Title: Robustness to deduplication and referral specifications appendix for "Research note: Examining potential bias in large-scale censored data"
Authors: Jennifer Allen (1), Markus Mobius (2), David M. Rothschild (2), Duncan J. Watts (3,4,5)
Date: July 26th, 2021
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix 3: Robustness to deduplication and referral *specifications*

A) Deduplication

One distinction between the Facebook URLs dataset and the Nielsen dataset is that the Facebook URLs dataset deduplicates engagement such the dataset only records a single action per user per URL, regardless of whether the user shared, clicked, or viewed the URL multiple times, whereas the Nielsen dataset does not deduplicate and tracks all clicks to the same URL by a given URL. Thus, the Facebook data is undercounting some of the engagement on URLs in the Facebook URLs dataset, which could affect the way that these numbers are interpreted (e.g., if fake news has a much higher click or share count in reality due to inauthentic activity or click farms).

For robustness, we decided to examine the number of repeated clicks in the Nielsen dataset. 93% of URLs from Facebook are clicked only one time per user. And, most critically, almost all (>99%) of the double-clicks are on non-news (neither fake or real) URLs and are to "evergreen" links gaming or weather. This does not affect the overall results, and convinces us that the reverse issue of Facebook missing sharing of fake news by only counting unique shares, is not an issue. We also redid our primary analysis and found no difference when using the deduplicated URLs (see Appendix 3C).

B) Referral specifications

One potential issue with our analysis is in the way we defined clicks from Facebook as URLs that contained the "fbclid" parameter. However, it is possible that URLs with the "fbclid" parameter are actually URLs that have been copy-pasted and sent to the user via different platforms (e.g., email) and not just Facebook. In order to assess the impact that this definition of Facebook clicks had, we examined the Nielsen browning data and found that 60.6% of URLs with "fbclid" in the title are immediately preceded by a click from "facebook.com". We suspect the clicks not immediately preceded by "fbclid" are more likely due to browser / tab switching away from facebook.com and then back to an earlier tab without reloading rather than users clicking to an "fbclid" from an email or messaging site, since we do not see any consistent pattern of URLs with "fbclid" coming from email or messaging clients.

We also redid our primary analysis and found no difference when using a stricter definition of Facebook clicks (see Appendix 3C).

C) Repetition of main analysis

We re-calculated the percentage of fake news, not-fake news, and real news on Facebook, based on the issues raised above. In particular, we 1) de-duplicated clicks from Facebook at the user-URL level and 2) restricted to the set of URLs that contained the "fbclid" parameter and were immediately preceded by a visit to "facebook.com." We found almost identical results to our main text, with 2.6% of clicks going to fake news, 24% of clicks going to not-fake news, and 73% of clicks going to non-news. This is almost identical to the 2.5% fake news, 24% not-fake news, and 74% not-news reported in our main text.