Supplementary Information For Do Users Adopt Extremist Beliefs from Exposure to Hate Subreddits?

Matheus Schmitz[1], Goran Muric[1], Daniel Hickey[2] and Keith Burghardt[1]

[1]Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, 90292, CA, USA.
[2]Department of Botany and Plant Pathology, Oregon State University, 2701 SW Campus Way, Corvallis, 97331, OR, USA.

Contributing authors: mschmitz@isi.edu; gmuric@isi.edu; hickeyda@oregonstate.edu; keithab@isi.edu;

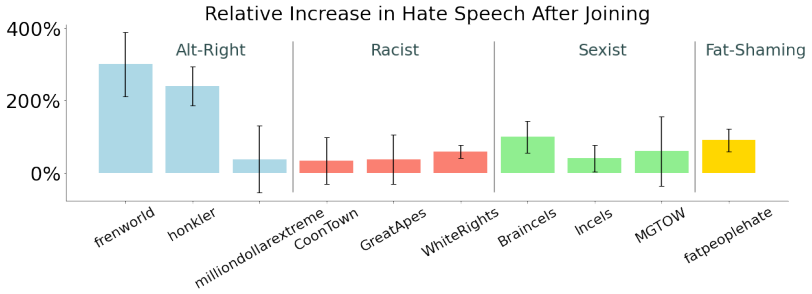

Fig. S1 Relative increase in the rates of hate speech immediately after users become active in a hate subreddit, as obtained from the interrupted time series model. Data is split by the subreddit type (alt-right, racist, sexist, and fat-shaming). Error bars are 95% confidence intervals in the mean.

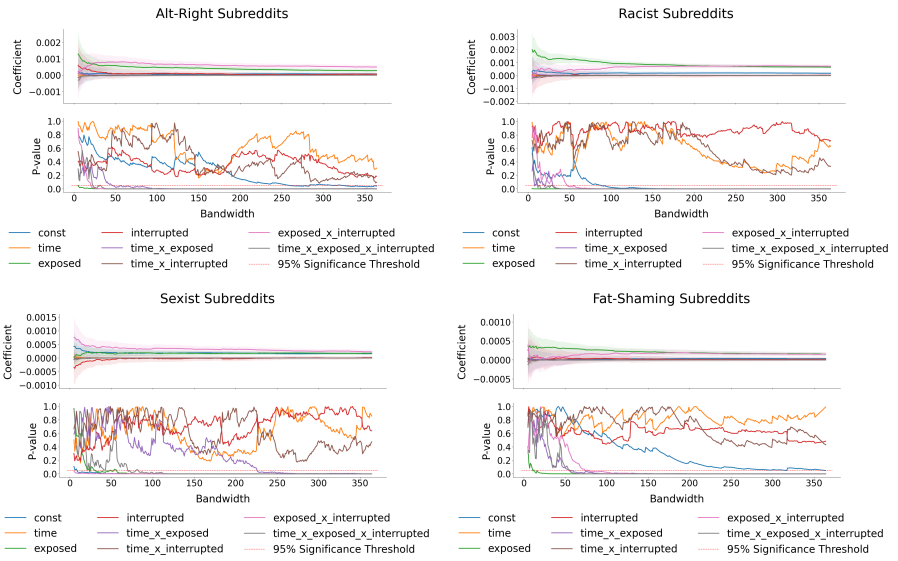

Fig. S2 Robustness of coefficients as a function of bandwidth for data aggregated to the subreddit category: alt-right, racist, sexist, and fat-shaming subreddits. Top row of each subfigure are coefficients as a function of bandwidth. Shaded area are standard errors. Bottom row of each subfigure are the significance (p-value) of each coefficient. Larger bandwidths allow for more data to be used, increasing confidence on the estimates (i.e. generally decreasing p-values), at the cost of considering potentially less relevant data points further away chronologically from the event of interest.

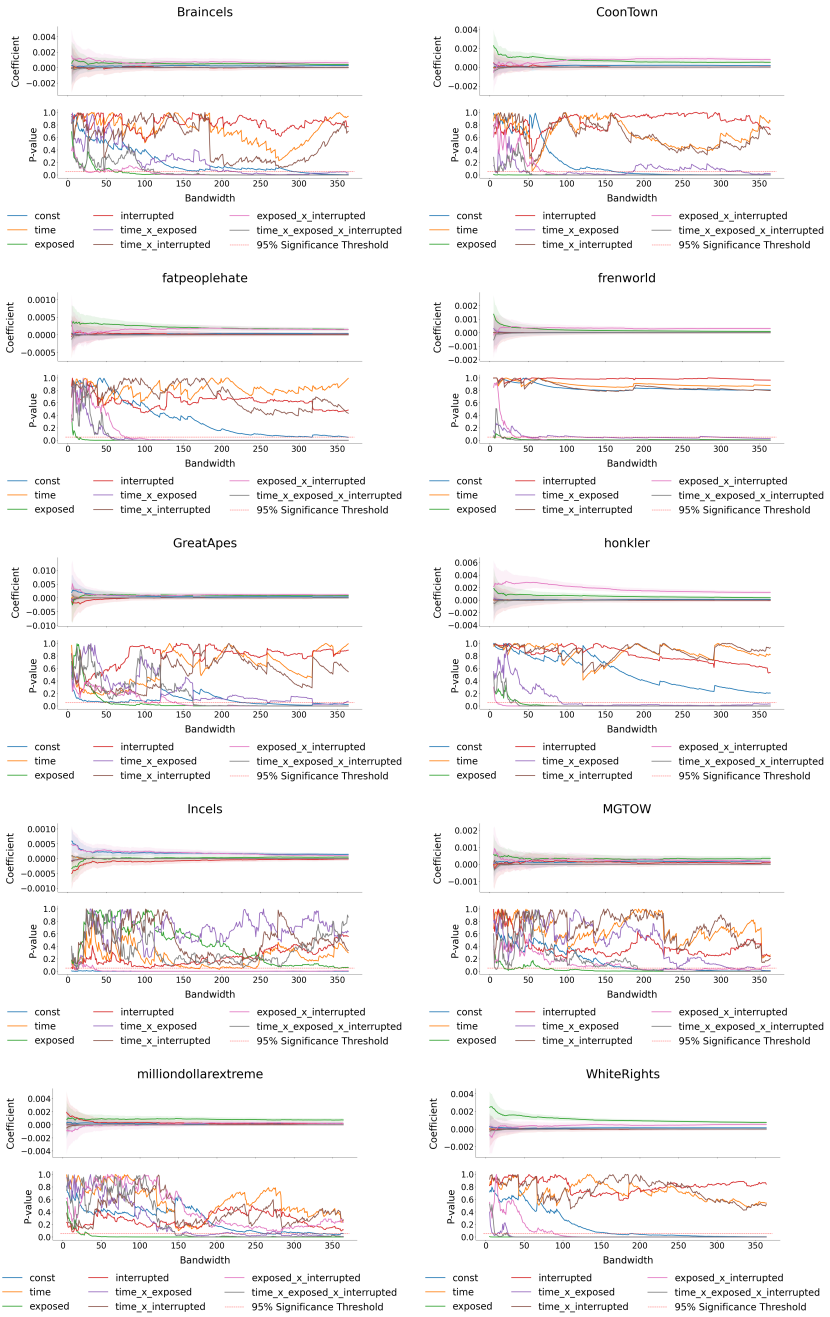

Fig. S3 Robustness of coefficients as a function of bandwidth for each subreddit. The subreddits are, in order, r/braincels (sexist), r/coontown (racist), r/fatpeoplehate (fat-shaming), r/frenworld (alt-right), r/greatapes (racist), r/honkler (alt-right), r/incels (sexist), r/MGTOW (sexist), r/milliondollarextreme (alt-right), r/whiterights (racist). Top row of each subfigure are coefficients as a function of bandwidth. Shaded area are standard errors, while the red line represents p-values equal to 0.05. Bottom row of each subfigure are the significance (p-value) of each coefficient.

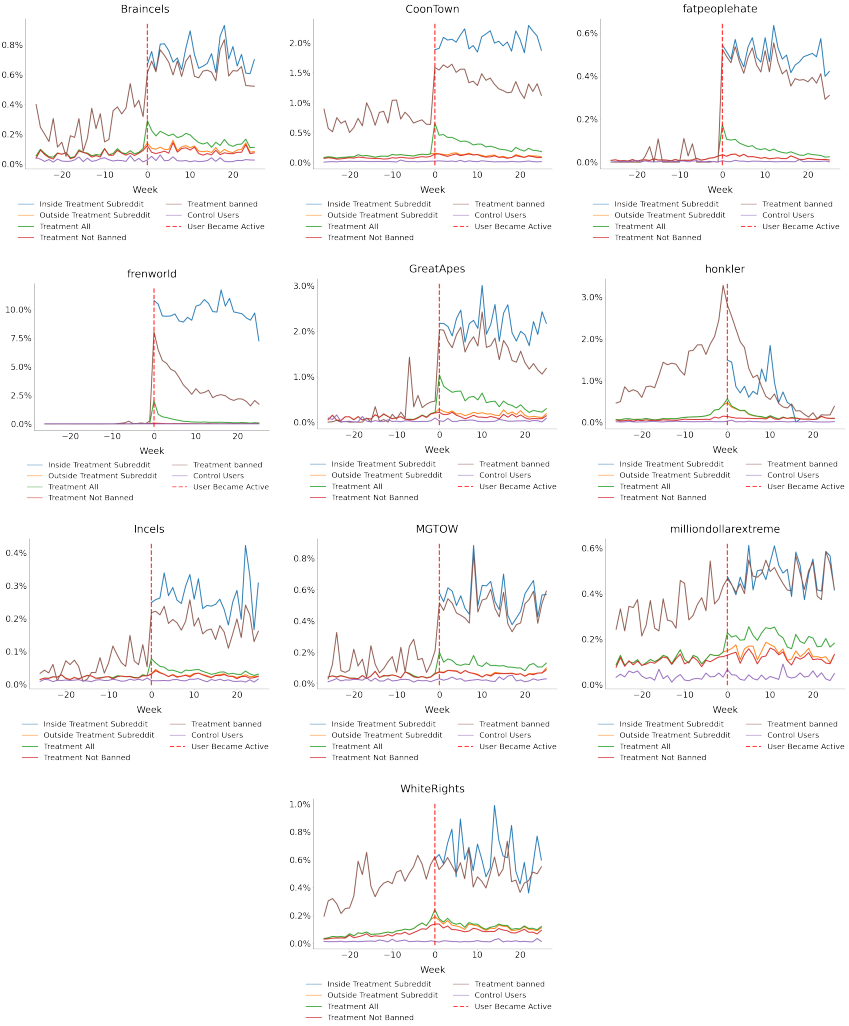

Fig. S4 Percent of hate speech over time. We plot the rate of hate speech over time (percent of words that are hate words) for: treatment users within the hate subreddit (blue), treatment users outside the treatment subreddit (orange), all posts for treatment users (green), non-banned subreddits for treatment users (red), banned subreddits for treatment users (brown), and control users (purple). The orange line in this plot is the same as the dots shown in Supplementary Figure S6. The red dashed line is when treatment users become active. The subreddits are, in order, r/braincels (sexist), r/coontown (racist), r/fatpeoplehate (fat-shaming), r/frenworld (alt-right), r/greatapes (racist), r/honkler (alt-right), r/incels (sexist), r/MGTOW (sexist), r/milliondollarextreme (alt-right), r/whiterights (racist).

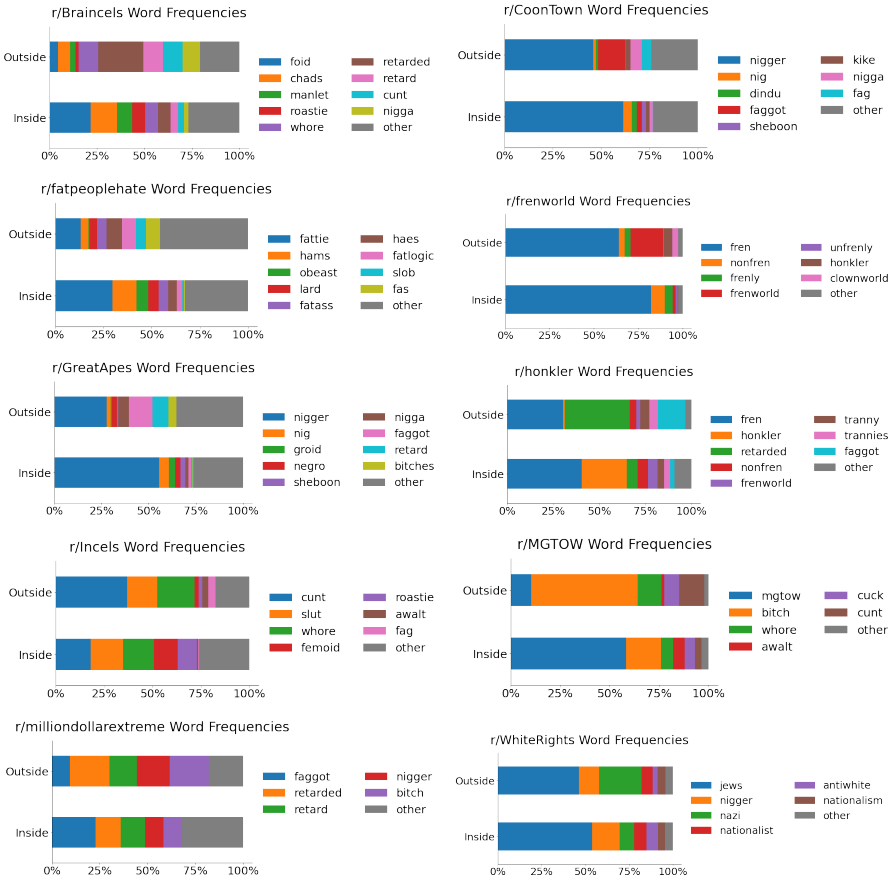

Fig. S5 Relative frequency of words for each subreddit. The subreddits are, in order, r/braincels (sexist), r/coontown (racist), r/fatpeoplehate (fat-shaming), r/frenworld (alt-right), r/greatapes (racist), r/honkler (alt-right), r/incels (sexist), r/MGTOW (sexist), r/milliondollarextreme (alt-right), r/whiterights (racist). Within each subfigure, "outside" refers to hate speech used outside of the hate subreddit, while "inside" refers to hate speech used within the subreddit. We explicitly label the most common words. Qualitatively a relative decrease in group lingo can be observed when treatment users move "outside", for example, an approximate four-times decrease in the usage of "foid" is observed when members of r/Braincels communicate outside the subreddit.

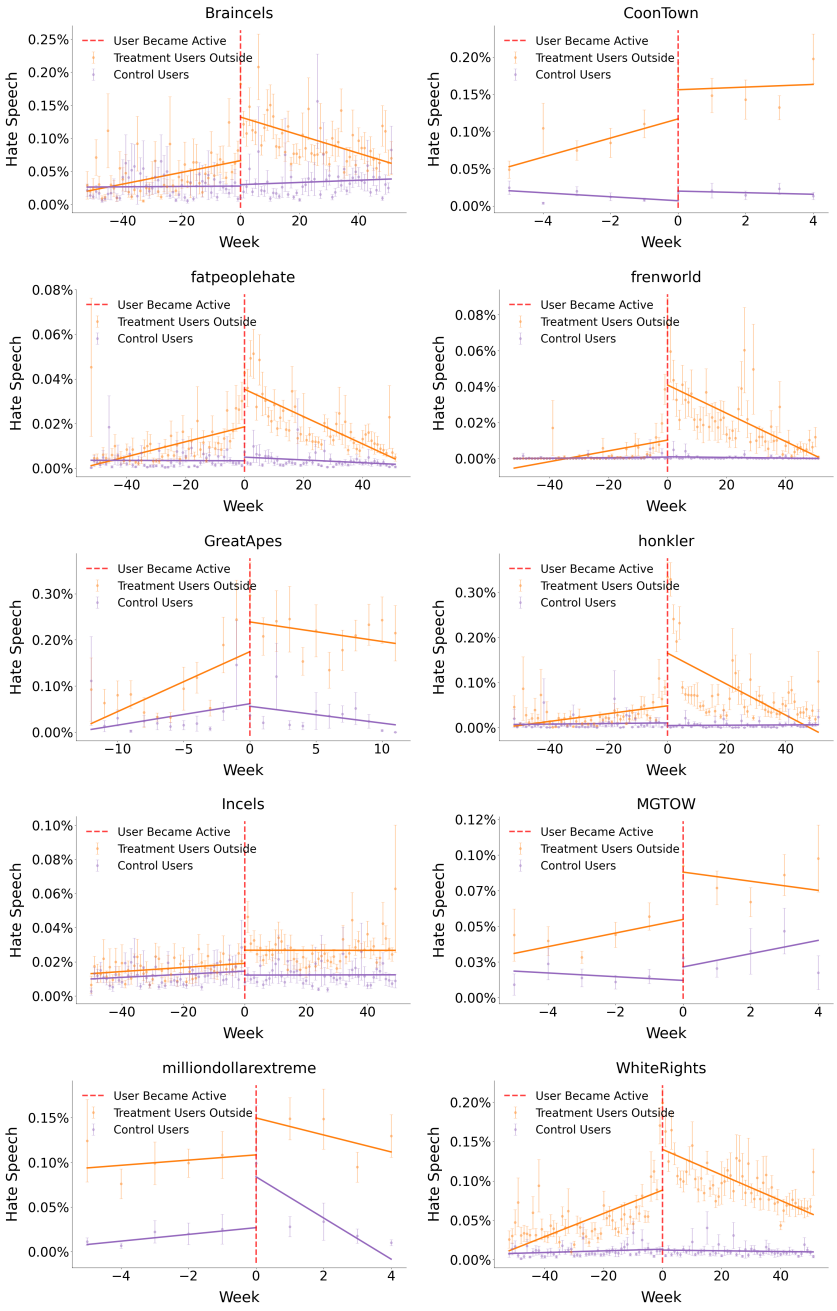

Fig. S6 Interrupted Time Series plots for the studied subreddits. We plot the rate of hate speech over time (percent of words that are hate words) outside of the subreddit in question for treatment users that join the hate subreddit and control users that never join. The subreddits are, in order, r/braincels (sexist), r/coontown (racist), r/fatpeoplehate (fat-shaming), r/frenworld (alt-right), r/greatapes (racist), r/honkler (alt-right), r/incels (sexist), r/MGTOW (sexist), r/milliondollarextreme (alt-right), r/whiterights (racist). A spike in hate speech by treatment users can be seen in all studied subreddits after the user becomes active in the community.