COMMUNICATION METHODS AND MEASURES
https://doi.org/10.1080/19312458.2025.2542725

Routledge
Taylor & Francis Group

The public that engages invisibly: what visible engagement fails to capture in online political communication

Yijing Chen[a], Zoltán Kmetty[b,c], Gerardo Iñiguez[a,d,e], and Elisa Omodei[a]

[a]Department of Network and Data Science, Central European University, Vienna, Austria; [b]CSS-Recens Research Group, Eotvos Lorand University, Centre for Social Sciences, Budapest, Hungary; [c]Sociology Department, Faculty of Social Sciences, Eotvos Lorand University, Budapest, Hungary; [d]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland; [e]Centro de Ciencias de la Complejidad, Universidad Nacional Autonóma de México, Ciudad de México, Mexico

ABSTRACT

Measurements of political polarization online have so far largely focused on visible traces accessible through platform APIs, neglecting invisible traces that are not recorded or otherwise unavailable to researchers, which can reveal key aspects of political engagement online. Our study addresses this gap by investigating the polarization measurement bias that arises when only visible engagement is analyzed, uncovering disparities at both the user and channel levels. Analyzing a combined dataset that links survey responses with YouTube digital traces through data donation from a sample of Hungarian Internet users ($N = 758$), we find that users who engage visibly through commenting are more politically polarized, and exhibit a greater level of selective exposure to content than users who engage invisibly through viewing. Moreover, ideologically heterogeneous channels are more likely to share viewers than subscribers or commenters. Thus, relying solely on public comment data may simplify, even overstate the segregation of political channels. Our results suggest that research using only visible engagement may overestimate the extent of polarization and the prevalence of echo chambers on YouTube. We highlight the benefits of using combined datasets to address measurement bias in online political communication, and contribute to the polarization literature by providing a fresh evaluation of potential biases in platform-focused research.

Introduction

Political polarization, the phenomenon of individuals diverging into distinct ideological camps, or becoming antagonistic toward out-group members, has become a critical global concern (Carothers & O'Donohue, 2019). The measurement of this phenomenon, which traditionally relies on survey responses, now increasingly incorporates online behavioral data from social media platforms, re-conceptualized in ways such as the extent to which users opt into ideologically congruent content that align with their preexisting beliefs (i.e., selective exposure) (Spohr, 2017), and how the overall user activities are segregated along the ideological line (Bright, 2018). While online datasets enable researchers to observe facets of polarization absent in self-reported data, it is essential to critically examine the observational perspective these datasets afford. When using social media datasets to

CONTACT Zoltán Kmetty kmetty.zoltan@tk.hu CSS-Recens Research Group, Centre for Social Sciences, Eotvos Lorand University, Budapest, 1019, Hungary

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19312458.2025.2542725

measure polarization, most researchers focus solely on activities that leave visible traces and are accessible by platform application programming interfaces (APIs), i.e., visible engagement, while neglecting activities that are harder to retrieve and remain hidden, i.e., invisible engagement (Sen et al., 2021; Wagner et al., 2021). However, can the analysis of visible traces alone provide a valid measurement of political polarization?

We address this question through a case study of political communication on YouTube, focusing on Hungary, one of the most severely polarized countries in Europe (Vegetti, 2019) that has so far received disproportionately little attention from researchers (Kubin & Sikorski, 2021). From a sample of Hungarian Internet users ($N = 758$), we compile an individually linked dataset of survey responses and digital traces. Our analysis centers on three forms of political engagement on YouTube: viewing, subscribing, and commenting, examining potential biases introduced when only visible engagement (i.e., commenting) is analyzed, while invisible engagement (i.e., viewing and subscribing) is overlooked. For each form of engagement, we identify the corresponding user group – viewers, subscribers, and commenters – who interact with Hungarian political content. Depending on how data is aggregated and what analysis unit is used, measurement bias can result in incomplete or deviated descriptions at different levels. Here, our study elaborates on two analytical levels: the user and the channel level, which represent the platform's demand and supply sides. This focus aligns with most prior research that analyzes either a specific user group [Wu & Resnick 2021; Bessi et al., 2016] or a set of channels e.g (Mislove et al., 2007; Rauchfleisch & Kaiser, 2020). More specifically, we define user- and channel-level biases as follows:

• **User-level bias** refers to disparities in measurement outcomes across user groups that engage in different forms or at different levels.

– We examine how users' political attributes (e.g., ideological leaning) correlate with their form and level of engagement. For example, if right-leaning users are more likely to comment, studies based solely on comment data may misrepresent the distribution of public opinion.

- We also assess how the ideological distribution of consumed content varies by engagement type. For example, if users comment primarily on ideologically aligned videos but view a more ideologically diverse set, reliance on comment data alone would yield a skewed picture of selective exposure.

• **Channel-level bias** refers to disparities in measurement outcomes across audience landscapes built on user groups that engage in different forms.

- We explore how patterns of audience overlap between channels vary depending on the observed form of engagement. For example, if left- and right-leaning channels attract distinct commenters but share many viewers, comment-based analyses would overstate the degree of polarization.

By examining the measurement differences between visible and invisible digital traces, our paper centers on the bias stemming from trace selection error, while interpreting results within a broader framework of digital data collection error (Sen et al., 2021). To clarify, we do not systematically examine other biases that may arise during the data collection phase (e.g., sampling bias), although in our later discussion of API-based and user-centric data collection these biases may become relevant. Our work contributes to the studies of online political communication as follows. Firstly, we reveal the relationship between political attributes and forms of online political engagement, helping researchers to reflect on their observed ideological space given the engagement form included in the analysis. Secondly, we connect the discussion of political engagement forms with reflections on the measurement processes that presume the selection of certain forms of engagement. By evaluating how excluding invisible engagement may impact downstream observations, we highlight the importance of articulating the scope of observations (e.g., only users who comment) when measuring polarization based on a limited dataset (e.g., only including the comments). Third, our analysis of a combined dataset showcases how obtaining richer and multi-dimensional user data can capture variations in ideologies and behaviors across different lenses of observation.

While our case study on Hungarian YouTube usage is unique, we caution readers about its generalizability, and hope to encourage explorations of more responsible approaches for collecting comprehensive datasets as well as analyses across different platforms and countries.

Biases from selecting digital traces of online political engagement

Regarding ways of online political engagement, existing works have distinguished categories such as passive and active political Internet use, and discussed ways in which they correlate with offline political engagement of various types e.g (Dimitrova et al., 2014; Kruikemeier et al., 2014; Vaccari et al., 2015). Yet, few have examined whether different categories of engagement lead to varying outcomes of polarization, ambiguating the ongoing debate about the relationship between social media usage and polarization (Matthes et al., 2023). Among the few that have considered this perspective, work by Matthes et al. theorizes about how active and passive social media use can shape affective polarization differently, and demonstrates that affective polarization is linked with active, but not passive political engagement (Matthes et al., 2023). Furthermore, when assessing online polarization through the lens of selective exposure – defined as a tendency for users to selectively consume ideologically congruent content (Stroud, 2010) – researchers have drawn on behavioral traces of passive consumption (e.g., browsing (Peterson et al., 2021)) and active consumption (e.g., liking and commenting (Cinelli et al., 2020)), but not yet jointly consider how different categories of engagement may encode varying implications for selective exposure.

In this paper, we connect the categorization of active/passive consumption with visible/invisible engagement, not only because they are practically overlapping (e.g., passive engagement such as viewing is usually invisible to researchers and to other users) but also because the visible/invisible framework helps articulate how limitations of data sources could impact findings derived from observational analysis. Given the scarce discussion on how various categories of engagement are accessible through platforms' APIs, it remains unclear how the visible and invisible engagement data would capture different polarization patterns. We now address these open tasks by leveraging a combined dataset of survey and digital traces on YouTube, focusing three forms of engagement with varying visibility (i.e., viewing, subscribing, commenting).

As one of the most popular video-sharing platforms, YouTube has been instrumental in facilitating online political communication through disseminating political content generated by politicians, news organizations, and grassroots content producers (Munger & Phillips, 2022). Studies collecting data from YouTube, mostly relying on its public API, have used three types of seeds – channels, videos and users – as the starting point for data queries. Many studies begin with retrieving relevant content, such as channels or videos grouped under a certain theme. Some borrow a pre-defined list of channels or videos e.g (Munger & Phillips, 2022; Rauchfleisch & Kaiser, 2020; Wu & Resnick, 2021), while others expand these lists through snowballing (Ottoni et al., 2018; Ribeiro et al., 2020; Stocking et al., 2020). A few also collect YouTube links sourced from external sites (Bessi et al., 2016; Lai et al., 2024). For works starting with users or utilizing user endpoints to expand samples, obtaining a random user sample appears impractical, and snowball sampling has been applied in this case as well (Clark & Zaitsev, 2020; Mislove et al., 2007). Very few have explored alternative pathways such as user-centric data collections (Ohme et al., 2024). One study that explores this pathway collects web-browsing data from a representative sample in the US to extract YouTube URLs viewed by the participants (Hosseinmardi et al., 2021).

Compared to Facebook or Twitter/X studies where users are usually the elementary units of collection and analysis, YouTube studies are more content-centered and content-driven. Investigations of the information landscape on YouTube typically begin by identifying the relevant content and, if needed, proceed to collect data on users who have engaged with it. As pointed out by Heft et al., content-centered approaches rely heavily on the dictionaries used to query for relevant content (Heft et al., 2024), which may introduce sampling bias if certain users deliberately avoid using dictionary terms (Massanari, 2017).

Because most of the aforementioned studies collect data via the YouTube API, the type of data accessible to researchers is almost always determined by what is available in the API scheme. Researchers have access to channel-level metadata (e.g., title, category, counts of views and likes) and video-level metadata (e.g., title, description, upload time), but not detailed subscriber or viewer

lists for these contents. For videos, researchers can further retrieve user comments; and for users, researchers can retrieve their subscription lists only if the user has made them public (Clark & Zaitsev, 2020). Thus, among all forms of user engagement, commenting has a more accessible – in our case, more visible – source of data than viewing and subscribing, making it a more common basis for measuring polarization. For instance, work by Bessi and his colleagues quantifies polarization as the proportion of comments a user leaves on content supporting one ideological side (Bessi et al., 2016). This underscores the importance of assessing whether, and how, a reliance on commenting-based engagement data introduces measurement bias in polarization research on YouTube.

Data and methods

In this section, we present our pipeline for data collection and analysis.[1] We start with broadly introducing the user-centric approaches for data collections, followed by our data donation procedure, where we outline the sampling strategies and describe the datasets we obtained. We then detail how we identify and label political content, and explain how we filter respondents during preprocessing and define respondent groups engaging via different forms. Finally, we describe our analytical steps to address user-level and channel-level biases

User-centric approaches of data collection

Among the various ways to access online political engagement data, retrieving data through platform APIs was the most common method in the 2010s. However, novel methods are needed as many popular platforms (e.g., Facebook, Twitter/X) increase their API restrictions (Breuer et al., 2023; Tromble, 2021). One promising alternative in recent years is user-centric approaches (Breuer et al., 2023), where participants are invited to voluntarily share access to their digital traces. There are two main user-centric approaches (Ohme et al., 2024). The first involves implementing tools that monitor participants' online presence, such as software that tracks browsing history or records content encountered on social media platforms (Haim & Nienierza, 2019). The second is data donation, which utilizes Data Download Packages (DDPs) that users can manually download from platforms (Boeschoten et al., 2022). Due to the General Data Protection Regulation (GDPR) law, technology giants such as Google, Meta, and Netflix are required to allow users to access and download their personal data stored on the platform. Through data donation, researchers invite participants to voluntarily donate their DDPs to legally and ethically retrieve their digital traces.

Compared to API-based access, the data donation approach is not bound by the restrictions of data availability from APIs, and provides richer – though not always complete – views of users' online activities. However, due to the complexity of the data donation procedure (e.g., recruiting participants, designing instructions and filtering out ineligible donations), sample sizes in these studies tend to be much smaller than those in API-based research. Additionally, ensuring sample representativeness can be challenging and is not easily addressed through survey design alone (Hase & Haim, 2024). One goal of our study is thus to showcase how to work through these challenges and obtain valuable insights from such datasets.

Collecting data through data donation

Now we detail our data collection procedure and the datasets we obtained from participants. Our data collection was conducted from February to June 2023.[2] After quality checks, we were able to obtain a combined dataset of survey responses and DDPs from a non-probability sample of Hungarian Internet users ($N = 758$). Participants who consented to donate were asked to upload their DDPs

[1]Our analysis code can be found at https://github.com/yijingch/invisible-public.
[2]The data collection project has been approved by the HUN-REN Centre for Social Science Ethical Board (1-FOIG/130–37/2022).

from YouTube and complete a survey questionnaire with demographic and political ideology questions. Details of recruitment, data collection and data availability are provided in Appendix F.

The population of interest is defined as Hungarian Internet users aged 16 and older who use the Internet for communication via chat or e-mail. To mitigate potential sampling biases, we applied individual weights to align the sample more closely with population distributions based on sociodemographic factors. These weights were generated using iterative proportional fitting (Bishop et al., 2007) to adjust for discrepancies. The weighting factors include gender, age, education level, type of settlement, and geographical region (see Kmetty & Stefkovics, 2025 for more details). Previous studies have mixed results on how biased these donated samples were (Hase & Haim, 2024; Gil et al., 2023; Kmetty et al., 2024. Additional analysis indicates that the data donation process does not introduce non-response biases in terms of respondents' political interests or ideological leanings (see Appendix F for details).

Through data donation, we obtain two sets of data from each participant: survey responses and digital traces on YouTube. In the survey, participants self-report their basic demographics (e.g., gender, age, education) and various political attributes (e.g., interest in politics, ideological positions). These data points are later used in user-level analysis.

For digital traces, we consider both visible (i.e., accessible via YouTube API, commenting) and invisible (i.e., inaccessible via YouTube API, viewing and subscribing) engagement that are included in the DDPs. Given the significant shifts in Hungarian's media environment under Orbán's evolving media policies (Bátorfy & Urbán, 2020), we restrict our analysis to user activities from the most recent 5 years of our dataset – May 2018 to May 2023.

Identifying (Hungarian) political content and assigning ideological labels

Given our focus on political engagement, we limit our analysis to political content within the donated dataset. To efficiently identify such content, we retrieved metadata via the YouTube API for all channels and videos that the respondents have engaged with. We first retrieved profile metadata for all 787,945 channels that appear in respondents' viewing, subscribing, or commenting activities. This metadata contains aggregated statistics (e.g., viewer and subscriber count) and YouTube-assigned topic tags. We classified channels as political if they included "politics" among their topic categories – approximately 1% of all channels). At the video level, we also collected API metadata for over 2.4 million videos viewed or commented by respondents, again using the presence of the "politics" tag to identify political videos.

After narrowing down to political channels and videos, we further focus on one specific context – Hungarian politics – that delivers consistent ideological implications. We manually labeled a set of political channels and channels containing political videos (i) to distinguish domestic (Hungarian) channels from international ones, and (ii) to assign political leaning labels for domestic channels along the anti-/pro-government spectrum, which is the most pronounced cleavage in today's Hungarian media system[3] (Bajomi-Lázár, 2021). The detailed labeling process is described in Appendix G. Among the 11,065 channels with political videos or tagged as "politics," we identified 626 Hungarian political channels, of which 139 were classified as pro-government, 276 as anti-government and 149 as neutral.

Filtering and extracting groups of respondent

Through the preprocessing above, our respondent sample has been incrementally narrowed: from the initial sample of 758 respondents, to the subgroup of 735 with some level of YouTube activities, to the smaller subgroup of 700 who have engaged with any political channels or videos via viewing,

[3]Although YouTube is non-traditional media, the most popular sites offering political content are either linked to politicians or to media outlets who have "traditional" nonsocial media platforms (e.g., online news sites, TV channels).

Table 1. Groups of respondent obtained in each filtering step.

Filter level (group abbreviation)	Size
All respondents who participate in data donation (ALL)	758
Respondents who have some level of YouTube activity (YTB)	735
Respondents who have engaged with political channel(s) or video(s) (POL)	700
Respondents who have engaged with Hungarian political channel(s) (POL-HU)	668

Table 2. Number of unique respondents, channels and videos for three forms of engagement (i.e., viewing, subscribing, and commenting) with Hungarian political channels.

Respondent group	Viewing	Subscribing	Commenting
# of unique respondents	640	299	72
# of unique channels	545	210	108
# of unique videos	57400	–	993

subscribing, or commenting, and finally to our focal group of 668 who have engaged with at least one Hungarian political channel via one of these engagement forms. We include all these groups in comparisons as a preliminary check, so that the biases introduced by selecting different engagement forms are not conflated with those caused by excluding non-active and nonpolitical YouTube respondents. We report the group size for each stage in Table 1.

Within the focal group of 668 respondents, we identify three respondent groups – viewers, subscribers, and commenters – based on whether one has engaged with at least one of the Hungarian political channels through viewing, subscribing, or commenting, respectively. We report the group size, and the number of unique channels and videos for each group in Table 2.

Addressing the user-level bias

We start our analysis with investigating user-level bias – the disparity in measurement outcomes across respondent groups engaging via different forms. First, we outline how we measure bias in self-reported ideologies across respondent groups. We then describe how we identify factors associated with levels of engagement, which may be biased if only highly engaged users are included in the analysis. Lastly, we illustrate how we compare the selective exposure pattern across different forms of engagement.

Analyzing bias in self-reported ideologies across different forms of engagement

First, we capture the bias in polarization measurement based on the self-reported ideological leanings from the survey. Focusing on respondents' anti-/pro-government position, we compare the polarization degree using two longstanding polarization metrics: variance and kurtosis (DiMaggio et al., 1996). Significant variations in these metrics across respondent groups indicate that studies focusing only on visible groups (e.g., commenters) risk measurement bias and should avoid generalizations about YouTube users. Further details of statistical tests used for the comparison are provided in Appendix A.

Analyzing bias in political attributes across varying levels of engagement

Besides comparing respondents who engage or do not engage in a certain form, we also examine the variation across different levels of engagement. To assess bias in political attributes, we compare these attributes across engagement levels using Negative Binomial Regression[4] to account for over-dispersion in the data. For each form of engagement, we construct a model comprising respondents who have engaged via this form with *any* content on YouTube[5] (i.e., from the respondent group YTB in Table 1) during the 5-year period. The dependent variable (DV) is the level of engagement,

[4]We tested four different models: Poisson, Negative Binomial, and the zero-inflated version of these models. Based on fit statistics, the negative binomial model was the best to use in this study.

Table 3. Dependent variables and sample sizes (i.e., number of respondents who have engaged in a certain form) for negative binomial models for each engagement form.

Model	DV	Size
view	number of Hungarian political videos viewed by a respondent	690
subscribe	number of Hungarian political channels subscribed by a respondent	680
comment	number of Hungarian political videos commented by a respondent	314

Table 4. Independent variables (IVs) for negative binomial models. The star sign (*) marks the reference variables.

IV	Encoding
gender	binary (1: male, 2: female)
age	continuous (divided by 10)
education	3-level (1: low, 3: high)
timespan	continuous (percentage of active months on YouTube; range from 0 to 1)
interest in politics	5-level (1: not interested, 5: very interested)
pro-gov	binary (1: belong to the pro-gov group, 0: does not belong)
anti-gov	binary (1: belong to the anti-gov group, 0: does not belong)
neutral*	binary (1: belong to the neutral group and NA, 0: does not belong)
left	binary (1: belong to the left-leaning group, 0: does not belong)
right	binary (1: belong to the right-leaning group, 0: does not belong)
center*	binary (1: belong to the center group and NA, 0: does not belong)

quantified by the number of recorded activities (e.g., the number of videos viewed, the number of channels subscribed) for the corresponding engagement form with Hungarian political content (see Table 3). For respondents who have engaged with some YouTube content but not Hungarian political content, their DVs would be zero.

For independent variables (see Table 4), we control for basic demographics (i.e., gender, age, and education) and timespan. Timespan quantifies the amount of time a respondent is potentially active on YouTube in the 5-year period, ranging from 0 (not active at all) and 1 (active all the time). We compute this by calculating the number of overlapping months between the respondent's activity interval (from the earliest to the latest activity of a respondent) and the 5-year interval, and normalizing it by the maximum length (i.e., 60 months).

The main focus of this analysis is to explore how political attribute varies across different engagement levels. To address this, we include the following political variables as IVs. Political interest is measured on a five-point scale (1 = not interested, 5 = very interested). Ideological variables are constructed using respondents' anti-/pro-government and left/right positions. The anti-/pro-government position is measured on a 0–10 attitude scale toward Fidesz, the incumbent party in Hungary (0 = strongly dislike, 10 = strongly like)[6] . As we do not expect a simple linear relationship between this variable and engagement levels, we recode it into three binary indicators corresponding to ideological groups: anti-government (0–2), neutral (3–7), and pro-government (8–10). The neutral group serve as the reference in the regressions. We impute the missing values (see Appendix E), and adopt these binary indicators to preserve the statistical power given the limited sample size. Similarly, the left/right positions are also recoded into three categories: left (1–3), 4 center (4), right (5–7), with the center group used as the reference. Additionally, we include extremity measures for anti-/pro-government and left/right to capture the effect of individual polarity. Extremity value equals the absolute distance between the center (5 for anti-/pro-government, 4 for left/right) and the actual value. Descriptive statistics for the main variables are available in Appendix I.

[5]For a robustness check, we tested different sample compositions, see Section *Engagement Levels and Ideological Characteristics* and Appendix D.

[6]In the sample, 46% of the respondents lean toward anti-government, while only 15% lean toward pro-government. Please note that such an imbalance reflects the nature of Hungarian Internet users being generally more anti-government than the general population. Pro-government people use the Internet less frequently and are under-represented in all online samples (Stefkovics, 2022).

For brevity, we present the results with the anti-/pro-government variables in the main text, since this dimension is the most salient one in Hungarian politics (Matuszewski & Szabó, 2019), and is correlated with left/right (Pearson's $r = 0.59$, $p < 0.001$), producing similar results downstream. Supplementary analyses using left/right variables are provided in the Appendix B.

Analyzing bias in observed selective exposure across different forms of engagement

Next, we investigate the variation in selective exposure – a tendency of users to selectively consume ideologically congruent content (Stroud, 2010) – across viewers, subscribers and commenters. Selective exposure has been shown to be causally linked to political polarization (Stroud, 2010), with social media potentially facilitating this process (Spohr, 2017). Evaluating differences in selective exposure across engagement forms helps identify potential biases in polarization-related findings based solely on visible engagement.

We characterize selective exposure by assessing the average and diversity of the content leanings for each respondent, and compare the distribution of these measures across respondent groups. For a given respondent i, we extract three sets of political content: videos they have viewed (V_i), channels they have subscribed to (S_i), and videos they have commented (C_i). For each set, we calculate the average and variance of content leanings,[7] which indicate the overall orientation and the ideological diversity of consumed content.

To assess whether selective exposure patterns in visible traces (i.e., commenting) differ from those in invisible traces (i.e., viewing and subscribing), we compare the distribution of leaning average and variance using the Kolmogorov – Smirnov (KS) test. We benchmark the empirical D statistics against a distribution of statistics derived from within-group resampling (see details in Appendix H), so that noises (i.e., within-sample fluctuation) are distinguished from what we examine (i.e., between-sample difference).

Addressing the channel-level bias

Last, we turn to channel-level analysis. We show that relying on a convenient form of engagement (e.g., commenting) and using only visible traces to characterize the user bases of political channels, researchers risk overlooking structural properties of channel landscapes captured through less visible traces such as viewing and subscribing. We illustrate this bias using both graphical presentations of channel networks and quantitative evidence of varying community structures of Hungarian political channels across different engagement forms.

Analyzing bias in degree of segregation in channel networks

We construct network projections for Hungarian political channels, based on three sets of engagement data from respondents' DDPs: viewing, subscribing, and commenting. These channel networks illustrate how channels are interconnected via shared user bases, with individual channels represented as nodes and links formed between channels that share similar audiences. To quantify audience similarity, we define a user vector $\vec{u}_i$ of length U (i.e., the number of unique users who engage in a certain form) for each channel i; the nth element of $\vec{u}_i$ equals the total number of times user n has engaged with channel i in a certain form. For any pair of channel i and j, we assess the audience overlap based on cosine similarity between $\vec{u}_i$ and $\vec{u}_j$.

$$\text{Similarity}(i,j) = \cos(\theta) = \frac{\vec{u}_i \cdot \vec{u}_j}{\| \vec{u}_i \| \| \vec{u}_j \|}$$

[7] As discussed in Section *Identifying (Hungarian) Political Content and Assigning Ideological Labels*, each channel is labeled as one of the following: pro-government (1), neutral (0), or anti-government (−1); and each video carries the label of its channel.

Edges in the network are weighted by these similarity scores. We set a non-zero threshold to filter out weak links for visualization purposes, and retain all links with positive weights in the assortativity analysis.

Measuring channel assortativity

Besides network visualizations, we also quantitatively evaluate the degree of segregation by assessing the extent to which ideologically similar channels share overlapping audiences. High segregation indicates that channels cater to distinct ideological audience groups, reflecting siloed consumption patterns within Hungary's political landscape.

We first compute an overall assortativity score of the leaning attributes for the entire network, then compute the EI homophily index (Krackhardt & Stern, 1988) at the node level. The overall assortativity score equals the weighted correlation between the leaning labels of two connected nodes (i.e., channels) with a positive edge weight. The higher the correlation, the more likely two channels with the same leaning have a similar user base, hence the more segregated the entire network. At the node level, the EI index measures the extent to which a given channel shares a similar user base (and thus is connected) with others on the same leaning side (i.e., internal links), versus others on different leaning side (i.e., external links). A lower EI index reflects stronger homophily, contributing to a more segregated network overall.

$$\text{EI index} = \frac{\sum \text{external link weights} - \sum \text{internal link weights}}{\sum \text{external link weights} + \sum \text{internal link weights}}$$

As with the user-level metrics, we compare the distribution of EI indices across different engagement forms using KS-test. To test robustness, we conduct 1,000 bootstrap rounds and report two p-values for the D statistics, accounting for within-sample fluctuations (see details in Appendix H).

Results

User-level disparity

We first present our results from the user-level analysis. In this section, we report how engagement forms interact with ideological distribution, how engagement levels associate with political attributes, and how selective exposure patterns differ across engagement forms.

Engagement forms and ideological polarization

We start with comparing respondent groups in terms of self-reported ideologies.[8] We see slight visible variations in the ideological distribution between viewers/subscribers and commenters, but not between all respondents and viewers (see Figure 1). Anti-government respondents constitute the majority across all three groups, with their dominance being especially pronounced among subscribers. Pro-government respondents, in contrast, are more dominant among commenters compared to subscribers and viewers. These differences are reflected in the variation in average anti-/pro-government positions across three groups (3.24 for viewers, 2.85 for subscribers, and 4.07 for commenters). The tendency for commenters to concentrate less on the center and spread further to the ideological extreme suggets that the commenters of Hungarian political content likely represent a more polarized fraction of the Internet users.

As expected, commenters have a significantly higher level of polarization, with greater variance and lower kurtosis than viewers, subscribers, and all respondents (see Figure 2). Moreover, the higher proportion of anti-government and left-leaning respondents in subscribers also leads to subscribers

[8]Since we do not see significant ideological variations in respondent groups in different filtering stages, we include the analysis for all respondents and respondent groups across engagement forms in the main text, and provide the results comparing respondent groups during filtering stages in Appendix A.

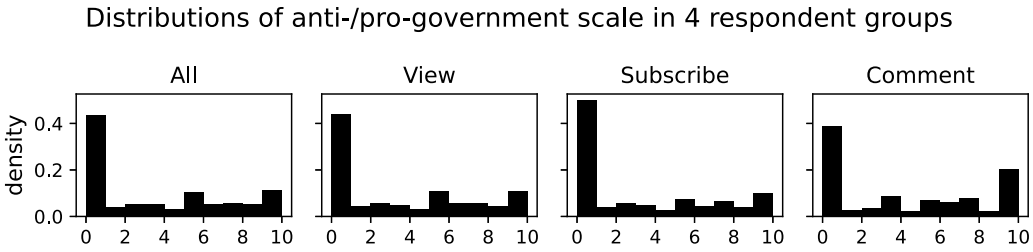

Figure 1. Distributions of respondent positions along the anti-/pro-government scale for four respondent groups (i.e., the entire sample, viewers, subscribers, and commenters).

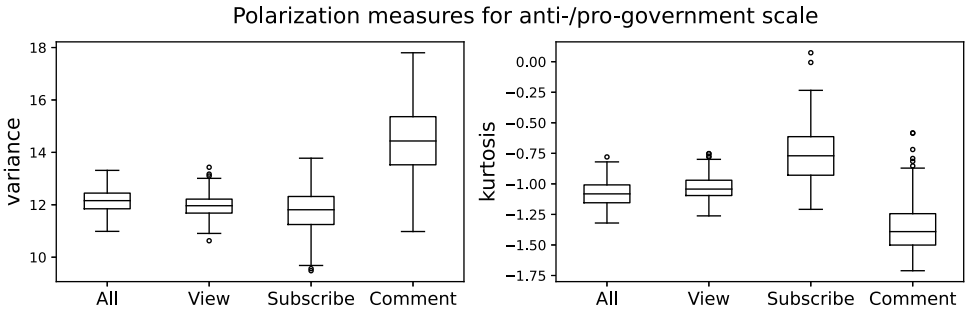

Figure 2. Variance and kurtosis of users' anti-/pro-government scale across four respondent groups (i.e., the entire sample, viewers, subscribers, and commenters). The value range comes from 200 rounds of bootstrapping with survey weights applied.

having a higher kurtosis than viewers. This finding, combined with the above observations, suggests that subscribers exhibit a more decentralized ideological spread, albeit in an imbalanced way. Statistical tests using Mann-Whitney confirm these differences (see Appendix H Table 6).

Engagement levels and ideological characteristics

The previous analysis compares users who do and do not engage in a certain form but does not assess the difference among users who engage at different levels. Thus, we further examine how different levels of each engagement form associate with individuals' political attributes. As shown in Figure 3, we find that the only persisting salient factor is interest in politics. Individuals with greater political interest exhibit higher levels of engagement with Hungarian political content across all forms.

Controlling for basic demographics and timespan, we find that pro-government respondents are likely to engage in lower levels of viewing, as indicated by a significant negative coefficient (−0.429, p = 0.025); anti-government respondents, while showing a similar trend, do not exhibit statistically significant predictive power (coefficient = −0.255, p = 0.052). This asymmetry might stem from pro-government respondents being more averse to ideologically incongruent content. When breaking down DVs into different channel categories, pro-government respondents would view significantly less neutral channels, while anti-government respondents would not (see Appendix C Figure 12). Models with left/right variables deliver similar findings (see Appendix B). Interestingly, when we assess the joint effect of political interest and anti/pro-government variable, we see that highly interested people view much more political content than those uninterested, regardless of their political leaning (Appendix B Figure 11). This further confirms that interest in politics, instead of ideology, is the main driver of higher engagement levels.

If belonging to one side of the ideology generally cannot predict the engagement level, can the extremity of ideological positions do so? The answer is also no, as we do not see any strong coefficients from the extremity variables included in separate models (Appendix B Figure 10).

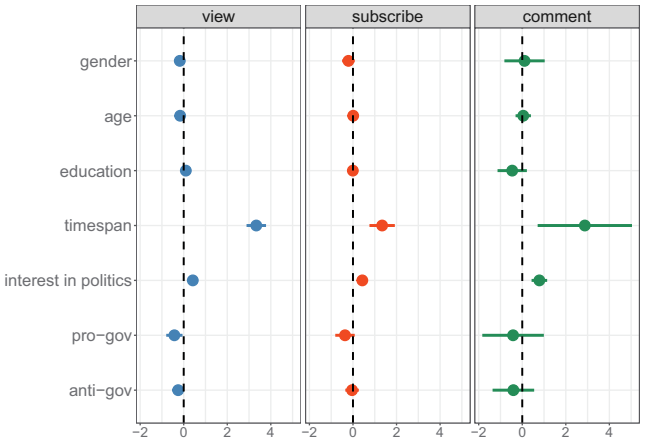

Figure 3. Negative binomial regression coefficients with 95% confidence intervals for viewing, subscribing and commenting activities on Hungarian political content. The DVs are the number of recorded activities for viewing, subscribing, and commenting from left to right for each panel.

For a robustness check, we re-run the models with an alternative filtering approach,[9] and a different imputation method.[10] Both approaches yield results consistent with our main findings.

Engagement form and selective exposure

We now turn to respondents' behavioral aspect and explore the variation in selective exposure across different engagement forms. For each respondent, we compute the average and variance of the leaning of the content viewed, subscribed, and commented, and show the distribution for respondents with the corresponding engagement records in Figure 4. In the left panel, the average leaning score for viewing is more centralized and less skewed than subscribing and commenting, with its mean falling around −0.166 (−0.540 for subscribing and −0.374 for commenting). This suggests that the selective

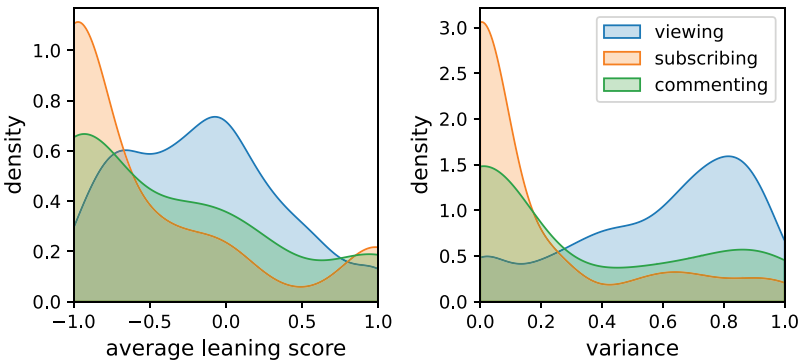

Figure 4. Average leaning score (left) and the leaning variance (right) for the Hungarian political content the respondents have engaged with through viewing, subscribing *or* commenting. An average leaning score of −1 (+1) means consuming only anti-government (pro-government) content, and 0 means consuming balanced or neutral content. Density functions are generated by kernel density estimate (KDE) methods.

[9]Instead of including only those who have engaged in a certain form (e.g., commenting), we also add those who have not engaged with a 0-value DV (see Appendix D).
[10]We used multiple imputation methods and analyzed pooled regression results (see Appendix E).

exposure characterized through commenting or subscribing behaviors appears to be more severe than through viewing. This finding is supported by the right panel: content viewed by respondents tends to be more ideologically dispersed, with a higher variance than content subscribed or commented by respondents. Kolmogorov–Smirnov (KS) tests confirm that these visual differences are statistically significant (see Appendix H Table 7).

Building on this observation, we further ask if the disparity in selective exposure outcomes stems from differences in sample composition or in forms of engagement. In other words, can we conclude that users who engage via viewing content are more likely to encounter mixed content than those who engage via subscribing or commenting, or is it that the content one views differs systematically from what one chooses to subscribe to or comment? To disentangle these possibilities, we re-generate Figure 4 using only the subset of respondents who engaged in all three forms ($N = 57$, see Appendix J Figure 18). The resulting pattern remains broadly consistent: the content viewed remains more mixed than content subscribed or commented, as indicated by a more central mean and a higher variance. However, these differences are less pronounced than those in Figure 4, suggesting that the disparities observed in Figure 4 are shaped by differences in both sample composition and engagement type.

Channel-level disparity

Finally, we turn to channel-level disparity and show how the channel audience overlap depicted based on different forms of engagement can display varying levels of segregation.

Engagement forms and audience segregation

Many studies that do not aim to infer a user-level landscape of political communication on YouTube, adopt a descriptive approach at the channel level instead, using channels as the basic unit of analysis. Accordingly, we examine channel-level biases that may arise from focusing only on comments when characterizing channels.

We first generate three channel networks based on viewing, subscribing, and commenting data from DDPs (see Figure 5). Compared to the viewing network, the subscribing and commenting ones appear more segregated along the anti-/pro-government cleavage, where ideologically aligned channels sharing more common users and clustering more closely together. We then assess these variations quantitatively. The network-level assortativity scores confirm that the viewing network has the lowest level of segregation, and that the commenting network is the most segregated among all (see Table 5). At the node level, the distribution of EI indices from networks built on

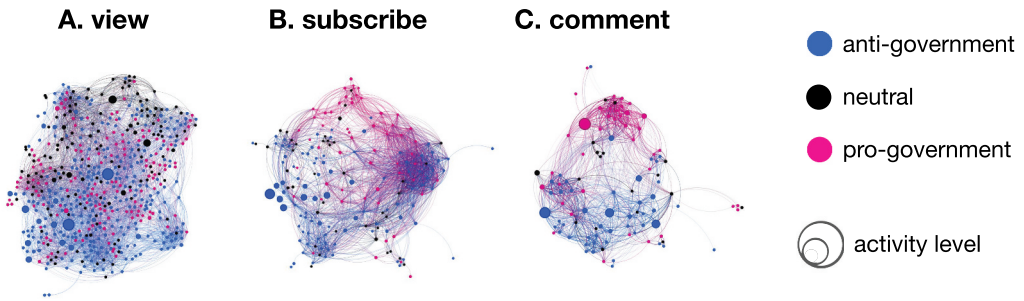

Figure 5. Networks of Hungarian political channels based on different types of user engagement. Each node represents one unique channel. Two nodes are connected if the users who engage with the two corresponding channels in a certain form overlap to a certain extent. Node color indicates the leaning labels for each Hungarian political channel, and node size is a function of activity levels for a given engagement type. Edges are filtered based on weights for visualization purposes (edge density from a to C: 0.02, 0.05, 0.1).

Table 5. Assortativity scores and average EI index for channel networks based on different forms of engagement.

Graph	View	Subscribe	Comment
Assortativity	0.0833	0.1289	0.3154
Average EI	0.1909	0.0801	−0.0298

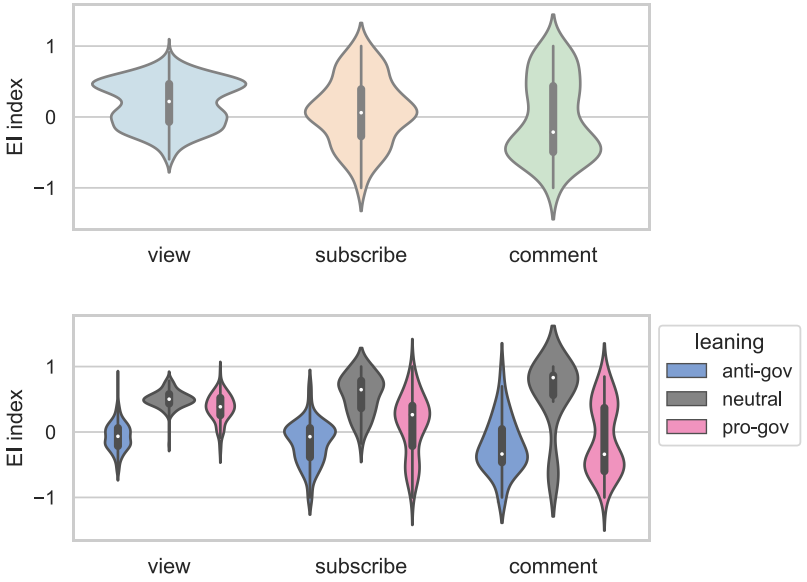

Figure 6. Node-level EI homophily indices across channel networks based on different forms of engagement. The upper figure shows the overall distribution; the bottom figure breaks down by channel leanings and shows the group-wise variation.

commenting data differs significantly from others: nodes in the commenting network generally have lower EI indices than those in the viewing and subscribing network. Again, we perform KS-test to formally test these differences and confirm that they are statistically significant (see Appendix H Table 8).

Additionally, when the distributions are segmented by leaning classes (see Figure 6), neutral channels show the highest degree of mixing with channels of different leanings, while anti-government channels are generally more segregated than pro-government channels.

Discussion

User-level: the incomplete picture of observed polarization

At the user level, we investigate whether and how engagement forms (i.e., how people engage – viewing, subscribing, or commenting) and levels (i.e., how much people engage) correlate with users' political attributes (e.g., ideological leaning, political interest) and patterns of selective exposure. Our findings suggest that, polarization measurements based solely on visible engagement (e.g., commenting) or excluding less politically active users can be biased in terms of ideological distribution and selective exposure.

First, for respondents who engage in different forms, we show that respondents who engage through commenting represent a subset more polarized in their self-reported ideologies compared to those who engage through viewing or subscribing. We therefore caution researchers who rely on YouTube public comments for user-level analysis, that such visible samples cannot represent all politically active users on YouTube, and are likely a more polarized fraction of the population.

However, among respondents who differ in their level of engagement, we find no significant differences in ideological leaning. Individuals who view more videos, subscribe to more channels, or comment under more videos, are not significantly more pro- or anti-government, or ideologically extreme, yet they are more politically interested. Thus, excluding users who engage at low levels risks omitting individuals who are less politically interested.

The above two observations jointly reveal the nuances of where and how user-level bias may occur. Those who only engage through commenting form a biased subsample in terms of ideological composition. However, this bias does not extend to differences between those who comment little and those who comment substantially. For levels of engagement, the clearest associated factor is political interest rather than ideology. While these outcomes diverge on whether self-reported ideologies is a primarily biased factor, they are not directly comparable due to discrepancies in the model design, such as sample variation, the presence or absence of control variables (e.g., political interest) and different binning of pro-/anti-government variables.

In addition, our analysis of respondent's selective exposure suggests that selective exposure is more pronounced in commenting and subscribing than viewing. Although respondents selectively comment under videos and subscribe to channels, they are fairly open to viewing ideologically neutral or diverse content on YouTube. This pattern aligns with Guess's finding that people have a moderate media diet when browsing news content online (Guess, 2021). Thus, characterizing selective exposure based solely on the commenting data e.g., (Bessi et al., 2016) can overstate the extent to which users opt into ideologically homogeneous content, consequently exaggerating the existence of echo-chamber on YouTube.

Channel-level: shared views but segregated comments

At the channel level, we find that audience overlap among Hungarian political channels depends on how *audience* is defined. Works that collect data via the YouTube API typically define audience as users who have made comments under videos of interest e.g (Bessi et al., 2016; Tran et al., 2022; Wu & Resnick, 2021). We argue that this approach can yield biased measurements of polarization, if polarization is quantified based on the degree to which certain channels form distinct communities and fragment the politically active users.

Indeed, our channel-level results show that the network of channels conditioned on commenting data does differ structurally from that conditioned on viewing data, which are not captured in YouTube API. Specifically, viewers of Hungarian political content show substantial overlap among channels with different leanings, while commenters are more segregated into ideological silos. This suggests that observations based on visible engagement (e.g., commenting), reflect a tilted, and potentially exaggerated picture of polarization. Therefore, when using only the commenting dataset, researchers should interpret evidence of "echo chamber" on YouTube (Marco et al., 2021). as reflective of commenting behavior rather than general content consumption, and refrain from extending findings to broader patterns of engagement.

Sources of biases

We contextualize our findings using Total Error Framework for Digital Traces of Human Behavior on Online Platforms (TED-On) (Sen et al., 2021), which provides a standard vocabulary to describe the source of measurement biases. By examining the differences between visible and invisible digital traces, our paper centers on measurement bias coming from trace selection errors, although we have dealt with trace reduction error and trace augmentation errors on the sideline.

The primary contribution of our paper is to quantify trace selection errors stemming from considering only visible political engagement data on YouTube. Trace selection error refers to the measurement error caused by selectively analyzing some type of digital traces and overlooking others. Our paper shows that when researchers select visible forms of engagement (i.e., comment) to construct

polarization measures, the outcome can be incomplete, often reflecting a greater degree of polarization compared to the outcome yielded by other invisible engagement forms.

Although our analysis does not focus on data pre-processing (e.g., labeling content), it is worth noting as a potential source of error in polarization metrics. Sen et al. define two error sources during pre-processing: trace reduction error and trace augmentation error (Sen et al., 2021). The former occurs when relevant traces are excluded or irrelevant traces are included, while the latter refers to errors caused by inaccurate annotations of digital traces.

In our study, we control these errors by covering as many popular channels as possible when filtering for political content (i.e., trace reduction) and assigning ideological labels (i.e., trace augmentation).

While automatic tagging methods are improving, off-the-shelf solutions for context-specific data remain limited. We opted for manual labeling, verified for accuracy, but given the time and resource restrictions, we had to limit the amount of content to be labeled. Although we inspected channels not explicitly tagged as "politics" but containing videos with "politics" tags, we did not examine the vast amount of channels without "politics" videos. Preliminary analysis indicates that most of this excluded portion consists of music content.

We also excluded political videos viewed by only one or two users, as these channels tend to have a significantly lower proportion of political content compared to channels with more viewers. To mitigate the risk of missing out widely viewed political content, we ensured that our labeled dataset covers more than 90% of total viewership.

Nevertheless, we acknowledge a potential source of trace augmentation error stemming from our channel-level (instead of video-level) labeling. Since not all videos uploaded by political channels are political, this may lead to an overestimation of user engagement with political content. This error could be better controlled by video-level annotation using context-specific classifiers.

Limitations and future work

We recognize several limitations of our study and encourage future researchers to explore these under-explored areas.

First, our analysis concentrates on political content identified using YouTube's topic tags. While convenient and commonly used in prior research e.g (Bärtl, 2018; Munger & Phillips, 2022), these tags may not serve as a definitive ground truth. Although we manually verified the list of Hungarian political channels for precision, the recall remains unclear – we may have missed political content not tagged as "politics" by YouTube. Additionally, seemingly apolitical content on YouTube may still be relevant to political communication, given the increasing politicization of pop culture on social media (Bay, 2018).

Second, we use manually assigned ideological labels for political channels and videos using a three-category spectrum (i.e., pro-government, neutral, anti-government). While this captures the most dominant political divide in Hungary, it offers limited granularity. Future work can expand this framework to reflect on additional ideological dimensions or topic cleavages. Even within our simplified categories, annotators have struggled with a low inter-rater agreement ($\kappa = 0.42$). However, weighted κ–adjusted for the number of watched videos – are substantially better (0.83), indicating strong agreement for more-viewed channels, which also carry greater weight in our analysis. We also assume that channel-level ideology aligns with the ideological slant of individual videos – a simplification that future work could test empirically.

Third, our analysis includes only Hungarian political channels, which may not capture the full scope of political content on YouTube. While our case study provides a fresh perspective into a relatively understudied political context, we caution readers about generalizing these findings to other countries. Furthermore, the engagement forms we analyze are specific to YouTube and those stored in DDPs, which can limit cross-platform comparisons. For instance, video likes were unavailable in the DDPs at the time of data collection, so we

cannot evaluate the polarization in liking behaviors. Other forms of engagement such as searching and re-posting may reflect different cognitive mechanisms and collective patterns not captured in our study. We hope this work motivates future research using more diverse datasets across other contexts and platforms.

Fourth, our analysis is limited by a small sample (especially in commenting data). For instance, our sample includes only 72 respondents who commented on Hungarian political videos. This small size increases susceptibility to sampling error and volatility in exposure patterns and network analysis. In addition, hyperactive respondents disproportionately influence the structural properties of channel networks. While we have implemented bootstrapping to statistically disentangle the within-sample fluctuation from between-sample variation, we caution researchers when interpreting such a small-sample results.

Fifth, our sample may suffer from coverage bias due to the requirements for data donation. Although our target population is Hungarian internet users aged 16+, we only included those who regularly use Google or Facebook. Despite the fact that both platforms are widely used in Hungary (penetration >85% based on our preliminary study, see Appendix F), this criterion may exclude relevant subgroups.

Finally, although the data donation approach provides a richer dataset than API-based collections, there can still be missing data if respondents having multiple YouTube accounts or have cleaned viewing history. In our sample, 46 respondents have comments or subscriptions but no views in their DDPs, likely due to disabled watch histories. Thus, we run analyses on alternative samples (see Section *Engagement Levels and Ideological Characteristics* and Appendix D) to ensure that our results are robust. However, this highlights the need to consider the methodological consequences of selecting specific traces to measure a construct.

Conclusions

Utilizing a combined dataset of survey responses and digital traces on YouTube, this study explores the potential bias in research that only focuses on visible engagement on YouTube (i.e., commenting) while neglecting invisible engagement such as viewing and subscribing. Here we focus specifically on the measurement of polarization, a heavily researched yet still contested issue regarding its relationship with social media.

We examine disparities at both the user and channel levels, and compare polarization measures across different engagement forms, and varying engagement levels. Our results reveal that (i) respondents who engage visibly (i.e., commenters) form a more polarized subset than those who engage invisibly (i.e., subscribers and viewers), (ii) individual engagement level does not associate with their ideological leaning or extremity, yet is significantly correlated with level of political interest; (iii) selective exposure is more pronounced in engagement through commenting and subscribing compared to viewing, and (iv) while politically divergent channels attract mixed viewership, their subscriber and commenter bases are more ideologically siloed. Overall, commenters and commenting behavior – which are more accessible via public APIs and thus more frequently studied – exhibit higher levels of polarization and ideological selectivity. Visible engagement on YouTube portrays an incomplete picture of political communication, potentially leading to inaccurate and often exaggerated estimate of polarization.

Therefore, we urge future researchers who rely on visible engagement traces to carefully define and interpret polarization measures within specific contexts. The observed ideological polarization based on visible engagement may be biased and exaggerated, as it often excludes more moderate and less politically interested users. While we do not advocate for a single ideal approach to measuring polarization, we hope this study provides a useful baseline for critically assessing current practices and for advancing more comprehensive frameworks that incorporate richer engagement forms.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The data collection was funded by the Eötvös Loránd Research Network within the framework of Flagship Research Projects: KÖ–32/2021. The work of Zoltán Kmetty was supported by the Digital Political Footprint project, funded by NKFIH, with a grant number [K-147329]; Nemzeti Kutatási Fejlesztési és Innovációs Hivatal [K-147329].

Notes on contributors

Yijing Chen is a PhD candidate in the Department of Network and Data Science at Central European University. Her work focuses on online political communication and polarization, using computational methods to analyze the segregation of agendas, beliefs, and communication contexts across the contemporary multimedia landscape.

Zoltán Kmetty is a Research Professor at the ELTE Centre for Social Sciences, CSS-RECENS Research group; and an Associate Professor at the Eötvös Loránd University Faculty of Social Sciences, Sociology department. He has diverse research interests, including computational social science, political sociology, and network studies. He is an expert in methodology, survey design, and quantitative analysis.

Gerardo Iñiguez is Associate Professor in complex systems at Tampere University and Leader of the Tampere Complexity Lab, as well as Visiting Professor at both the Department of Network and Data Science of Central European University and the Center for Complexity Science in the National Autonomous University of Mexico. His research aims at a theoretical understanding of complex adaptive systems, particularly the interplay between processes of information spreading and social network evolution online, with tools from statistical mechanics and network science.

Elisa Omodei is an Associate Professor at the Department of Network and Data Science at the Central European University in Vienna, Austria. Her research integrates network science and data science approaches to analyze societal issues, from food security to political polarization and the spread of misinformation on social media.

Ethical approval

The Data Donation study has been approved by the Center for Social Science Ethical Board (1-FOIG/130–37/2022).

References

Bajomi-Lázár, P. (2021). Hungary's clientelistic media system. In Morrison, J., Birks, J., Berry, M. (Eds.), *The Routledge companion to political journalism* (pp. 103–110). Routledge.

Bärtl, M. (2018). "YouTube channels, uploads and views: A statistical analysis of the past 10 years". *Convergence*, *24.1*(1), 16–32. https://doi.org/10.1177/1354856517736979

Bátorfy, A., & Urbán, Á. (2020). State advertising as an instrument of transformation of the media market in Hungary. *East European Politics*, *36.1*(1), 44–65. https://doi.org/10.1080/21599165.2019.1662398

Bay, M. (2018). Weaponizing the haters: The Last Jedi and the strategic politicization of pop culture through social media manipulation. *First Monday*. https://doi.org/10.5210/fm.v23i11.9388

Bessi, A., Zollo, F., Del Vicario, M., Puliga, M., Scala, A., Caldarelli, G., Uzzi, B., Quattrociocchi, W. (2016). Users polarization on Facebook and YouTube. *PLOS ONE*, *11.8*(8), e0159641. https://doi.org/10.1371/journal.pone.0159641

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media.

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4.2*(2), 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES

Breuer, J., Kmetty, Z., Haim, M., Stier, S. (2023). User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research. *Information Communication & Society*, *26.14*, 2649–2668. https://doi.org/10.1080/1369118X.2022.2097015

Bright, J. (2018). Explaining the emergence of political fragmentation on social media: The role of ideology and extremism. *Journal of Computer-Mediated Communication*, *23.1*(1), 17–33 https://doi.org/10.1093/jcmc/zmx002.

Carothers, T., & O'Donohue, A. (2019). *Democracies divided: The global challenge of political polarization*. Brookings Institution Press.
Cinelli, M., et al. (2020). Selective exposure shapes the Facebook news diet. *PLOS ONE*, *15.3*(3), e0229129. https://doi.org/10.1371/journal.pone.0229129
Clark, S., & Zaitsev, A. (2020). Understanding YouTube communities via subscription-based channel embeddings. *ArXiv Preprint arXiv 2010.09892* https://doi.org/10.48550/arXiv.2010.09892.
DiMaggio, P., Evans, J., & Bryson, B. (1996). "Have American's social attitudes become more polarized?". *The American Journal of Sociology*, *102.3*(3), 690–755. https://doi.org/10.1086/230995
Dimitrova, D. V., Shehata, A., Strömbäck, J., & Nord, L. W. (2014). The effects of digital media on political knowledge and participation in election campaigns: Evidence from panel data. *Communication Research*, *41.1*(1), 95–118. https://doi.org/10.1177/0093650211426004
Gil-López, T., Christner, C., de León, E., Makhortykh, M., Urman, A., Maier, M., & Adam, S. (2023). Do (not!) track me: Relationship between willingness to participate and sample composition in online information behavior tracking research. *Social Science Computer Review*, *41.6*(6), 2274–2292. https://doi.org/10.1177/08944393231156634
Guess, A. M. (2021). "(Almost) everything in moderation: New evidence on Americans' online media diets". *American Journal of Political Science*, *65.4*(4), 1007–1022. https://doi.org/10.1111/ajps.12589
Haim, M., & Nienierza, A. (2019). Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Computational Communication Research*, *1.1*(1), 79–102. https://doi.org/10.5117/CCR2019.1.004.HAIM
Hase, V., & Haim, M. (2024). Can we get rid of bias? Mitigating systematic error in data donation studies through survey design strategies. *Computational Communication Research*, *6*(2), 6.2. https://doi.org/10.5117/CCR2024.2.2.HASE
Heft, A., Buehling, K., Zhang, X., Schindler, J., Milzner, M. (2024). Challenges of and approaches to data collection across platforms and time: Conspiracy-related digital traces as examples of political contention. *Journal of Information Technology & Politics*, *21.3*(3), 323–339 https://doi.org/10.1080/19331681.2023.2250779.
Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D., Watts, D. J. (2021). Examining the consumption of radical content on YouTube. Proceedings of the National Academy of Sciences. (Vol. 118/32). pp. e2101967118 https://doi.org/10.1073/pnas.2101967118.
Kmetty, Z., & Stefkovics, Á. (2025). Validating a willingness to share measure of a vignette experiment using real-world behavioral data. *Scientific Reports*, *15.1*(1), 9319. https://doi.org/10.1038/s41598-025-92349-2
Kmetty, Z., Stefkovics, Á., Számely, J., Deng, D., Kellner, A., Pauló, E., Omodei, E., Koltai, J. (2024). Determinants of willingness to donate data from social media platforms. *Information Communication & Society* 28(7) , 1324–1349. https://doi.org/10.1080/1369118X.2024.2340995 .
Krackhardt, D., & Stern, R. N. (1988). Informal networks and organizational crises: An experimental simulation. *Social Psychology Quarterly*, *51*(2), 123–140. https://doi.org/10.2307/2786835
Kruikemeier, S., Van Noort, G., Vliegenthart, R., de Vreese, C. (2014). "Unraveling the effects of active and passive forms of political internet use: Does it affect citizens' political involvement?". *New Media and Society*, *16.6*(6), 903–920. https://doi.org/10.1177/1461444813495163
Kubin, E., & Sikorski, C. V. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, *45.3*(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070
Lai, A., Brown, M. A., Bisbee, J., Tucker, J. A., Nagler, J., Bonneau, R. (2024). Estimating the ideology of political YouTube videos. *Political Analysis* 32(3) , 345–360 https://doi.org/10.1017/pan.2023.42.
Marco, N. D., Cinelli, M., & Quattrociocchi, W. (2021). https://doi.org/10.48550/arXiv.2106.08684 Infodemics on YouTube: Reliability of content and echo chambers on COVID-19. *ArXiv Preprint arXiv 2106.08684*.
Massanari, A. (2017). # gamergate and the fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media and Society*, *19.3*(3), 329–346. https://doi.org/10.1177/1461444815608807
Matthes, J., Nanz, A., Kaskeleviciute, R., Reiter, F., Freiling, I., Neureiter, A., Stubenvoll, M., Sherrah, S. E., Juricek, S., Munzir, A., Noronha, I. (2023). The way we use social media matters: A panel study on passive versus active political social media use and affective polarization." In: *International Journal of Communication (19328036)*, *17*. https://ijoc.org/index.php/ijoc/article/view/19015.
Matuszewski, P., & Szabó, G. (2019). Are echo chambers based on partisanship? Twitter and political polarity in Poland and Hungary. *Social Media+ Society*, *5.2*(2), 2056305119837671. https://doi.org/10.1177/2056305119837671
Mislove, A., et al. (2007). Measurement and analysis of online social networks. *In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. San Diego, California, USA (pp. 29–42).
Munger, K., & Phillips, J. (2022). Right-wing YouTube: A supply and demand perspective. *International Journal of Press/Politics*, *27.1*(1), 186–219. https://doi.org/10.1177/1940161220964767
Ohme, J., Araujo, T., Boeschoten, L., Freelon, L., Ram, N., Reeves, B. B., Robinson, T. N. (2024). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18.2*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319
Ottoni, R., Cunha, E., Magno, G., Bernardina, P., Meira Jr, W., Almeida, V. (2018). Analyzing right-wing You Tube channels: Hate, violence and discrimination. Proceedings of the 10th ACM conference on web science. Amsterdam, Netherlands (pp. 323–332).

Peterson, E., Goel, S., & Iyengar, S. (2021). Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign. *Political Science Research and Methods*, *9.2*(2), 242–258. https://doi.org/10.1017/psrm.2019.55

Rauchfleisch, A., & Kaiser, J. (2020). "The German far-right on YouTube: An analysis of user overlap and user comments". *Journal of Broadcasting and Electronic Media*, *64.3*(3), 373–396. https://doi.org/10.1080/08838151.2020.1799690

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., Meira Jr, V. (2020). Auditing radicalization pathways on YouTube. *In Proceedings of the 2020 conference on fairness, accountability, and transparency* Barcelona, Spain (pp. 131–141).

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, *85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review*, *34.3*(3), 150–160. https://doi.org/10.1177/0266382117722446

Stefkovics, Á. (2022). The role of mixed-mode surveys in political surveying and election studies [PhD thesis]. Eotvos Lorand University.

Stocking, G., Van Kessel, P., Barthel, M., Matsa, K. E., Khuzam, M. (2020). *Many Americans Get News on YouTube, Where News Organizations and Independent Producers Thrive Side by Side* (Pew Research Center)https://www.pewresearch.org/journalism/2020/09/28/many-americans-get-news-on-youtube-where-news-organizations-and-independent-producers-thrive-side-by-side/

Stroud, N. J. (2010). "Polarization and partisan selective exposure". in. *Journal of Communication*, *60.3*(3), 556–576. https://doi.org/10.1111/j.1460-2466.2010.01497.x

Tran, G. T., Nguyen, L. V., Jung, J. J., Han, J. (2022). Understanding political polarization based on user activity: A case study in Korean political YouTube channels. *SAGE Open*, *12.2*(2), 21582440221094587. https://doi.org/10.1177/21582440221094587

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media+ Society*, *7.1*(1), 2056305121988929. https://doi.org/10.1177/2056305121988929

Vaccari, C., Valeriani, A., Barberá, P., Bonneau, R., Jost, J. T., Nagler, J., & Tucker, J. A. (2015). Political expression and action on social media: Exploring the relationship between lower-and higher-threshold political activities among Twitter users in Italy". in: *Journal of Computer-Mediated Communication*, *20.2*(2), 221–239. https://doi.org/10.1111/jcc4.12108

Vegetti, F. (2019). The political nature of ideological polarization: The case of Hungary. *Annals of the American Academy of Political and Social Science*, *681.1*(1), 78–96. https://doi.org/10.1177/0002716218813895

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595.7866*(7866), 197–204. https://doi.org/10.1038/s41586-021-03666-1

Wu, S., & Resnick, P. (2021). Cross-partisan discussions on YouTube: Conservatives talk to liberals but liberals don't talk to conservatives. Proceedings of the International AAAI Conference on Web and Social Media. Virtual conference (Vol. 15. pp. 808–819). https://ojs.aaai.org/index.php/ICWSM/article/view/18105/17908.