Supplementary Materials for

Who is Curating My Political Feed?

Characterizing Political Exposure of Registered U.S. Voters on Twitter

Assaf Shamir[1], Jennifer Oser[2], and Nir Grinberg[1]

[1] Department of Software and Information Systems Engineering, Ben-Gurion University, Israel
[2] Department of Politics and Government, Ben-Gurion University, Israel

This file contains:

Appendix A: Sample Socio-demographics

Figure A1 shows the demographic characteristics of the set of active panel members analyzed in this work. In terms of age, there is a larger fraction of panel members between the ages of 30-49 and the sample average is 40.00 [95% Bootstrapped CIs of (39.98, 40.02)]. In terms of ethnicity, Caucasians comprise 83.66 (83.57, 83.75) percent of the sample. Moreover, 56.16 (56.05, 56.28) percent of the registered voters in the sample are Democrats, and 31.66 (31.55, 31.78) percent are Republicans. The demographic breakdown shown below is very similar to Figure 1 in Hughes et al. (2021), which indicates that our sample is reflective of the broader population of registered U.S. voters on Twitter.

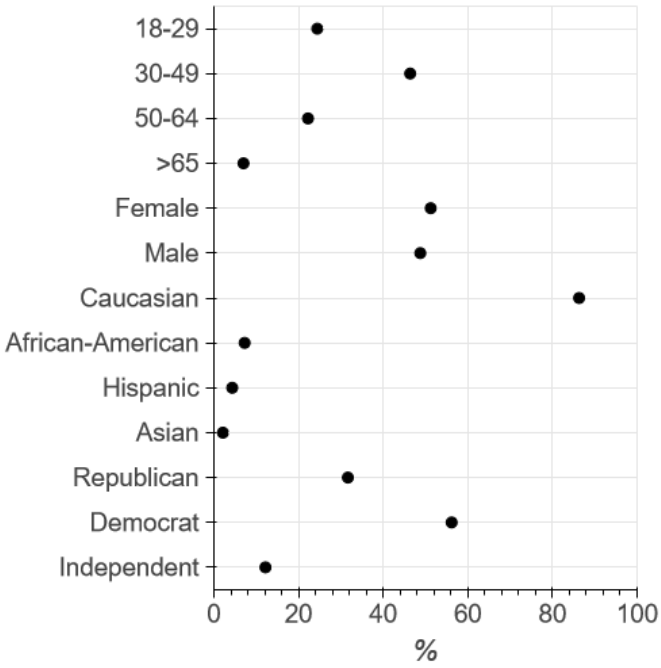

Figure A1: Demographic characteristics of our sample.
Note: The average demographic characteristics (age, gender, race/ethnicity, and political affiliation) for the set of 606,112 active panel members in our analysis. Party estimates are based on TargetSmart scores estimates.

Appendix B: Detection of Political Content

A central element to our analysis is the political classifier that distinguishes political from non-political content. We used the same keyword expansion approach utilized by prior work (Bakshy et al., 2015; Grinberg et al., 2019), and updated it for the 2020 U.S. Presidential election. The updating involved selecting high-specificity seed keywords that are likely to identify tweets about the U.S. election or politics more generally. Like prior work, we used a combination of general political keywords, hashtags, and candidate names to form our seed list. Then, we trained our classifier daily on a balanced set of political and non-political tweets (identified through a seed keyword list) to enable the classifier to identify additional words that co-occur with known political terms or figures.

We evaluated the political classifier using a stratified sample of 2065 tweets covering the entire study period and manual labeling by two raters on Amazon Mechanical Turk. Crowdworkers assigned the tweets into one of the following categories: (1) U.S. Presidential Election, (2) U.S. Politics, (3) Non-U.S. politics, (4) Other, or (5) I don't know. One of the authors resolved conflicts whenever they occurred. We find that the classifier can retrieve nearly all U.S. Presidential Election tweets with a recall of 96.4%. When collapsing categories (1) and (2) into one class of political content, we find that the classifier has a precision of 88.8% and a recall of 80.0%. These results are on par with the performance reported by Grinberg et al. (2019).

Appendix C: 116th Members of Congress Twitter Accounts

We compiled a list of 927 Twitter personal and campaign accounts for 533 representatives and 5 non-voting members from the 116th U.S. Congress. The 116th Congress convened on January 3, 2019 and ended on January 3, 2021, thus presiding throughout the study period. We started with a list of representatives from the official congress.gov website and then merged it with a list of politicians' Twitter account usernames (Wrubel and Kerchner, 2020). We manually validated that all accounts matched either a personal or a campaign account of an active MoC. Finally, Twitter account IDs were extracted by using the Twitter API. Two MoC are omitted from our analysis, one who closed their account and the other who did not post a single tweet throughout the election cycle. Therefore, our list includes 498 Democrats, 424 Republicans, four Independents, and one Libertarian MoC.

Appendix D: Identifying Opinion Leaders Accounts

As detailed in the Data and Methods section, the basis of our list of opinion leaders is a manually labeled list by Mukerjee et al. (2022), which we extend using the approach by Bail et al. (2018) that considers any user who is followed by 15 or more politicians as an opinion leader. This approach resulted in a set of 3,686 accounts that did not appear on any of the existing lists of media organizations, journalists, or politicians.

We further disentangled the set of 3,686 opinion leaders to make sure that they are strictly distinct from other categories of political actors in our analysis. We use inexact string matching to look for media and journalists' accounts by searching for names of media organizations (e.g., CNN) in the account name and profile description, which enabled us to find multiple accounts related to the same media outlet (e.g., *CNN, CNNNewsRoom, CNNBusiness*) and identify prominent journalists associated with it. We identified this way 127 additional media organization accounts, and 95 additional journalist accounts, which we manually validated.

We then trained four classifiers to identify each of the primary curator types with high precision based on account names and descriptions. In particular, we used the available lists of media organizations, journalists, politicians, and nonpolitical opinion leaders (from Mukerjee et al. 2022) to train four separate logistic regression classifiers, one for each curator type, to predict whether an account belongs in each category. In addition to a standard keyword-based model, we also fed our classifier with information about named entities in the account's name, which was extracted using standard Named Entity Recognition algorithms in the NLTK package. This improved the classifier's ability to distinguish between people and organizations. We used a

validation set that consists of a third of the accounts to set a score threshold for our classifiers to achieve 95% precision in their classifications.

Applying our classifiers to the set of 3,686 opinion leaders enabled us to identify a total of 1,933 additional accounts that belong to media organizations, journalists, politicians, and nonpolitical opinion leaders. Moreover, we manually labeled the top 300 accounts (by the total volume of exposure) in the remaining set of accounts that none of our classifiers identified. These top 300 accounts were responsible for 91% of exposure by the entire set of 3,686 opinion leaders, and thus contributed significantly to the proper attribution of political exposure. The remaining accounts were labeled as opinion leaders. To validate our inference about opinion leaders, we annotated a random sample of 100 accounts. We found that 80% of them were assigned to the correct category. For reference, a naive classifier based on the proportions of categories would have produced an accuracy of only 31%. In the following section (Appendix E), we validate the accuracy of our inferences for Opinion Leader accounts as well as for the other curating actor categories.

Appendix E: Validating Account Inferences and Robustness

In this section, we assess the robustness of our results to the inclusion of inferred accounts. First, Table E1 shows that the overall compositions of curator categories with and without inferred accounts are largely the same. The largest difference (6%) is for opinion leaders, which is expected due to direct change to this category. Consequently, we expect that there would be no considerable changes in the composition of political exposure due to the inclusion of inferred accounts. Figure E1 assesses this directly and compares political diets calculated with the inferred accounts (bars on the left) and without them (bars on the right). For example, for the nonpolitical cluster, we see that including the inferred accounts in the analysis reduces the indirect exposure to politicians from 38.1% to 37.5%. None of the differences in the figure is larger than 4.5% for all actor types and clusters. While small differences do exist, the overall pattern of similar breakdown is evident, and therefore we conclude that our results are not considerably affected by the inclusion or exclusion of inferred accounts.

	Media Outlets	Journalists	Politicians	Opinion Leaders
Manually Curated	7%	42%	20%	32%
Including Inferred Accounts	5%	38%	19%	38%

Table E1: Composition of curator accounts used in our analysis.
Note: The table shows a comparison of the composition of the pool of curator accounts used in our analysis, with and without the inclusion of inferred accounts

Figure E2 below provides an additional version of Fig 2 in the main body, with the ordering of the bars altered to better distinguish between direct (solid-colored bars, surrounded as a group by a black border) and indirect (lighter-colored bars without border) potential exposure.

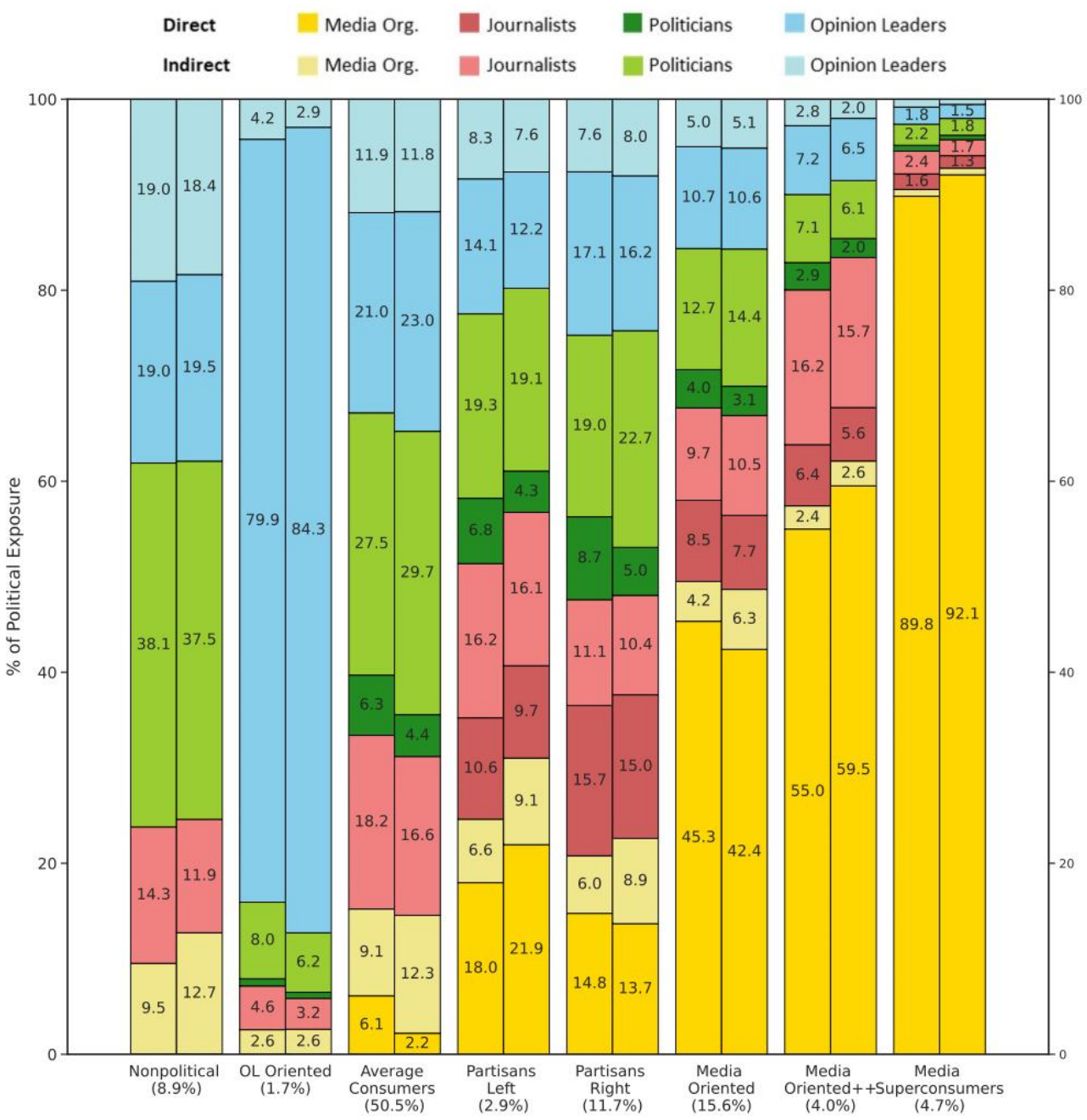

Figure E1: Political exposure when including or excluding inferred accounts.

Note: for each of the clusters, the bars on the left represent the political exposure obtained by including the set of inferred accounts; the bars on the right represent the political exposure while excluding inferred accounts.

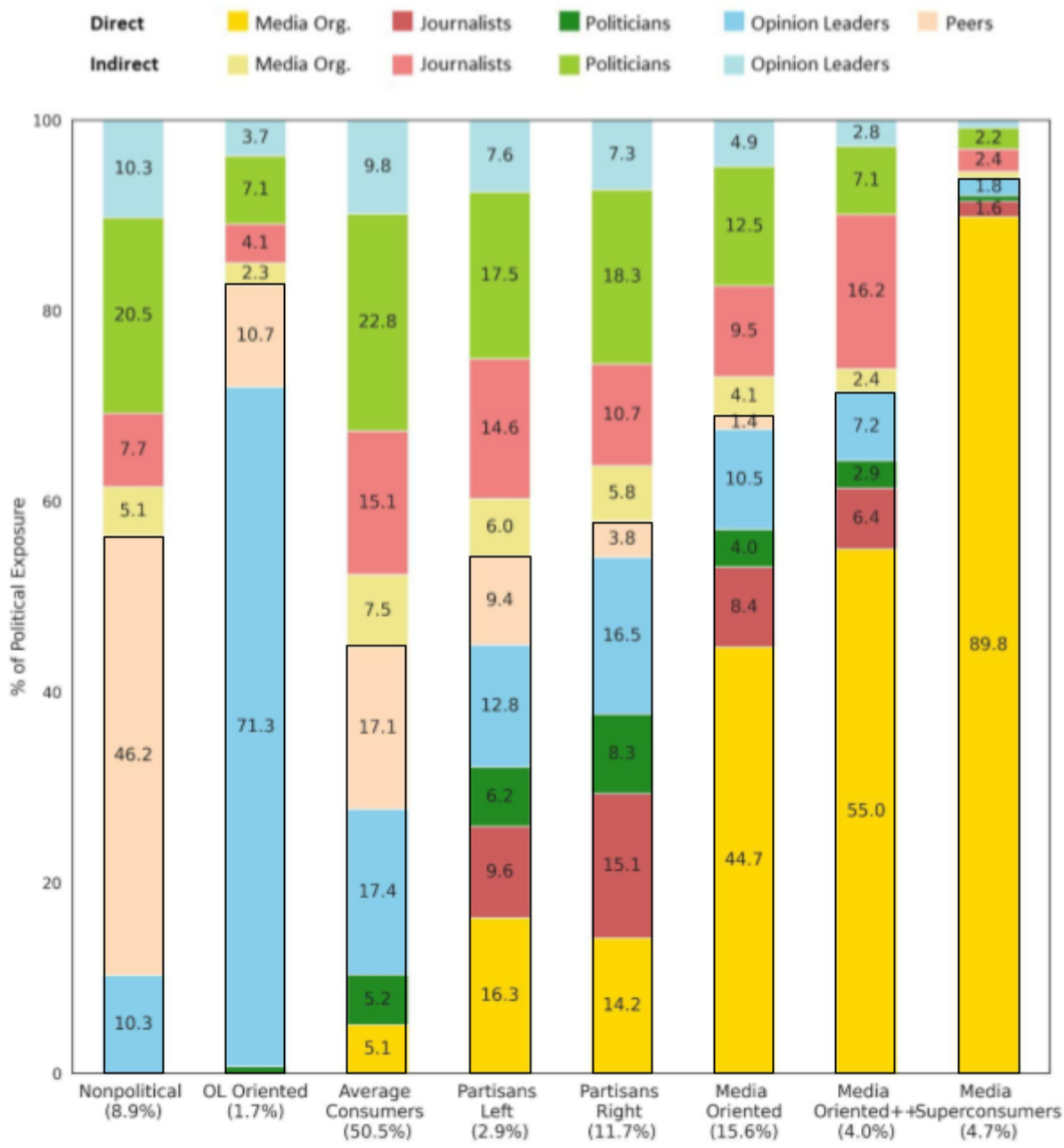

Figure E2: The composition of political exposure across clusters.

The share of politics curated by different actor types (y-axis) across clusters (x-axis). This figure parallels Figure 2, with the distinction that it is arranged according to whether the sources of information are direct or indirect. Darker-colored bars represent direct exposure to media organizations, journalists, politicians, opinion leaders, and social peers. A black border groups the bars representing direct exposure. Lighter-colored bars represent indirect exposure to media organizations, journalists, politicians, or opinion leaders through social peers.

Appendix F: Curating Actors

This section summarizes the different sources of curating actor types and provides examples of accounts on these lists. Table F1 shows the number of accounts used in this work from each source (McCabe et al., 2022; Mukerjee et al., 2022; Wojcieszak et al., 2022; Wrubel and Kerchner, 2020). Subtotals show that a sizeable proportion of accounts originated from manually curated lists, and that our inferences substantially expanded them (see the previous section for validation and robustness checks).

List Source	Media Outlets	Journalists	Politicians	Opinion Leaders
McCabe et al.	76	-	-	-
Wojcieszak et al.	80	1951	51	-
MoC + Wrubel and Kerchner	-	-	927	
Mukerjee et al.	-	-	-	1342
Opinion Leaders (n = 3908)				
Inferred	36	1084	621	1625
Manual Annotation	190	112	16	224
Subtotal (Manually Curated)	270	2063	994	1566
Total	306	3147	1615	3191

Table F1: Number of curator accounts used in this work by source.

Note: for each set of curating actors (media organizations, journalists, politicians, and opinion leaders) we present the total number of accounts used in our analysis and the corresponding sources. We used a total of 8,335 accounts in our analysis, out of which 3,366 were inferred using our classifiers.

Table F2 presents the top 10 examples (account name and Twitter handle) from each curator category when ordered by their number of followers from the panel. In terms of face validity, all of the accounts shown in the table belong to their respective categories. For example, the opinion leaders list includes examples of organizations (e.g. *NASA*), business magnates (e.g. *Bill Gates*), celebrities (e.g. *Jimmy Fallon*), and well-known public figures (*Michelle Obama*).

Media		**Journalists**	
Name	**Screen Name**	**Name**	**Screen Name**
The New York Times	nytimes	Anderson Cooper	andersoncooper
CNN Breaking News	cnnbrk	Rachel Maddow MSNBC	maddow
CNN	CNN	Jake Tapper	jaketapper
NPR	NPR	Ana Navarro-Cárdenas	ananavarro
The Washington Post	washingtonpost	Dan Rather	DanRather
The Wall Street Journal	WSJ	Chris Hayes	chrislhayes
BBC Breaking News	BBCBreaking	Jim Acosta	Acosta
HuffPost	HuffPost	Maggie Haberman	maggieNYT
Fox News	FoxNews	Yamiche Alcindor	Yamiche
TIME	TIME	Ezra Klein	ezraklein
Politician		**Opinion Leader**	
Name	**Screen Name**	**Name**	**Screen Name**
Barack Obama	BarackObama	Ellen DeGeneres	TheEllenShow
Joe Biden	JoeBiden	Jimmy Fallon	jimmyfallon
Kamala Harris	KamalaHarris	Michelle Obama	MichelleObama
Alexandria Ocasio-Cortez	AOC	Elon Musk	elonmusk
Hillary Clinton	HillaryClinton	Conan O'Brien	ConanOBrien
President Trump 45	POTUS45	Oprah Winfrey	Oprah
Vice President Kamala Harris	VP	NASA	NASA
Bernie Sanders	BernieSanders	Bill Gates	BillGates
Elizabeth Warren	ewarren	Justin Timberlake	jtimberlake
Bill Clinton	BillClinton	Melania Trump 45	FLOTUS45

Table F2: Example accounts for each of the curating actor categories

Note: These accounts include the top 10 accounts in each category in terms of the number of followers among panel members

Appendix G: Political Alignment of News Sites

We use an audience-based approach to estimate the political alignment of news sites similar to the one used by Bakshy et al. (2015). We follow a three-step approach. First, we compute a representation (a 100-dimensional embedding) for web domains based on their co-sharing patterns. We use a standard Word2Vec model (Mikolov et al., 2013) to obtain these representations where the sequence of domains shared by a user mimics words in a document in the original model. Second, we compute ideological alignment scores for domains shared by 30 or more panel members as the share of registered Democrats and Republicans who shared links for the domain. To enhance the robustness of our results, we remove users that share more than 100 domains per day or less than 10 tweets throughout the entire study period. We also remove the top 1% of sharers, regardless of domain sharing, to focus our inference on the majority of people. Finally, we use a neural network to learn the association between domain representation and alignment scores, and use it to extend the scores to the remaining domains shared by fewer than 30 people. Validating our inferences against the alignments scores provided by Bakshy et al. (2015) shows a Pearson correlation of 0.82, which demonstrates consistency with prominent work in the field. Following Guess (2021), we estimate the ideological slant of panel members' news diet using the average alignment score of the domains in their feed. To capture hyper-partisan consumption more directly, we quantify the fraction of sites that are shared almost exclusively (90% or more) by Democrats or Republicans.

Appendix H: List of Political Exposure Features

The full list of features used in our model for inferring the prototypical modes of political exposure is described in Table H1. We measured each of the features daily and averaged, for each panel member separately, across the study period (August to November 2020, inclusive). Note that in relation to the three categories of features discussed in the article in the final section of the Data and Methods that describes our clustering methodology, the relevant feature items for each category are as follows:

(i) The overall magnitude of political potential exposure (Features 1 & 2)
(ii) The curating sources partitioned by direct and indirect exposure (Features 3 through 12)
(iii) The ideological leaning of news sites in the feed (Features 13-15)

1	Number of political tweets per day (Log2)
2	Fraction of political tweets from Twitter feed
3	Fraction of political tweets from opinion leaders (direct)
4	Fraction of political tweets from opinion leaders (indirect)
5	Fraction of political tweets from politicians (direct)
6	Fraction of political tweets from politicians (indirect)
7	Fraction of political tweets from conservative opinion leaders & MoC
8	Fraction of political tweets from liberal opinion leaders & MoC
9	Fraction of political tweets from media (direct)

10	Fraction of political tweets from media (indirect)
11	Fraction of political tweets from journalists (direct)
12	Fraction of political tweets from journalists (indirect)
13	Fraction of political tweets from Left-leaning Hyper-Partisan websites
14	Fraction of political tweets from Right-leaning Hyper-Partisan websites
15	Average alignment score

Table H1: List of political exposure features used to identify prototypical types of exposure.

Appendix I: Age Distribution within Clusters

Figure I1 presents density plots of the age distribution within each of the clusters identified in our analysis. The y-axis is consistent across distributions, but its scale is omitted due to the uninterpretable nature of kernel-density estimates. Dashed vertical lines represent the cluster means, which exhibit the same positive correlation in cluster averages between age and political exposure (increasing from top to bottom) as discussed in the main body. For example, the nonpolitical cluster has a considerably lower average age than the average in any of the media clusters that have more political content available to them (as shown in Fig. 3 and discussed in the main body). Fig. I1 further shows that this trend also applies not only to the mean of each distribution but also to the mass of each distribution.

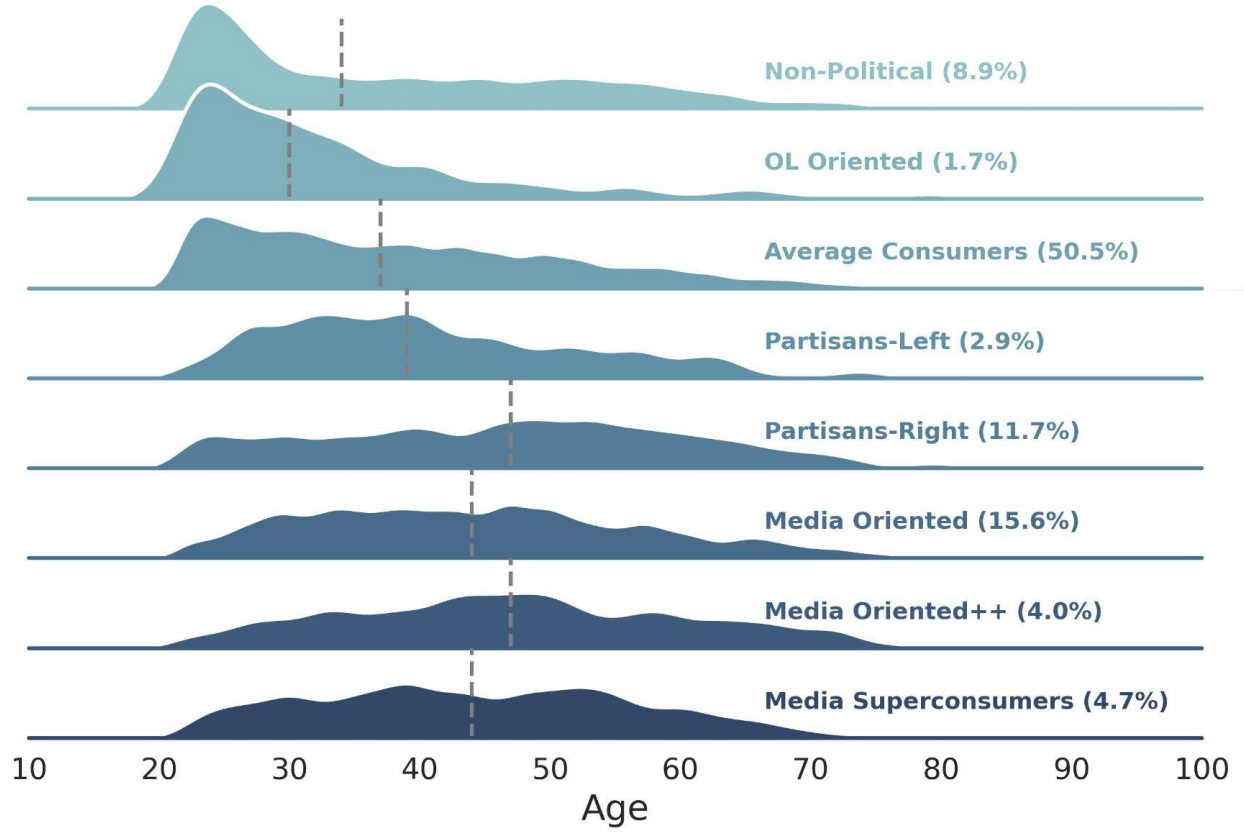

Figure I1: Age distribution within each cluster. *The y-axis is consistent across clusters, but its scale is omitted due to the uninterpretable nature of kernel-density estimates. X-axis shows age in years.*

Appendix J: Reference List for the Supplementary Materials

Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A., 2018. Exposure to opposing views on social media can increase political polarization. Proc Natl Acad Sci USA 115, 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bakshy, E., Messing, S., Adamic, L.A., 2015. Exposure to ideologically diverse news and opinion on Facebook. Science 348, 1130–1132. https://doi.org/10.1126/science.aaa1160

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., Lazer, D., 2019. Fake news on Twitter during the 2016 U.S. presidential election. Science 363, 374. https://doi.org/10.1126/science.aau2706

Guess, A.M., 2021. (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. American Journal of Political Science ajps.12589. https://doi.org/10.1111/ajps.12589

Hughes, A.G., McCabe, S.D., Hobbs, W.R., Remy, E., Shah, S., Lazer, D.M.J., 2021. Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets. Public Opinion Quarterly 85, 323–346. https://doi.org/10.1093/poq/nfab020

McCabe, S., Green, J., Wan, A., Lazer, D., 2022. New Tweetscores. OSF, Available at: osf.io/794va.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mukerjee, S., Jaidka, K., Lelkes, Y., 2022. The Political Landscape of the U.S. Twitterverse. Political Communication 39, 565–588. https://doi.org/10.1080/10584609.2022.2075061

Wojcieszak, M., Casas, A., Yu, X., Nagler, J., Tucker, J.A., 2022. Most users do not follow political elites on Twitter; those who do show overwhelming preferences for ideological congruity. Science Advances 8, eabn9418. https://doi.org/10.1126/sciadv.abn9418

Wrubel, L., Kerchner, D., 2020. 116th U.S. Congress Tweet Ids. https://doi.org/10.7910/DVN/MBOJNS