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

Disentangling algorithms and user preferences with counterfactual bots

Keywords: recommendation algorithm, socio-technical systems, randomized experiments

Extended Abstract

Sociotechnical systems have raised puzzling questions about whether what we observe on platforms such as Instagram or YouTube is the result of the platform's design, i.e., its algorithms, or a reflection of society as seen through the lens of that platform. It is difficult to determine which is more likely, but when it comes to technology's disturbing consequences, it is easy to overlook the role that human behavior plays. What is the extent to which recommendations influence engagement? This study seeks to disentangle the role of recommendation algorithms from the intentions of users. Specifically, we design experiments utilizing logged in accounts (bots) that mimic user behaviors, personalized by following real YouTube viewership trajectories.

A commonly held perception is that recommender systems systematically lead users to increasingly ideologically biased content, yet previous research on the subject has not reached consensus on the matter. On the one hand, audits simulating user behavior on YouTube [1, 2] have found that, as users blindly follow the recommender system, they indeed receive recommendations that are increasingly ideologically biased. On the other hand, studies with real user traces show that the consumption of highly partisan content on YouTube is driven by a combination of user preferences and platform features [3, 4]. Ref. [3], for instance, indicates that far-right content is often accessed through external websites and not in the end-tail of long, recommendation-driven sessions; and Ref. [4] has further found that subscriptions (i.e., when users ask to receive updates about a YouTube channel) play a big role in how users consume extreme content on the platform.

A possible explanation for these seemingly contradictory findings is that existing audits of YouTube's recommender system do not meaningfully disentangle the role of the algorithm from user preferences, e.g., a majority of YouTube users might prefer right-leaning content. Using bots, i.e., automated programs that simulate user behavior on YouTube, we estimate the influence of recommender systems on what users consume by contrasting recommendations obtained when bots follow real user behavior (which we obtain from user traces) and when they blindly follow algorithmic recommendations. Unlike previous work, we are able to compare the content users are recommended by randomly following the recommender system with a meaningful counterfactual – the content they have consumed while surfing on the Web (and, naturally, interacting with the algorithm according to their own preferences). Specifically, we trained a set of four bots to watch real user histories for $N_{train} = 60$ videos (the personalization phase), while three bots continue watching $N_{heldout} = 60$ videos based on a predefined rule and one continues watching based on the real user's trace. The experiments are conducted on ten users representing different news viewership archetypes, and we repeat each experiment 20 times per user. We observe that in all experiments the highest concentration is around the center. The recommended videos are more moderate when bots are guided by one of the algorithmic paths, Fig. 1 and 2.

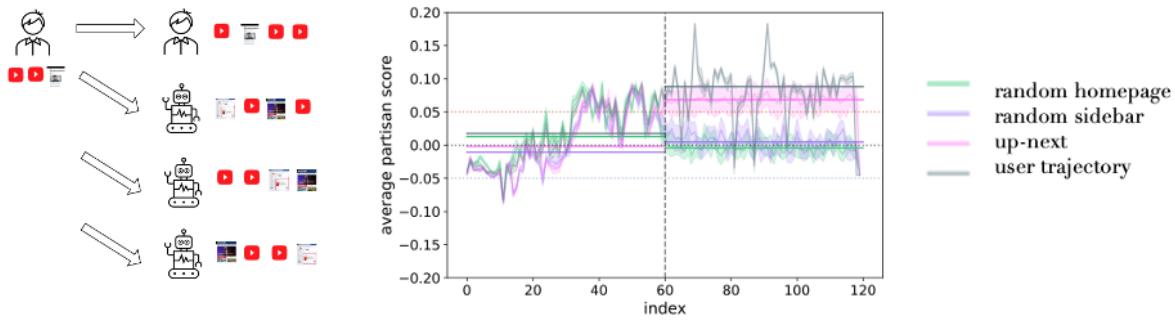

Figure 1: Y-axis shows the average partisan scores (over 20 iterations) of videos collected (after personalization) from the sidebar recommended videos for bots who followed (i) the user trajectory, (ii) the first video in the sidebar recommendations (upnext), (iii) a random video from the sidebar recommendations, and (iv) a random video from the homepage recommendations. On average, the partisanship of recommended items is the highest for bots who follow the user trajectory, and the least for bots who choose from either homepage or sidebar recommended items. This result is for one user trajectory, with high consumption of partisan right content in the personalization phase.

References

[1] M. A. Brown, J. Bisbee, A. Lai, R. Bonneau, J. Nagler, and J. A. Tucker, "Echo chambers, rabbit holes, and algorithmic bias: How youtube recommends content to real users," *Available at SSRN 4114905*, 2022.

[2] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak, "Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations," *arXiv preprint arXiv:2203.10666*, 2022.

[3] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, and D. J. Watts, "Examining the consumption of radical content on youtube," *Proceedings of the National Academy of Sciences*, vol. 118, no. 32, p. e2101967118, 2021.

[4] A. Y. Chen, B. Nyhan, J. Reifler, R. E. Robertson, and C. Wilson, "Subscriptions and external links help drive resentful users to alternative and extremist youtube videos," *arXiv preprint arXiv:2204.10921*, 2022.

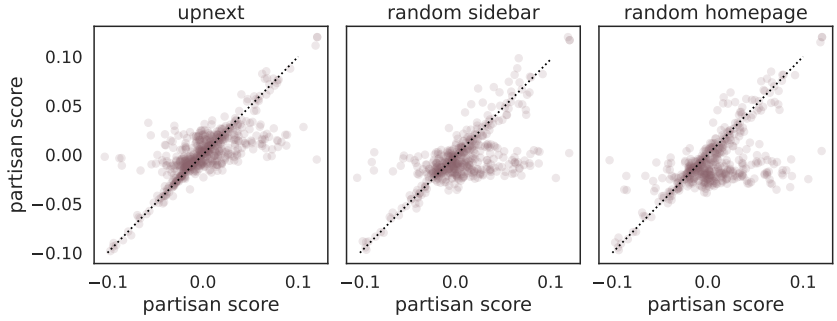

Figure 2: On the Y-axis are the partisan scores of videos collected (after personalization) from the bots who followed (i) the first video in the sidebar recommendations (upnext), (ii) a random video from the sidebar recommendations, and (iii) a random video from the homepage recommendations. A partisan score is shown on the X-axis based on the videos collected from sidebar recommendations of bots that continue to follow the exact user trajectory after being personalized.