Behavioral classes of radicalization within the QAnon conspiracy on Twitter

QAnon, radicalization, social networks, Twitter,

Extended Abstract

Social media platforms are becoming increasingly instrumental in the spread of fringe conspiracy theories, which often originate from close-knit, niche online communities and eventually gain traction in mainstream media. While some conspiracy theories are not harmful, others create distrust towards institutions and, in turn, result in violence offline [5, 6].

One prominent example in recent times is QAnon, a conspiracy supporting a range of extreme and uncorroborated ideas, e.g., the "Pizzagate" and the "Obamagate" conspiracies, and many more recent conspiracies surrounding the COVID-19 pandemic and the 2020 US Presidential election [3, 7]. Beyond political influence, QAnon support escalated into real-world acts of violence such as the January 6th, 2021 attack on the US Capitol, where several QAnon followers participated. It has thus become increasingly important to understand how individuals may become radicalized online, starting from social media platforms and later shifting to real-world action [4].

In this work, we identify and characterize a diverse set of behaviors tied to users' engagement with QAnon theories on Twitter, by proposing a set of continuous metrics to capture signals of radicalization. To this end, we leverage a dataset of over 240 million election-related tweets collected in the run-up to the 2020 US Presidential election, in a period that captures Twitter ban on QAnon (July 2020). We measure signals of radicalization in users' tweets and profiles as well as through their social connections.

Our contribution is manifold. We provide a framework for measuring users' signals of radicalization within the QAnon conspiracy, looking both at content production and social interactions. We construct a suite of four continuous metrics of radicalization, which can separate different aspects of engagement with QAnon and are agnostic to the platform and group under analysis, thus allowing for generalizability to other scenarios and platforms.

Our findings suggest that radicalization processes should be modeled across multiple dimensions, as one single dimension cannot capture the complex facets of radicalization (see Figure 1). Apply clustering techniques, we discover six distinct behavioral classes of radicalization, each associated with different behaviors (see Figure 2). The main archetypes of radicalized users include conspiracy amplifiers, self-declared supporters, and hyper-active promoters. Besides, the three most radicalized classes comprise a significant proportion ($\sim$9.4%) of the users in the dataset under analysis.

We observe that users in the most radicalized classes tend to share fewer reliable URLs and are significantly more likely to be suspended by Twitter. Hyper-active promoters are most persistent in sharing QAnon content over time and engage with a large variety of QAnon topics. Amplifiers and hyper-active promoters tend to rebroadcast content originating from their own groups significantly more than expected (see Figure 3). Comparing the interactions between QAnon accounts and other users, before and after Twitter intervention against QAnon, we find that engagements with hyper-active QAnon promoters are reduced substantially, whereas QAnon amplifiers were not significantly affected by Twitter moderation.

The results of this paper convey a two-pronged message. On the one hand, we observe that mitigation strategies adopted by social media platforms can be effective to quell hyperactive users spreading conspiracies. On the other hand, we show that less evident radicalized behaviors can escape moderation, leading to the persistence of problematic content on online platforms. Based on these premises, our findings can inform social media providers, regulators, and policymakers to formulate strategies to counter the circulation of conspiracy theories and fringe narratives on social media.

Our findings show that radicalization within a given conspiracy on a social media platform can generate diverse behaviors, which should be observed through the lens of a heterogeneous set of indicators. The signals of radicalization presented in this work are a primary example of the diverse facets that can be captured to model radicalization processes. Indeed, our methodological framework is not comprehensive and can be augmented by encompassing a larger variety of metrics. This work represents a first building block in the large-scale modeling of radicalization within fringe communities and can pave the way to research validating and augmenting our methodology. In the future, we aim to develop new metrics quantifying beyond the “Network” stage proposed in radicalization theories, i.e., “Need” and “Narrative”. These metrics may allow us to include factors as anxiety, isolation, and anger, thus enhancing the identification of topics and opinions in users’ messages.

We also note that content moderation and user bans appear to reduce conspiracy content, but they may be counterproductive, leading users to migrate to platforms with lower diversity of ideas and no moderation against problematic content [1, 2]. To moderate divisive issues, actions like diversifying content and user recommendations may be more far-reaching, effective, and fair.

References

[1] S. Ali, M. H. Saeed, E. Aldreabi, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini. Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pages 187–195, 2021.

[2] M. Cinelli, G. Etta, M. Avalle, A. Quattrociocchi, N. Di Marco, C. Valensise, A. Galeazzi, and W. Quattrociocchi. Conspiracy theories and social media platforms. *Current Opinion in Psychology*, page 101407, 2022.

[3] E. Ferrara. What types of covid-19 conspiracies are populated by twitter bots? *First Monday*, 25(6), 2020.

[4] M. Hannah. Qanon and the information dark age. *First Monday*, 2021.

[5] D. A. Stecula and M. Pickup. Social media, cognitive reflection, and conspiracy beliefs. *Frontiers in Political Science*, 3:647957, 2021.

[6] J. Tollefson. Tracking qanon: how trump turned conspiracy-theory research upside down. *Nature*, 590(7845), 2021.

[7] K.-C. Yang, F. Pierri, P.-M. Hui, D. Axelrod, C. Torres-Lugo, J. Bryden, and F. Menczer. The covid-19 infodemic: Twitter versus facebook. *Big Data & Society*, 8(1):20539517211013861, 2021.

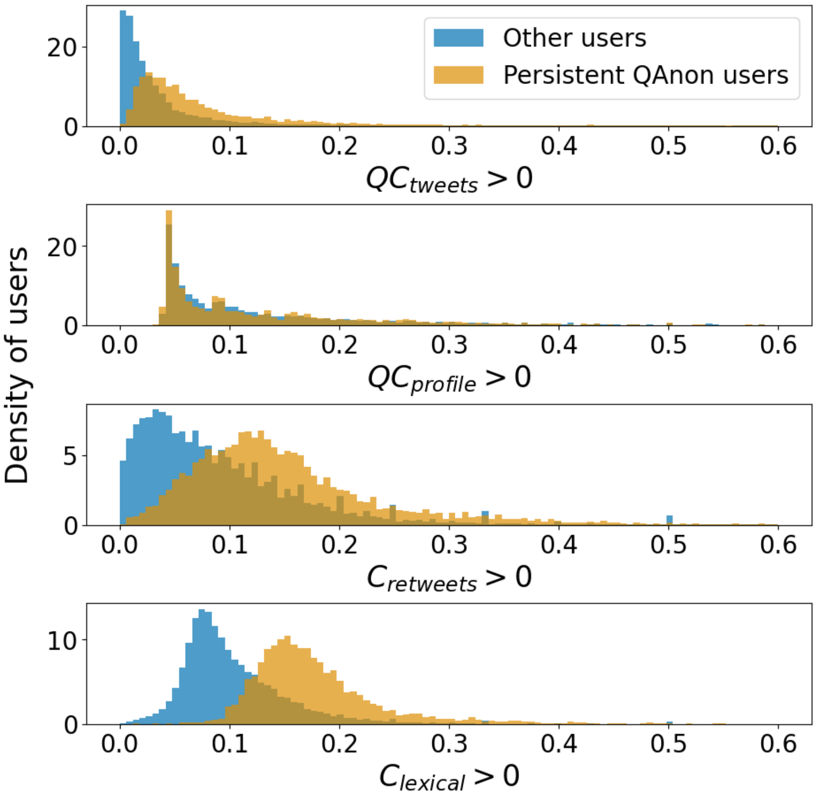

Figure 1: Distribution of the four metrics (QC_{tweets}, $QC_{profile}$, $C_{retweets}$, and $C_{lexical}$) for *persistent QAnon users* and other users. Only values > 0 are plotted. *Persistent QAnon users* were identified as those users that produced a considerable amount of original QAnon content and were persistently active in the weeks before the ban of Twitter on Qanon.

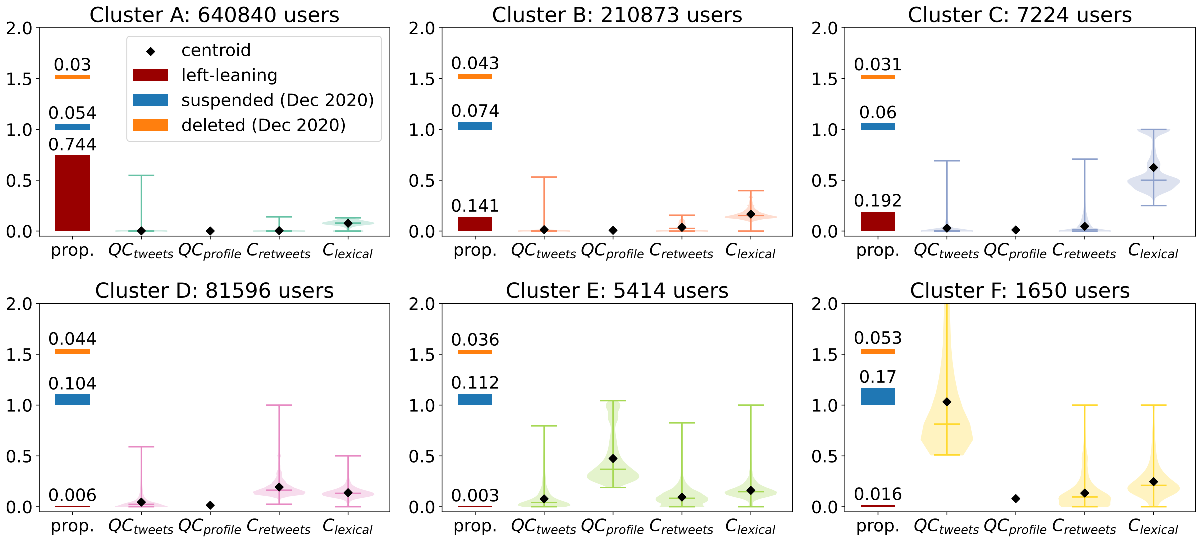

Figure 2: Distribution of the four metrics (QC_{tweets}, $QC_{profile}$, $C_{retweets}$, and $C_{lexical}$) for each cluster. The proportion of left-leaning, suspended, and deleted accounts within each cluster is also reported. Diamonds correspond to the values of the centroid in each cluster. Horizontal lines of the violin plot correspond to the minimum, median, and maximum values.

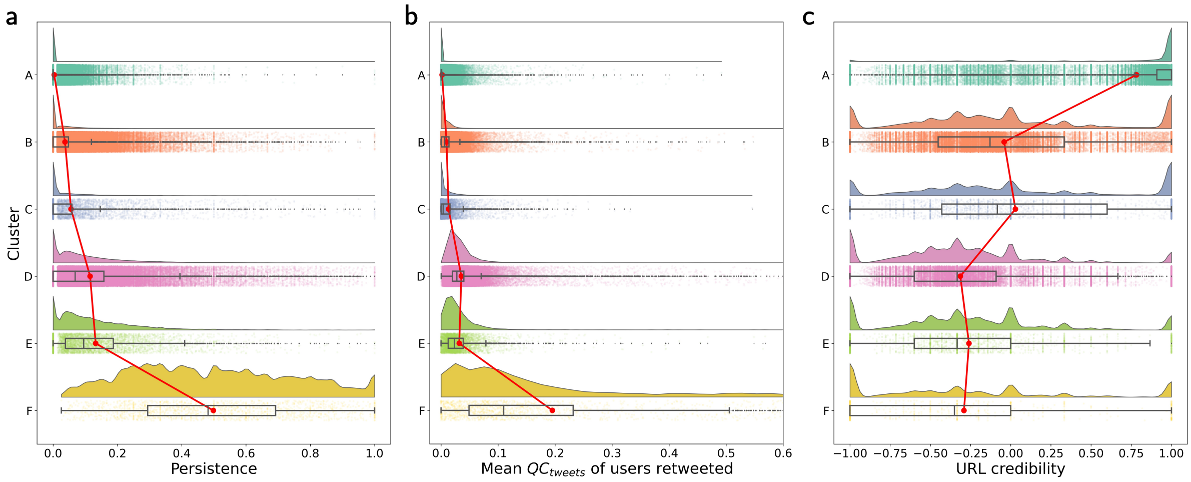

Figure 3: Clusters characteristics in terms of (a) Persistence, (b) Mean QC_{tweets} of retweeted users, (c) URL credibility