of the ban-induced platform migration from Twitter to Gettr on the US political right. Our dataset covers the near-complete evolution of Gettr from its founding in July 2021 to May 2022 including 15 million posts from 790 thousand active users ("Not verified" cohort in figures). Of these users 6,152 are verified, 1,588 of which self-declare as active on Twitter ("Matched" cohort). For these 1,588 self-declared Twitter users with a verified Gettr account, we download their Twitter timeline from July 2021 to May 2022 totalling 11 million posts ("Twitter" cohort). For other verified Gettr users, we use the Twitter API to identify accounts who have been suspended by Twitter, totalling 454 accounts ("Banned cohort").

Analyzing these cohorts, we first show that retention of migrated users varies significantly: users banned from Twitter are more active (between 3–6 times), and remain on Gettr for significantly longer, than users who have not been suspended from Twitter. Second, we analyse the structure and content of Gettr, demonstrating that there is little difference between those users suspended on Twitter, and those still active on Twitter, see figure 1(a); both groups, and Gettr in general, are broadly representative of the US far-right. Third, we show that Gettr content is, on average, significantly more toxic than the Twitter content of matched users. However, the most toxic Twitter content is more toxic than on Gettr, see figure 1(b/c). To better understand this phenomenon, we focus on the behaviour of the matched cohort on Twitter and study the Twitter users with whom they interact, see figure 2. We define the "quote-ratio" as the number of matched accounts who quote-tweet a user on Twitter, normalised by all mentions. This shows that the cohort of matched users on Gettr broadly align with far-right media sources and differ significantly from left-leaning groups. Studying the toxicity of posts on Twitter, we find that posts who mention users who are disproportionately quote-tweeted than retweeted are significantly more toxic than other posts. We find that many of these posts target US Democrat politicians, and specifically female politicians. Finally, we discuss the broader impact of fringe platforms on democracy by showing the impact of Gettr on the Brasilia insurrections in January 2023. Together, these results emphasise the critical importance of carefully considered social-media moderation policies, and the need for high quality data on, and monitoring of, fringe social-media platforms.

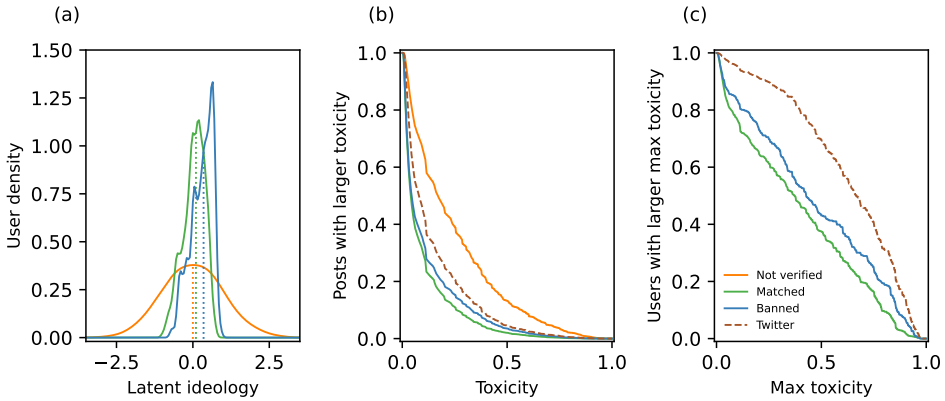

Figure 1: **The latent ideology of Gettr users, the toxicity of posts on Gettr and Twitter, and the maximum toxicity of each user on Gettr and Twitter.** (a) The latent ideology is calculated using interactions from the "not verified" cohort with the "banned" and "matched" cohorts. All three distributions are unimodal when tested using Hartigan's diptest (multimodality not statistically significant for not-verified group, $p = 0.99 > 0.01$, matched group, $p = 0.49 > 0.01$, and banned group, $p = 0.91 > 0.01$). Structural data required for calculating the latent ideology for Twitter is not available. (b) The fraction of posts from each user cohort on Gettr (and matched Twitter) with a toxicity value larger than the value shown on the x-axis. Toxicity is calculated using the Google Perspective API. (c) The distribution of the most toxic posts for each user from each cohort on Gettr and Twitter.

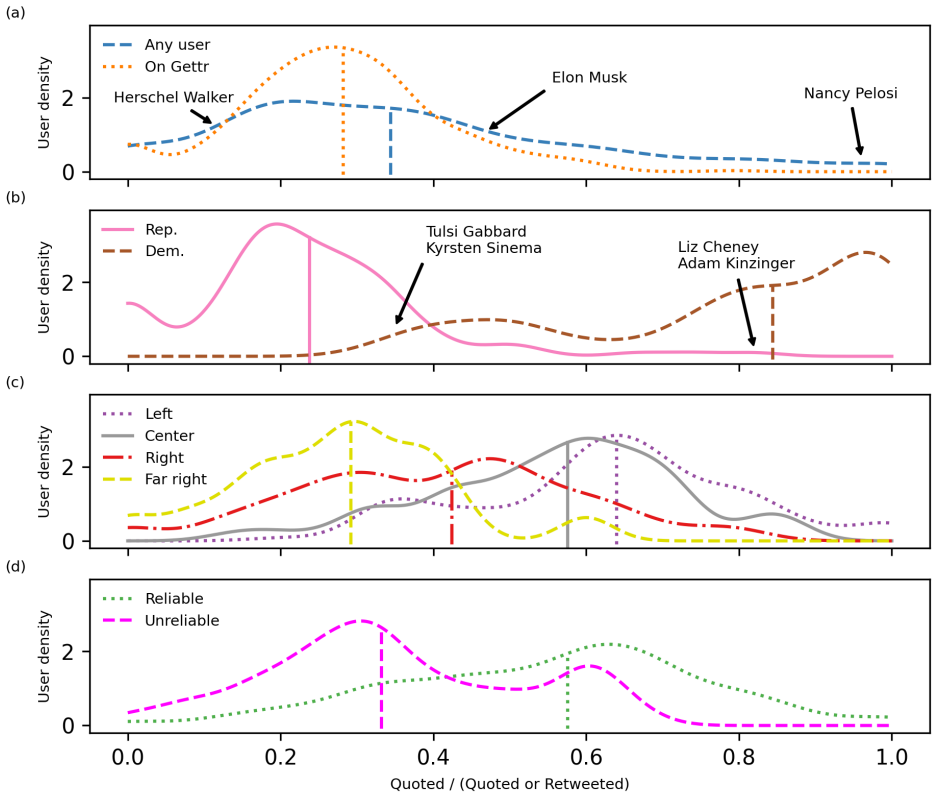

Figure 2: **The distribution of the quote-ratio of accounts mentioned on Twitter by the matched cohort.** The quote-ratio is defined as the number of matched users who quote-tweet an account, normalised by the number of matched users who quote-tweet or retweet the same account. Only Twitter accounts who are mentioned by at least five unique matched accounts are included. (a) The quote-ratio distribution for all mentioned accounts (blue dashed), and for mentioned accounts who are part of the matched cohort of users (i.e., a matched user mentioning another matched user; orange dotted). (b) The quote-ratio distribution for Twitter accounts belonging to known elected US Republican (pink solid) and known elected US Democrat (brown dashed) politicians. (c) The quote-ratio distribution for Twitter accounts belonging to news media organisations who have been labelled with a political leaning by MediaBias/FactCheck. Organisations are classified as left (purple dotted), least-biased (grey solid), right (red dot-dashed), or far-right (yellow dashed). (d) The same news media organisations, but broken down according to whether they are classified as a reliable or unreliable. Vertical lines mark the median of each distribution. Annotations indicate users of particular interest: in panel (a) the accounts with the smallest and largest quote-ratios ($>$ 100 mentions), Herschel Walker and Nancy Pelosi respectively, and the account with the most absolute mentions, Elon Musk. In panel (b), we label outliers from each political party.