Supplemental Appendix for Estimating the Ideology of Political YouTube Videos

Angela Lai,[1,4‡], Megan A. Brown,[1], James Bisbee,[1]
Joshua A. Tucker[1,2,4], Jonathan Nagler[1,2,4], Richard Bonneau,[1,3,4,5]

[1]Center for Social Media and Politics, New York University

[2]Politics Department, New York University

[3]Computer Science Department, New York University

[4]Center for Data Science, New York University

[5]Department of Biology, New York University

[‡]To whom correspondence should be addressed: angela.lai@nyu.edu

December 7, 2023

1 Identifying Political Subreddits

In Figure 1, we provide an example of the type of plot that aided us in filtering non-political subreddits. Note the gap between the defined cluster of subreddits and "straggling" subreddits. In this example, we filter out subreddits with scores exceeding 5 on the y-axis. We repeat this process with a few pairs of coordinates.

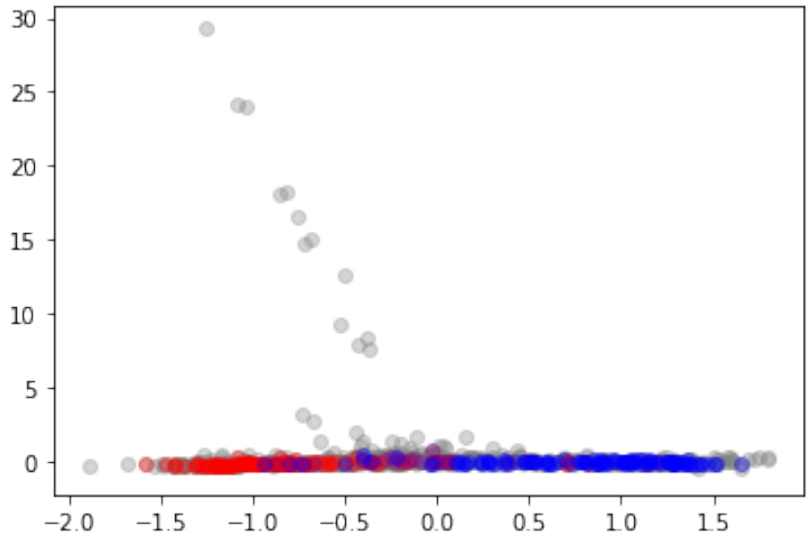


Figure 1: Correspondence analysis coordinates for political and non-political subreddits. Some points representing political subreddits are colored based on manual labels for a set of seed political subreddits.

2 Validation

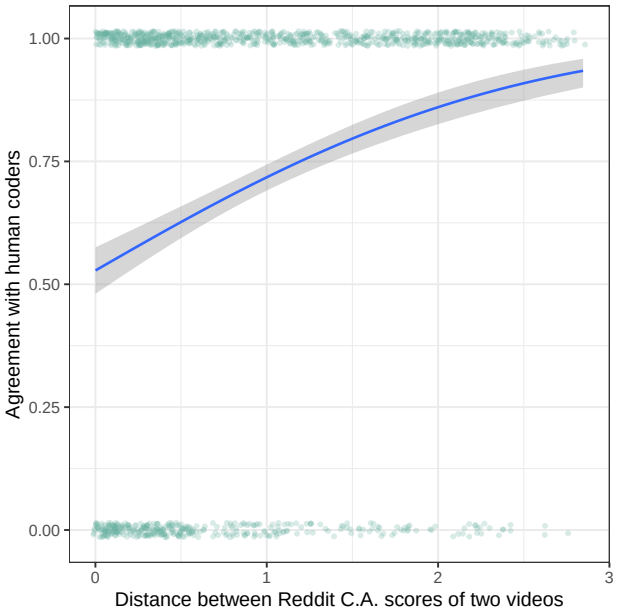
2.1 Human coding

In Figure 2, we validate our correspondence analysis ideology scores against human-coded pairwise comparisons as described in Section 3.2 of the main text. This validation exercise analyzes 1,194 unique pairs of videos where one or more coders compared video ideologies. 1,712 unique videos were used to construct these pairs.[1]

[1]We start with a total of 1,763 pairs of videos and exclude 14 pairs where coders were evenly split for a total of 1,739 pairs of compared videos. The BERT model used for ideology prediction does not see any videos in this set during training, so we use the same labeled videos for validation of the correspondence analysis and text-based

Score distance	% Agreement
(0.0, 0.25]	53.04
(0.25, 0.5]	61.28
(0.5, 1.0]	67.86
(1.0, 1.5]	75.90
(1.5, 2.0]	82.80
(2.0, 3.0]	88.66

(a) Agreement with human coders increases with the ideology score distance.



(b) Probit regression where agreement is a binary outcome.

Figure 2: In (a), the score distance is the absolute value of the difference between the ideology scores of two videos. Percent agreement is the percentage of labeled video pairs where the ideology scores aligned with the label and is calculated for videos falling within each score distance bin. In (b), each point is a labeled video pair, where the x-coordinate represents the score distance and the binary y-coordinate is whether the ideology scores of the videos agreed with the human label. We fit a probit regression to these points and find that it trends upward, increasing with score distance.

As discussed in Section 3.2 of the main text, we also validate our text-based ideology estimates with 937 human-binned videos. We here provide further detail on said validation. Some human coders were asked to bin the same video more than once, resulting in 476 cases where one coder binned a video multiple times. When one coder labeled the same video multiple times, we took the mean of that coder's labels and treated it as one observation.

This appears to be a difficult task for humans. We first exclude videos deemed non-political or unrelated to U.S. politics by at least two-thirds of coders. For the remaining videos (n=659), 19 percent were placed in conservative *and* liberal ideology bins, meaning at least 25 percent of coders labeled the video as conservative and at least 25 percent of coders labeled the video as liberal. Only

estimates. We have correspondence analysis-based estimates for both of the videos for 1,194 pairs. The remaining pairs contain at least one video which does not have such an estimate because it was not in our subreddit-video matrix—these are used to further validate the text-based estimates.

41 percent of videos labeled by exactly three coders (n=509) were placed in bins on the same side of the aisle by all three coders (meaning all conservative, all liberal, or only moderate). Interrater reliability as measured by Krippendorff's alpha is low at 0.519. We suspect that labeling videos is more difficult than other coding tasks, such as those involving text. We further note that our initial approach was based on a body of work demonstrating that pairwise comparisons are easier for non-expert human coders than conventional formats such as Likert scales [King et al., 2004, Oishi et al., 2005, Carlson and Montgomery, 2017].

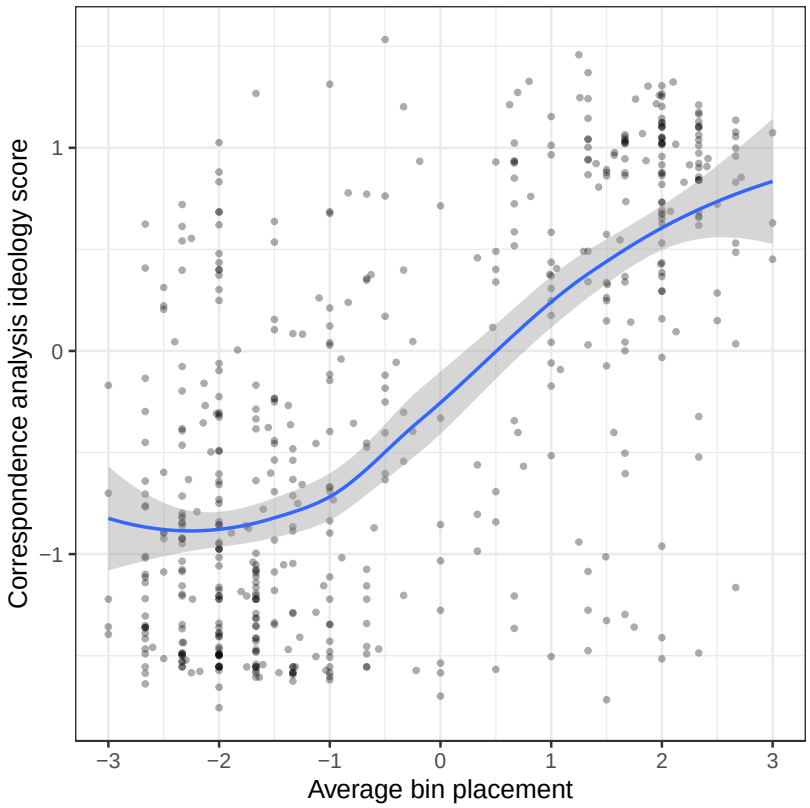


Figure 3: The average of the ideological bin placements by human coders vs. correspondence analysis ideology scores for videos where coders do not have significant cross-aisle disagreement. A local polynomial regression fitting is plotted on top of the points to show a trend.

In Figure 3, we show videos' average bin placements plotted against the correspondence analysis (CA) ideology scores. Figure 3 here is analogous to Figure 5 in the main paper but with CA scores instead of text-based scores. In both figures, we exclude videos where coders had cross-aisle disagreement, or where at least 25 percent of coders labeled the video as conservative and at least 25 percent of coders labeled the video as liberal, for a total of 535 video score comparisons. The

Pearson correlation between the CA scores and coders' video bins is very similar to the correlation between coders' bins and the text-based scores at around 0.66. For the text-based scores, note that we get a correlation of 0.61 if we do not exclude videos with cross-aisle disagreement.

Additionally, we looked at the correlation of unique bin values assigned to the same video by the same coder. If every coder was completely consistent and labeled a video the same way each time they saw it, we would have a correlation of 1. We calculate the correlation between x, "bin label 1," and y, "bin label 2", assigned by the same coder to the same video. If the same coder assigned the same label to the same video every time they labeled that video, that would be recorded as, for instance, [x=1,y=1]. A case where a coder placed a video in two unique bins might be [x=1,y=2]. If a coder labeled a video with more than two unique bin values, we sample two of those bin values. The correlation between bins assigned to the same video by the same coder was 0.84 and could be as low as 0.79 depending on how we sample labels from the 48 cases where coders placed the same video in more than two unique ideological bins.

Further, to better understand the cases where our text-based ideology scores disagree with human coders, we examine videos where the distance between our ideology score and the average human label is largest. We sort videos based on this distance and categorize roughly the top five percent, or 30 videos (the differences start to become more marginal once we go beyond that), based on their content and whether our score or the human label appears more reasonable.

In summary, we contend that these disagreements are largely in favor of our method and reflect the difficulty of this task. Coders labeled six far left videos as conservative while our score placed them as liberal. These videos often criticized mainstream Democrats like Joe Biden and Kamala Harris in the context of the 2020 U.S. primaries for president and came from progressive channels such as The Young Turks and The Jimmy Dore Show. Coders labeled three libertarian videos as liberal whereas our method scored them as conservative. The remaining videos covered a variety of topics and included four parodies, four videos with fringe commentary (in a couple cases coming from sources accused of spreading Russian propaganda), three videos with footage but no commentary (e.g. footage of Russian flags being thrown at Trump), two videos from foreign sources, and more in miscellaneous categories. We did not find evidence that our method's ideology

estimates systematically disagreed with humans on certain types of videos: there may be noise from videos our estimator has difficulty with but this does not appear to be directional.

2.2 Validation of text as model input

Before using these videos as training data, we perform an additional validation check. In addition to the human-labeled validation at the video level, we run a structural topic model (STM, Roberts et al. [2019]) on the transcripts of the videos found on Reddit and scored via correspondence analysis. We predict the probability that each out of 100 topics is associated with the video's ideology, modeled as a flexible spline over the support of the predicted ideologies, the logged number of likes each video received, and an indicator for the category to which YouTube assigned the video. Formally, the STM shares much in common with a standard latent Dirichlet allocation (LDA, Blei et al. [2003]) method in the sense that both approaches calculate the probability of topics across documents ($\theta_{d,k}$), and the probability of words across topics ($\phi_{k,w}$), based on the observed distribution of words across documents. The difference is that LDA starts from a set of shared Dirichlet priors to generate θ_d and ϕ_k, whereas STM models these priors with generalized linear models that incorporate document-specific covariates. These covariates both improve the estimation of the parameters of interest from a technical perspective, and also allow the researcher to estimate relationships of substantive interest in a more principled manner in which uncertainty at each step of the estimation is appropriately incorporated.[2]

The results of our STM analysis are summarized in Figure 4, which depicts the top five topics that are most strongly associated with liberal videos (left column), moderate videos (center column) and conservative videos (right column). To generate this plot, we first run the STM model predicting topic probability as a function of a flexible spline of the continuous measure of video ideology. We then aggregate over liberal (L), moderate (M), and conservative (C) videos to calculate the net predicted probability of a topic conditional on these parts of the ideological space. Liberal-owned topics are thus those who are disproportionately associated with liberal videos and

[2] An alternative approach would be a two-stage procedure in which the researcher first generates a topic model with LDA, and then regresses the resulting topic proportions on document-specific covariates.

disproportionately *not* associated with conservative videos. We summarize the substantive content of each topic by including the unique union of the top-5 highest scored words for each topic along the dimensions of probability ($\phi_{k,w}$), score (the log frequency of the word in a given topic divided by the log frequency of the same word in all other topics), and FREX (the weighted harmonic mean of each word's rank in terms of FRequency and EXclusivity).

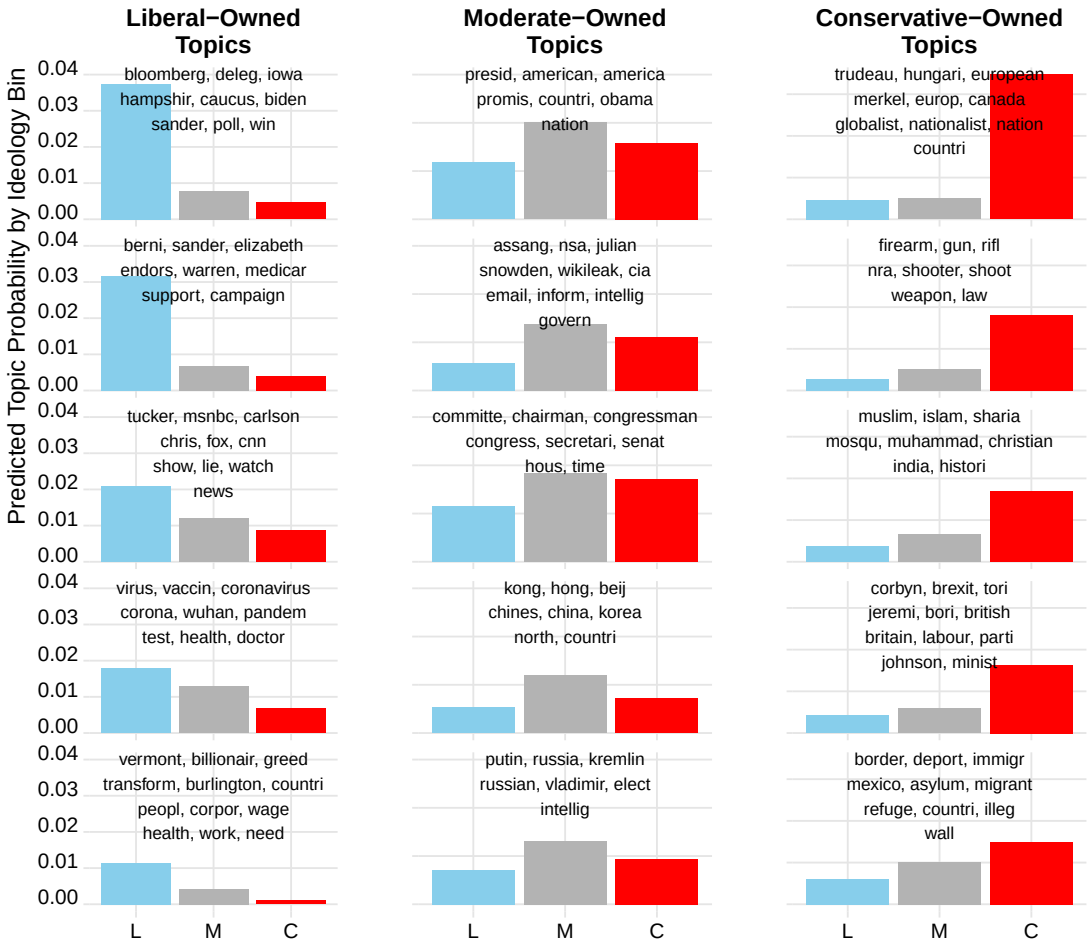


Figure 4: Results of STM analysis of video transcripts where topics are modeled as a function of the video's predicted ideology. Each column focuses on topics that are disproportionately associated with an ideology, calculated as the difference between the predicted topic loading among liberal (left column), moderate (center column), and conservative (right column) videos, and the minimum predicted category. I.e., the top-left plot is the topic most strongly associated with liberal videos (L, indicated on the x-axis) and least strongly associated with conservative videos (C, indicated on the x-axis). The topics are labeled by the unique union of the top five highest scoring terms in terms of probability, score, and FREX.

As in the preceding validation tests, we again find reassuring evidence in support of our pro-

posed measure of ideology. The most liberal topic is about the electoral fortunes of Bernie Sanders, the very popular liberal presidential candidate in 2016 and 2020, and Senator from Vermont. Three of the remaining four liberal-owned topics are about other progressive politicians, including Andrew Yang, Pete Buttigieg, Kamala Harris, Joe Biden, Elizabeth Warren, and Alexandra Ocasio-Cortez. Conversely, the most conservative topic is about Muslim extremists and religious-motivated terrorism, followed by the US-Mexican border, guns, the FBI investigation into Donald Trump, and finally gender identity. While not a bulletproof validation of our proposed method, these results are in concert with the other pieces of evidence give us confidence in the accuracy of our measure of the ideology of YouTube videos that appear on political subreddits.

3 Sensitivity Analysis

We test the robustness of the ideology estimates resulting from running correspondence analysis by permuting the data that goes into the underlying subreddit-video matrix. In Figure 5, we illustrate the effects of randomly dropping subsets of the Reddit post data used to construct the subreddit-video matrix. It is clear that dropping a sizeable portion of subreddits or posts seems to have a negligible effect on the ideology estimates of the videos overlapping between each set.

4 Uncertainty bounds

We calculate standard errors for the ideology scores obtained by running correspondence analysis on a subreddit-video matrix. We do not put uncertainty bounds on the text-based ideology estimates: it would be difficult to interpret the final outputs in addition to being highly computationally intensive since we would need to retrain a transformer language model hundreds of times. Instead, we here focus on the scores used as training label inputs for the final model and show that the text-based scores closely hew to the correspondence analysis scores.

We run non-parametric bootstrap 250 times and record the standard errors. For every run, we sample rows, or videos, with replacement from the subreddit-video matrix and then run

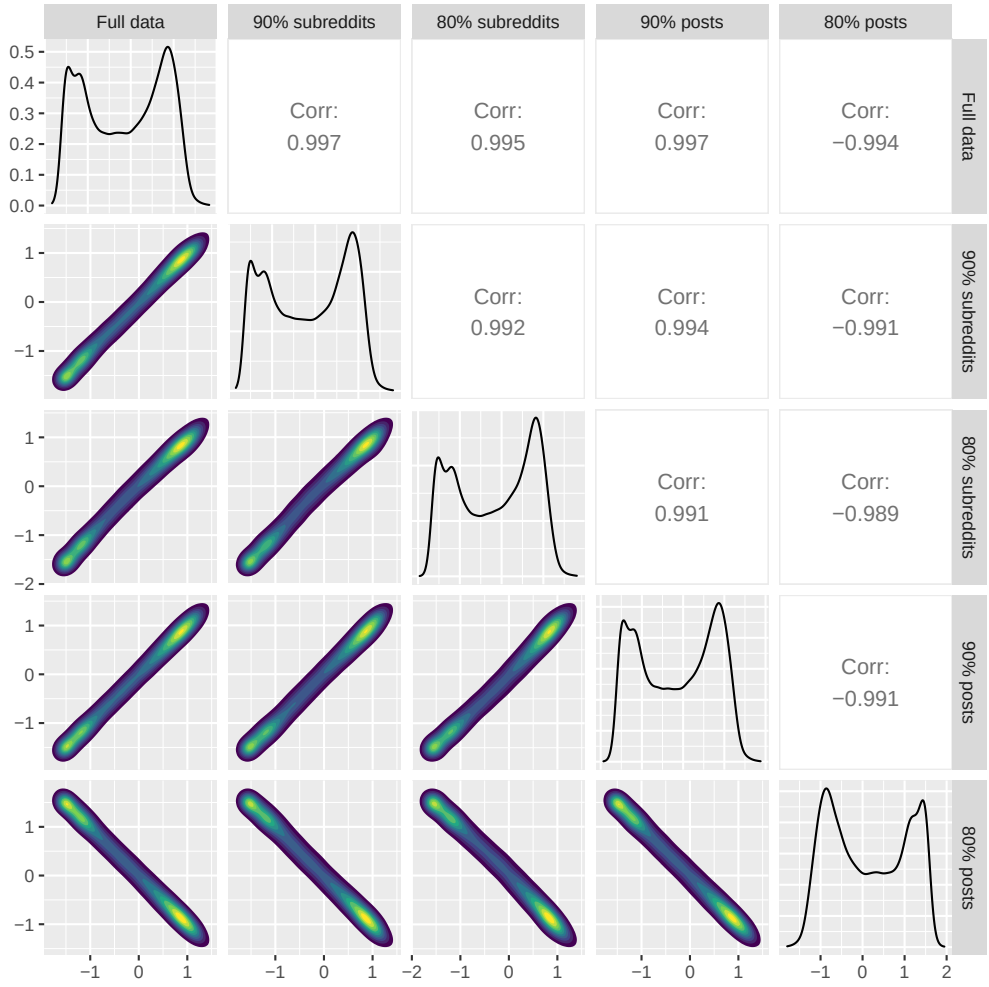
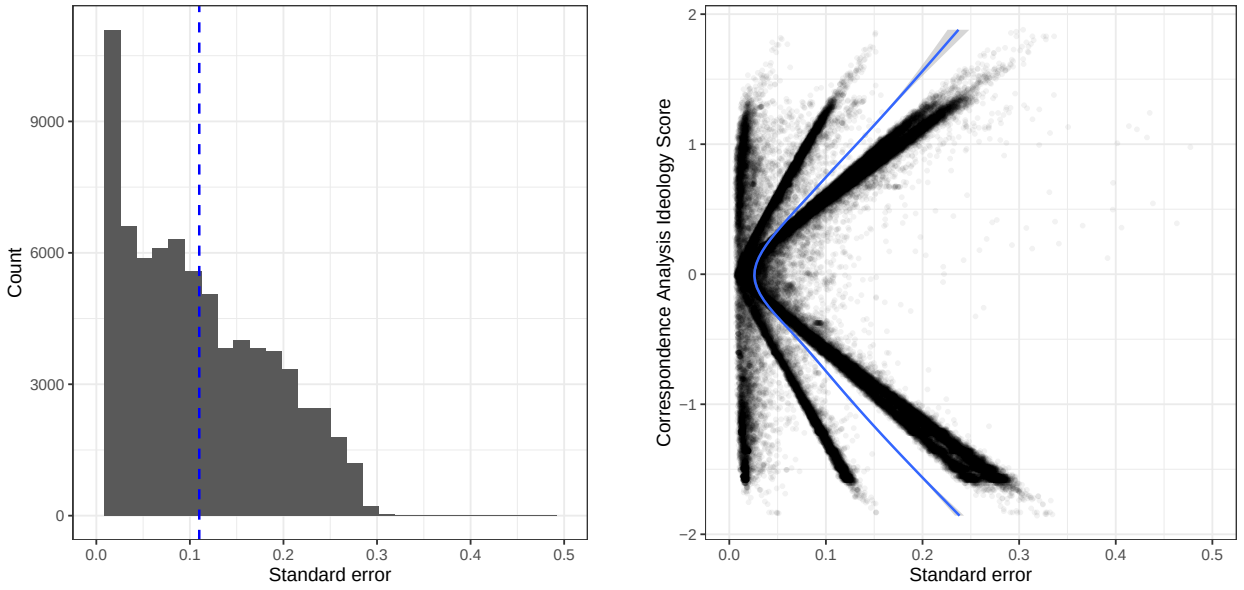


Figure 5: The resulting ideology estimates from correspondence analysis run on the subreddit-video matrix is robust to permutations in the data. Ideology estimates from our full data, data where posts from 10% of subreddits are dropped, data where posts from 20% of subreddits are dropped, data where 10% of posts are dropped, and data where 20% of posts are dropped have very high correlation.

correspondence analysis on this resampled matrix. We then calculate the standard error of each video's correspondence analysis scores. Figure 6 shows (a) the distribution of these standard errors and (b) standard errors vs. ideology estimates obtained by running correspondence analysis on the full data set. [3] The range of the original scores, or the result of running correspondence analysis on the full matrix, is -1.86 to 1.88, and the standard deviation of the same is 0.91. The mean of the videos' bootstrap standard errors is comparatively low at 0.11, though Figure 6 (b) shows that videos at the extremes tend to have higher standard errors.

[3]Note that 32 videos with standard errors > 0.5 were excluded from these plots for visualization purposes.



(a) (b)

Figure 6: (a) Histogram of the bootstrap standard errors for each of the 74,038 videos. The dashed line indicates the mean of those standard errors at 0.11. (b) The bootstrap standard error vs. the correspondence analysis ideology score for each video. This shows that the standard error tends to be higher, though still relatively low, for videos at the extremes. Note that we did not find meaningful associations between the standard errors and the number of subreddits each video appears in nor the number of bootstrap samples each video was in.

As discussed in Section 3.3 of the main paper, the text-based model's ideology estimates closely replicate the correspondence analysis scores with a mean absolute error of 0.295 and a correlation of 0.89 on the test set of videos. These estimates are plotted against each other in Figure 7.

5 Additional applications

To get a sense of respondents' media diets, we create a network of channels based on these watch histories and color the nodes with channel-level scores from Section 3.1 of the main text. Channels are linked if a respondents watched at least two videos from both channels. The edge weight is the number of such respondents. The resulting network in Figure 8 shows clustering based on channel ideology, which is consistent with the idea that most online users prefer ideologically congruent information. Channels with scores farther to the left appear to cluster around mainstream news

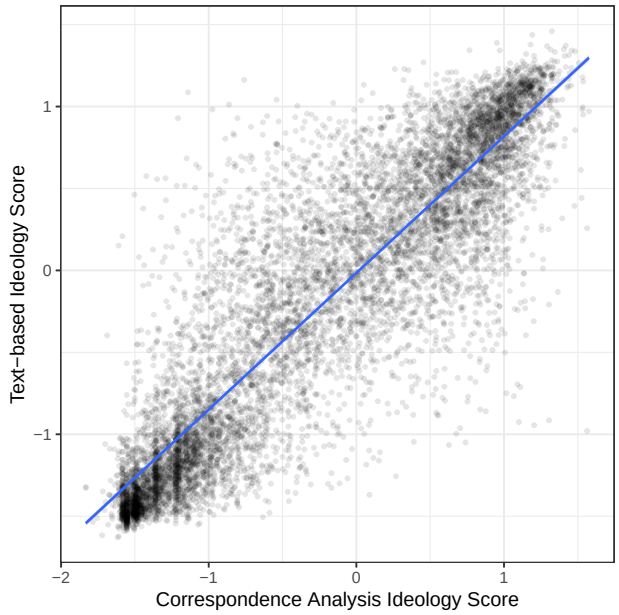


Figure 7: Correspondence analysis ideology scores vs. ideology predictions from the BERT model for a held out set of videos.

channels rather than in a separate cluster, whereas a group of channels with scores on the right appears somewhat distant, and less well-connected, to the cluster of mainstream news channels.

However, videos from a channel may meaningfully vary in their relevance to different ideological groups, so these channel-level labels could obscure a more complex picture. A benefit of our method is that it provides video-level ideology estimates. We therefore test for echo chambers by focusing on the viewing histories of the Democrats and Republicans. On average, respondents view similar numbers of videos regardless of party identification: 50% of Democrats view between two and 10 videos while 50% of Republicans view between two and 11.5 videos. Both have a median video count of four. 3,402 unique videos were viewed by Democrats and 1,635 unique videos were viewed by Republicans. Of those videos, only 106 were viewed by both Democrats and Republicans. We consider this intersection of videos separately and remove it from the Democrat and Republican video groups.

In Figure 9, we plot the distribution of ideology scores for videos viewed by Democrats only, Republicans only, and both Democrats and Republicans. The ideologies of the videos viewed by Democrats only and Republicans only have respective means of -0.324 and 0.458. The small fraction

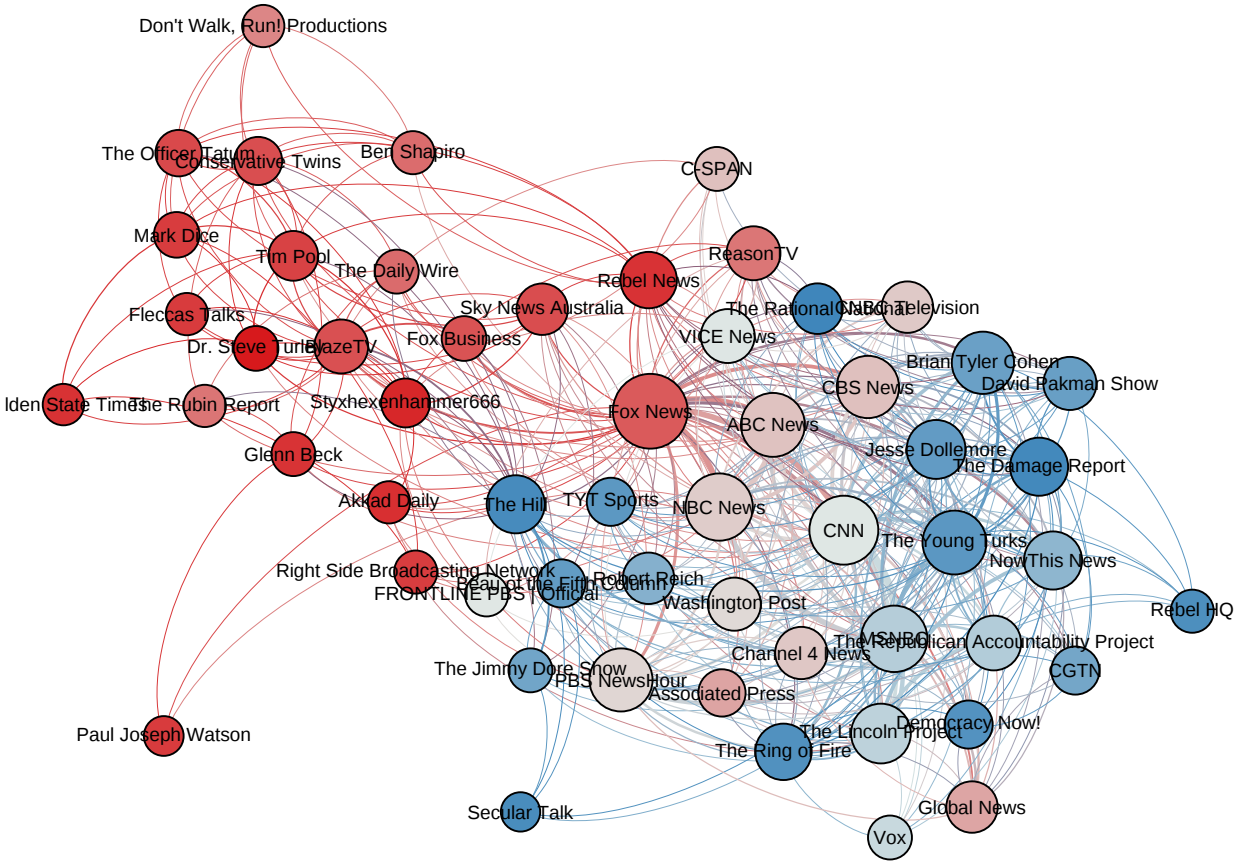


Figure 8: We draw a link between two channels if the same respondent has watched more than one video from both channels, and we plot the subset of channels which have at least 25 videos in the data set used in Step 1 of our method (Section 2.3 of the main text). The edge weight is the number of respondents who watched more than one video from both channels. Channels are colored according to the mean of their videos' ideology scores and nodes are sized based on degree.

of videos which were viewed by both Democrats and Republicans, on the other hand, has a mean ideology score of 0.014. Thus, we find that videos viewed by members of both parties appear to generally be more moderate while videos viewed by Democrats only and Republicans only tend to center around the left and right.

Note also, however, that the ideology distribution for videos viewed by Republicans only appears to peak around 0.75 whereas it peaks around -0.5 for videos viewed by Democrats only. If we exclude outliers when calculating the mean as shown in the overlaid boxplots, the mean of the Republican videos is farther from the center than the mean of the Democrat videos. Figure

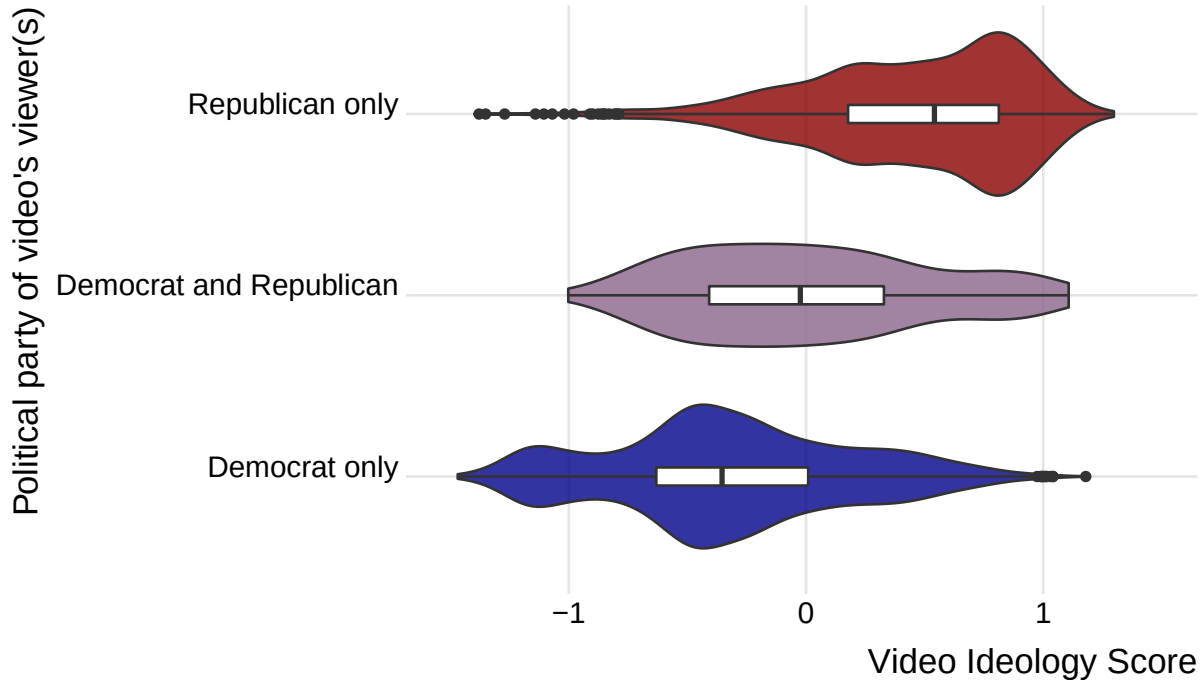


Figure 9: Distribution of video ideologies grouped by the self-reported party identification of the viewers. From top to bottom, we plot the ideology distribution of videos viewed only by Republicans, by both Democrats and Republicans, and by Democrats only.

8 suggests that some right-leaning channels cluster further away from mainstream channels while left-leaning channels tend to be closer — perhaps this accounts for some of the differences between the distributions in Figure 9.

Nevertheless, we find substantial areas of overlap in the ideological content consumed by Democrats and Republicans. This finding underscores an important benefit of our method: namely that video or channel-level analyses of echo chambers risk overstating their prevalence. Even though Democrats and Republicans may be watching different videos, there remains substantial overlap in the ideological content of what they watch. By estimating ideology as a latent measure, and by applying this to the video level, we can paint a more nuanced picture of the extent of ideological echo chambers on YouTube.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

D. Carlson and J. M. Montgomery. A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *American Political Science Review*, 111(4):835–843, 2017.

G. King, C. J. Murray, J. A. Salomon, and A. Tandon. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, 98(1):191–207, 2004.

S. Oishi, J. Hahn, U. Schimmack, P. Radhakrishan, V. Dzokoto, and S. Ahadi. The measurement of values across cultures: A pairwise comparison approach. *Journal of research in Personality*, 39(2):299–305, 2005.

M. E. Roberts, B. M. Stewart, and D. Tingley. Stm: An r package for structural topic models. *Journal of Statistical Software*, 91(1):1–40, 2019.