(Mis)measurement of Political Content Exposure within the Smartphone Ecosystem: Investigating Common Assumptions

DANIEL MUISE
Stanford University, USA

DAVID M. MARKOWITZ
Michigan State University, USA

BYRON REEVES
Stanford University, USA

NILAM RAM

THOMAS N. ROBINSON

The affordances of the smartphone are shifting individuals toward ever smaller and more fragmented units of political experience that are primarily provided by social contacts and non-traditional sources and voices. Accordingly, granular investigations of individuals’ exposure to political content must motivate political communication theories away from theories based on exposure to extended, uninterrupted political narratives that are constructed by news professionals and formal information outlets. In this piece, we use a novel approach to granular assessment of political

Daniel Muise: dmuise@alumni.stanford.edu
Date submitted: 2024-03-03

1 At the time of publication, author’s affiliation is Screenlake, Inc.

Copyright © 2024 (Daniel Muise, David M. Markowitz, Byron Reeves, Nilam Ram, Thomas N. Robinson). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at: http://journalqd.org
exposure on smartphones, to reveal significant complexity in modern political content diets that is at odds with simple assumptions made in political communication research. Based on five million screen-recording frames taken from 115 smartphones over two weeks, we find clear challenges to three common assumptions in the literature, with clear impacts on new theories about fragmented political media use. The assumptions we challenge are (1) unique encounters with political content can be aggregated by tabulation as though they are equivalent experiences; (2) durations of exposure to political content can be aggregated (e.g., at a monthly or daily level) without regard for how those time units are clustered at smaller timescales; and (3) singular political formats or sources (particularly for news) are sufficient proxies for measuring and manipulating overall political content exposure. Regarding both (1) and (2), our findings suggest that the majority of political content exposure occurs in rapid bursts lasting only a few forgettable seconds, and that the durations of political content exposure events are distributed according to a power-law curve. We find that this power-law distribution is robust to across apps and individuals, but also that these durations exhibit extreme variability both within and across individuals. Regarding (3), we find the vast majority of political content is encountered from formats and sources other than news and social media. We articulate how these results fit within and augment literature focused on political content exposure.

Keywords: selective exposure, political communication, Screenomics, digital media, screenshots, validity
Smartphone users can access political content\(^2\) at any time, for any duration, from any location, and in essentially any format (Molyneux, 2018; Reeves et al., 2021; Shim et al., 2015). This possibility poses a radical shift in political news exposure and a stark change from how political content was accessed in the preceding century. Whereas the broadcast era had an incredibly small menu of news sources, the present era has virtually infinite sources. Whereas the cable era had fixed programming timeslots and channels, the present era allows for exposure to politics throughout the day and across multiple platforms and devices. These changes have greatly expanded the possible sources of political influence. To keep up with this new fragmented environment, political communication researchers have greatly expanded use of new digital trace data to measure natural media interactions far more precisely than was possible with self-reports and laboratory experiments (Freelon, 2014; Jungherr, 2018; Lazer, 2020).

New media and new data have meant that certain topics in political communication have garnered special attention across popular and academic discourse. Chief among these is mass polarization, along with its hypothesized catalytic systems based on flexible media technology; namely, echo chambers and filter bubbles (A. Guess et al., 2018). These topics come into direct contact with the lived experience of many Americans and are built on long-held academic concerns about media failing to inform the public (Pariser, 2011; Terren & Borge, 2021).

In spite of the concerns, however, empirical research on filter bubbles, echo chambers, and technology-fueled polarization has yielded surprisingly mixed results regarding the intensity or even existence of these phenomena (Bruns, 2017; Dubois & Blank, 2018; A. Guess et al., 2018). There is a multitude of large-scale, observational, computational, and experimental studies that have revealed that the typical American consumes little news via smartphones or social media (Allen et al., 2020; González-Bailón et al., 2023), and moreover, there is little evidence that echo chambers and filter bubbles work as hypothesized, if at all (Bakshy et al., 2015; Barberá, 2020; Dubois & Blank, 2018; Eady et al., 2019; Flaxman et al., 2016; A. Guess et al., 2018; A. M. Guess, 2021; Morales

\(^2\) Political content is broadly defined as content of any format containing direct reference to governmental and societal affairs, broadly speaking. See full definition and operationalization in the Methods section.
et al., 2015; Muise, Hosseinmardi, et al., 2022; Peterson et al., 2019). The newest studies of echo chambers and filter bubbles, enabled with unprecedented cooperation from a large social media company, found that manipulation of the frequency and character of participants’ political content exposure did not yield strong impacts on their political attitudes or beliefs (A. M. Guess et al., 2023b, 2023a; Nyhan et al., 2023).

How can we make sense of this gap between theoretical expectations and popular understanding on the one hand, and still-unsettled empirical evidence on the other? In this paper, we advance the argument that techniques for using digital trace data to measure exposure to political content face greater limitations than are commonly recognized. These limitations stem, at least in part, from implicit acceptance of measurement assumptions tied to prior media eras that are inappropriate for the study of infinitely complex portable digital media. We demonstrate challenges in using large cross-sectional digital trace data through use of a new rich continuous smartphone screenshot method that directly captures moment-by-moment changes in screen content. These data allow examination of the fallibility of behavioral inferences commonly made in analyses of browser and application logs related to political media consumption.

**Three Problematic Assumptions About Aggregating Media Consumption**

We focus on three closely-linked assumptions about aggregating observations of exposure to political content. The first assumption regards the linear aggregation of encounters with political content, both across and within individuals. One conventional approach in the literature is to treat unique instances of media exposure (e.g., measured as the opening of a news article) as commensurable units. This is a holdover from prior eras of mass communication when programming occurred in uniform time blocks (e.g., hour-long television programs) (Lazer, 2020; Tewksbury et al., 2001). At odds with that approach is the temporal flexibility afforded by smartphones that allows users to access political content for any duration and with varying levels of attention. Skimming, scrolling, and switching to non-political content can occur at-will, in an instant and with minimal interface hassle (Matthes et al., 2020). Therefore, when encountering political content,
smartphone users may attend for hours or for a few seconds --- a range representing levels of attention that create radically different user experiences (Lazer, 2020; Van Damme et al., 2019).

The second problematic assumption relates to the aggregation of time units of exposure without consideration of when the exposure might have occurred. That is, aggregate measures of time-spent on political content over a month, day, or hour necessarily erase crucial information because they do not account for how those seconds are grouped into different segments. Across a day, one hour of exposure to political content can occur in a single one-hour-long segment, or across 360 ten-second segments, or in various combinations. Measurement strategies that aggregate political content exposure time without regard for the durations and timing of individual segments risk equating entirely different psychological processes that are used to process messages of different lengths (Lemke, 1998). The processing of political content depends on the speed at which (and thus the duration over which) processing occurs (Matthes et al., 2020; Stoker et al., 2015), as well as the user’s goals and willingness to spend time on elaboration (Kaye & Johnson, 2004; Van Damme et al., 2019). This is especially relevant in political communication research, given the greater cognitive resources needed to affect knowledge gain or opinion change (Gil de Zúñiga et al., 2021), and the variability of cognitive resources over the course of a day. Studies of repeated exposure that do not differentiate between durations are ignoring such differences, and doing so at the expense of useful tests of theories about how political information processing works.

The third assumption is that measurement of singular media formats (e.g., what falls under the umbrella term “news”), can be used as proxies for exposure to political content. The news format and industry was once the major delivery mechanism for political content, but they no longer comprise the most common instance of political content exposure in today’s media environment. Therefore, measurement strategies based on news articles, domains, platforms, or other relatively formal sources risk measuring not political content, but something else entirely. Research that does not consider this restructuring of content delivery mechanisms cannot adequately describe exposure to political content today (Matthes et al., 2020).
Missing from the literature about each of these three assumptions is a high-resolution, ground-truth understanding of smartphone users’ experiences of political content that embraces its natural complexity. We address this opportunity with a rigorous description of the rapidity and idiosyncrasy of political content exposure on smartphones. We provide rich quantitative description as a research outcome, with an aim to theoretically ground the ongoing study of complex evolving systems through empirical baseline inquiry (Jebb et al., 2017; Munger et al., 2021). We focus extensively on the political content exposure patterns both between- and within-persons. Guided by the Screenomics framework and research pipeline, we use rich observational data to measure the duration, frequency, and variation, both within and between individuals, of political experiences on smartphones, and relate our findings to popular literature.

**Smartphones and Screenomics**

We focus our analysis on smartphones as the increasingly dominant media platform of our era. The flexibility of the smartphone allows for the consolidation of virtually any task or topic onto a single screen, fragmented across time to suit moment-level changes in a user’s goals, interests, or emotional states (Brinberg et al., 2021; Oulasvirta et al., 2005). The result for each user at each moment is a stream of rapidly sequenced content and behavior that is idiosyncratic, impossible for a human to fully recall, and essentially irreproducible over time (Reeves et al., 2021).

To properly map idiosyncratic smartphone experiences, we follow recent research in utilizing the Screenomics framework to capture real-time, nuanced smartphone data from users across segments of the population. This framework is a passive and granular smartphone screenshot logging technique (Reeves et al., 2021) that cuts across formats and platforms (such as news/non-news or Instagram/X) by passively capturing all screen activity agnostically across all software applications as people use their devices. Importantly for this study, the framework provides a seconds-level temporal granularity, necessary for studying quick exposures (Reeves et al., 2020).

In essence, the Screenomics framework lets researchers watch and record media
users via an “over-the-shoulder” perspective in natural settings, enabling inspection of core media behaviors and patterns at the micro-temporal level. For example, researchers have detected emotional balancing in media selections over time (Cho, Reeves, Ram, & Robinson, 2023), documented consistency in micro-temporal usage choices (such as usage session durations) across highly dissimilar media environments (Muise, Lu, Pan, & Reeves, 2022), and even identified how payday loan advertisements cause upticks in psychological stress among vulnerable populations, by measuring subtle linguistic and structural differences in media engagement before and after ad exposures (Lee, Hamilton, Ram, Roehrick, & Reeves, 2023). Hence, the granularity of digital trace data available through the Screenomics framework facilitates investigation of several research questions related to the core theoretical and methodological interests of this paper. Given the potential sensitivity of data collected in this manner, the present work was conducted only upon attaining and developing the necessary resources for data security, data management, researcher privacy training, and clear informed consent, as described further in the Supplementary Materials and prior literature (Brinberg et al., 2021; Cho et al., 2023; Lee et al., 2024; Muise et al., 2022; Reeves et al., 2020).

As we are focused on common practices used in aggregating political content exposure both between and within persons, we particularly scrutinize the temporal variability of instances of exposure, as that variability that might challenge or clarify the reliability of substantive findings in the political communication literature. First, to establish the baseline prevalence of encounters with political content on the smartphone, we sought to measure the frequency of these encounters.

*RQ1: What is the frequency with which political content appears on the smartphone screen?*

The answer to this question is then useful for identifying how patterns of political content exposure differ across individuals. These variations are often entirely overlooked in aggregate measurements of audiences but are central to assessments of the idiosyncratic behavior enabled by smartphones. If the pattern of political content exposure is very
different across individuals, extrapolation from group-level estimates to the individual-level predictions would risk commission of ecological fallacy.

*RQ2: How does political content exposure frequency vary across and within individuals?*

This question pertains to the within-person variation that is central to the uniqueness and temporality of smartphone measurements. Summary metrics used to describe group-level and even individual-levels exposure to political content can mask meaningful complexity, risking substantive conclusions driven by outlier individuals and moments respectively. We expand on this question in Supplementary Materials, directly comparing within-person variability to between-persons variability as a benchmark.

Political content exposure on flexible media can be very short-lived relative to conventional political media coverage. We aim to provide direct evidence as to the distribution of political exposure durations, evidence that guides theorizing about how political content is processed psychologically.

*RQ3: What is the distribution of durations of political content segments experienced on the smartphone?*

As with RQ2-3, we can examine within- and between-person variation in political content duration distributions. Conventional wisdom suggests that some individuals with greater interest in politics may spend longer amounts of time viewing, but this is unclear in the literature. Examining between-person and within-person variation in durations also allows for investigation of ergodicity within variables underlying political content exposure. Ergodicity, as a property of the relationship between two variables, is the extent to which the between-person correlation between two variables is equivalent to the within-person correlation of those same variables.

*RQ4: How do political content segment durations vary across and within
Finally, we examine how political content exposure varies across applications. Smartphone users have numerous sources for political content, including private communication, social media, and news applications. The field lacks direct evidence of how political content exposure is distributed across these sources in natural usage, both within and across individuals. The final research question is:

*RQ5: How does political content exposure vary across application categories?*

Altogether, we hope to highlight the utility of the Screenomics framework for tracking political content via digital media, and also to draw attention to concerns about how selective exposure to politics is measured – each with the goal of more tightly coupling theories about political communication with changes in how content is naturally experienced. We provide the first detailed descriptions of how much political content appears on smartphones over time, and of how different application categories (e.g., news, social media) are associated with political content exposure.

**Method**

*Data Collection and Recruitment*

In the Screenomics framework, screenshots are recorded remotely and passively from a user’s smartphone at the onset of use and then every five seconds that the smartphone screen is activated. This process is accomplished with a custom data-collection application for Android phones that captures screenshots as PNG files, minimally compresses them as JPG files, and then uploads the files through secure encryption to a single cloud server. Raw screenshot files were stored exclusively on approved cloud storage or associated virtual machines and accessed only in accordance with IRB approved training and certification (please see the Supplementary Materials for additional
Commentary on how participant privacy was maintained during data collection and analysis. In addition to images, the data collection application provides a separate log identifying the name of the app in the smartphone’s foreground. Text is extracted from each frame using a task-specific optical character recognition pipeline (Chiatti et al., 2017).

The 115 research participants in this study were recruited through the Qualtrics panel aggregation service in 2019, using a quota-screening method focused on age, gender, and region to create a demographically and geographically dispersed convenience sample. Given the known demographics of the sample, the set of participants is not representative of the US or global adult population. Rather, the sample’s composition is sufficiently varied for the demonstration of screen interaction characteristics of interest in this paper; recent research has indicated that smartphone screen interaction patterns are not significantly different across demographic groups at the micro-temporal level (Muise, Lu, et al., 2022).

The recruited individuals completed a survey that included a set of additional screening questions confirming that they owned a compatible Android phone, the phone was not shared, and they were interested in participation for two weeks in exchange for a $30 incentive payment. Survey respondents who agreed to these arrangements were forwarded to a university-hosted website on which they could provide signed, written consent as approved by the Stanford University Institutional Review Board, and then follow instructions for installing the Screenomics data collection application on their own smartphones. Upon confirmation that an individual’s smartphone was uploading screenshots and application log data, a respondent was considered actively enrolled for an intended two-week period. Table 1 shows a summary of participants’ screen contributions to the raw data used in this paper.

All data collection and analysis were conducted with approval of the Stanford University Institutional Review Board (Protocol 38485).
Table 1. Overall measures of sample scale and granularity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Total</th>
<th>Mean</th>
<th>Min.</th>
<th>Med.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants</td>
<td>115</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Screenshots</td>
<td>4,907,091</td>
<td>42,670</td>
<td>1,069</td>
<td>37,819</td>
<td>178,094</td>
</tr>
<tr>
<td>Screentime (hours)</td>
<td>6,815</td>
<td>59.3</td>
<td>1.5</td>
<td>52.5</td>
<td>247.4</td>
</tr>
<tr>
<td>Participant days</td>
<td>1,463</td>
<td>13</td>
<td>6</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Daily Screentime (hours/day)</td>
<td></td>
<td>4.6</td>
<td>0.1</td>
<td>4.0</td>
<td>17.7</td>
</tr>
</tbody>
</table>

Notes. Screenshots are the basic unit of data. Screentime, shown in hours, is the approximate total screen activity of all participants estimated through screenshot count. Participant-days is the number of days of smartphone usage included in the subsample. The sample is trimmed such that all participants contribute a minimum of 6 consecutive days wherein the screen was never inactive for longer than 48 hours.

Participants

Table 2 summarizes the demographic information (available for 48 of our 115 participants) collected in a survey conducted prior to screenshot data collection. Per available demographic information, participants were split nearly evenly by gender; no participants reported nonbinary gender. Age ranged from 21 to 68 years, with a median age of 44 years. Most participants had at least some higher education past a high school diploma. We emphasize that our intent in this analysis is not to generalize specific content-exposure attributes to a broader population, but rather to demonstrate the ease with which scrutiny can reveal intense and meaningful between-person and within-person variability in content exposure.
Table 2. Sample demographics per available information

<table>
<thead>
<tr>
<th>Variable</th>
<th>Key</th>
<th>Value</th>
<th>Variable</th>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender (%)</td>
<td>Female</td>
<td>56%</td>
<td>Race (%)</td>
<td>White</td>
<td>69%</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>44%</td>
<td></td>
<td>Black</td>
<td>14%</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Minimum</td>
<td>21</td>
<td>Hispanic or Latino/a (%)</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>34</td>
<td></td>
<td>2%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>44</td>
<td>Asian (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>59</td>
<td>Multi/other (%)</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>68</td>
<td>&lt;$25k (%)</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[$25k, $50k) (%)</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[$50k, $100k) (%)</td>
<td>25%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>≥$100k+ (%)</td>
<td>17%</td>
<td></td>
</tr>
<tr>
<td>Education (%)</td>
<td>HS or less</td>
<td>27%</td>
<td>English at home (%)</td>
<td>Always</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td>Some college</td>
<td>35%</td>
<td>Mostly</td>
<td>8%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bachelor</td>
<td>25%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduate</td>
<td>13%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marital status (%)</td>
<td>Single</td>
<td>30%</td>
<td>Partisanship (%)</td>
<td>Democrat</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>Married</td>
<td>52%</td>
<td></td>
<td>Independent</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>Divorced</td>
<td>18%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-partisan thermometer (0 to 100)</td>
<td>Minimum</td>
<td>0</td>
<td>Attention to politics (%)</td>
<td>Not much interested</td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td>25&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>11</td>
<td></td>
<td>Somewhat interested</td>
<td>40%</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>23</td>
<td></td>
<td>Very much interested</td>
<td>44%</td>
</tr>
<tr>
<td></td>
<td>75&lt;sup&gt;th&lt;/sup&gt; percentile</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Demographic features for 48 out of 115 participants. Information is self-reported. Percentages that do not add to 100 are affected by rounding.

**Political Content classification**

To begin the analysis, we identified whether there was political content in each screenshot. Political content refers to any content containing any discussion of the
following topic areas (Bode, 2015; Matthes et al., 2020): the U.S. presidential administration; politicians, including prospective, elected, or appointed government officials; political satire and political satirists; partisan groups, including political parties and think tanks; elections and election campaigns (at any level from national to local); policy debates and/or decisions (economic policy, foreign policy, social policy, other policy); political ideology and/or partisanship e.g., liberalism, conservatism, socialism, progressivism, fascism, populism, libertarianism, communism, republicanism or democratism; hate speech, racial inequity, identity threats or representation; terrorism, current American wars, US foreign policy, foreign affairs; analysis of or reactions to contemporary political events. These topics are based on national, state, or municipal governance (e.g., legislation, ideology), relevant system roles (e.g., Presidents, candidates, pundits), and matters of social justice (e.g., racism, terrorism) without accounting for source, format, or factuality.

Because manual coding of 4.9 million screenshots was impractical, we employed automated classification methods. To do so, we began with a smaller subset of screenshots that was manually labelled, and then used to train a supervised classification model. The initial set of political content ground-truth consisted of 125,473 screenshots captured by 69 participants. These screenshots were manually tagged by seven research assistants (Cohen’s $\kappa = 0.59$) using the coding rules above. Screenshots were randomly selected, stratified by subject and time, by weekday, time of day, and 30-minute blocks within time of day, with an additional oversample of high-likelihood political screenshots based on the presence of basic political keywords.

We used this ground-truth set to train and measure the performance of binary classifiers of political screenshots in the broader sample. Most features associated with screenshots were not suited for an application-agnostic classifier. For example, inputting image data into a political content classifier would bias a model toward visual formats by which political content is more often consumed (Muise, Lu, et al., 2022). Therefore, text alone was chosen as a mostly format-agnostic asset for the classification of political content. In the ground truth set, just 59 political screenshots contained no words (operationalized as containing fewer than five characters), comprising 2.4% of political
screenshots and 0.3% of all wordless screenshots in the ground truth sample.

We preprocessed text features within the data to optimize classification performance and limit the influence of artifacts. Spellcheck was applied to the text from each screenshot using the hunspell package in R (Ooms, 2016). This removed obvious optical character recognition format-disclosing artifacts (e.g., logos) and other symbols (e.g., reading of the smartphone’s battery symbol) and corrected words to nearby suggestions. We trained a random forest binary classifier on our hand-tagged screenshots to identify political screenshots, using as inputs the combined presence or absence of 168 carefully selected political word-stems within each screenshot. This process resulted in an $F$ score of 0.78 (versus a performance of $F$: 0.74 for an optimized Bag-of-Words approach and $F$: 0.50 for a text-embedding based approach). The confusion matrix of our classifier presented the following breakdown: TN: 95.9%, TP: 3.6%, FN: 1.4%, FP: 0.6%. Through manual inspection, we found that false positives were driven by political terms used in surprising and edge-case contexts. Through the same inspection, we found that false negatives were driven partially by edge-case political discussions of rare topics that made no use of any of our 168 stems, and to a lesser extent, fully unlabeled political imagery. We detail our model development and selection process in full the Supplementary Materials.

**Operationalizing Political Content Exposure Measures**

For clarity, we use the following terminology throughout. A political content encounter is an event in which political content appears on-screen when it was not on-screen previously. This was measured by the presence of a screenshot (i.e., roughly five seconds of screentime) containing political content that is preceded by inactivity or a screenshot that does not contain political content. A segment (of political content exposure) is the span of time over which political content stays on screen following an encounter. For example, a political content encounter followed by seven screenshots, each five seconds apart, represents a segment of political content exposure lasting 40 seconds (8 political
screenshots total × 5 seconds per screenshot$^3$). The number of segments is definitionally equal to the number of encounters. A single segment can traverse multiple applications or URLs (but usually they do not). Segments are bounded by temporal separation of at least one screenshot not identified as containing political content, or any gap in screen activity (i.e., a break between sessions). This method is based on current literature which acknowledges multiple methods for determining gaps in smartphone behavior, but also acknowledges that the measurement value of more sophisticated approaches is unclear (Peng et al., 2020; van Berkel et al., 2016; Zhu et al., 2018).

Throughout this study, we focus on the prevalence of segments lasting only five seconds, which is the minimum detectable duration of political content exposure in our data. These short durations, while not instantaneous, are representative of extremely rapid exposure to politics.$^4$

**Smartphone Application Category Classification**

The 2314 unique applications encountered by participants were grouped into 24 categories. The app categorization schema is based on app developers’ self-declared categorizations on the Google Play Store, the default source for Android applications. We manually adjusted some categorizations to better fit the goals of this work. For example, to categorize applications used by participants to earn or gamble small amounts of cash, we created a category ‘survey/cash’ which supplants the myriad labels of these applications created by individual app makers (e.g., “Lifestyle” for Swagbucks, “Social” for YouGov, and “Productivity” for SurveyMonkey).

The primary use of this app classification system was to examine the degree to

---

$x^3$ Given our five-second frame-interval, we treat each political screenshot as representing five seconds of political content exposure. Specifically, we assume that a political content segment captured by an instantaneous screenshot at moment $t$ has a uniform probability of having begun at any point from $t - 4.9$ seconds to point $t - \epsilon$ and having ended at any point in $t + \epsilon$ to $t + 4.9$ seconds, assuming no preceding or following political screenshot is captured, respectively. This reduces to bounds of $t \pm 2.5$ seconds, i.e., a five-second segment.

$^4$ We also performed a robustness check by flexing the operationalization of a political content segment. Specifically, we introduced an extra 5 second tolerance in segment separation, so that, for example, political content segments separated by a non-political screenshot would be combined into one segment. This robustness check produced no substantive change in the findings reported in the current paper.
which political content is encountered via news applications or social media applications. The former category (e.g., the CNN app, the Google News app, or the New York Times app) is of interest given the industry’s historic role in conveying political content. The latter category (including Facebook, Instagram, and Reddit) is of interest given the attention it has garnered from academics and the public alike, specifically as a common source of low-quality political content.

The category Browser (including Chrome, Safari, and Edge) can represent a variety of behaviors, given that any website with any content can be accessed via URL (including, e.g., CNN.com or Facebook.com). While we treat mobile browser applications as a standalone category, behavior on mobile browser applications follows an unknown distribution across all other application categories.

Results

RQ1: Political Content Exposure on Smartphones

To address RQ1, we describe sample statistics regarding the frequency of viewing political content for the entire sample. Political content appeared infrequently. Out of 4,907,091 screenshots captured across all 115 participants, only 92,988 contained political content of any type. This is equivalent to 1.89% of all smartphone use, or one minute and eight seconds out of every hour of screen activity, a value similar-to but lower than the estimate of 2.78% by Allen et al. (2020). The number of political content encounters on the smartphone and equivalently, the number of segments, was 26,238 — spread across 12,753 unique sessions.

Is the frequency of political content encounters uniform over time? Figure 1 shows a histogram of encounter times throughout the day in the entire sample, broken into 15-minute blocks, standardized to each subject’s local time. On the 24-hour clock (midnight to midnight), encounters are broadly distributed through the standard waking hours and beyond. There is not a single block of 15 minutes in which a political content encounter did not occur. This picture contrasts sharply with older models of political communication.
founded on morning and evening primetime news consumption. In particular, the prevalence of political content encounters in the overnight and workday hours makes clear that political content is untethered to traditional divisions of a media day.

Figure 1. Political content encounters on the smartphone screen across 115 participants. Note. Encounters are bucketed into 15-minute blocks spanning the entire day. Note that ‘encounter’ refers to only the first screenshot in a series of continuous screenshots in which political content is identified.

The variability of political content exposure across participants and across participant-days is shown in Figure 2, which depicts a week of smartphone screen activity, spanning 7 contiguous days (rows) for five participants (columns) selected manually to illustrate between-person within-person variation. Each of the 35 individual panels shows a 24-hour clock (midnight to midnight) along a single horizontal timeline. Black color indicates smartphone screen activity of any kind; red color indicates the presence of political content on screen. Across the five participants (columns) of Figure 2, no two participants share similar smartphone behavior, and within that behavior, they do not share similar political content exposure.
Equally interesting is the degree of variation *within* individuals. Some participants show significant fluctuations in screentime across consecutive days. With increased granularity in time — days to hours to minutes — the likelihood of political content being consumed with any particular frequency is increasingly random.

![Figure 2](image)

**Figure 2. Between-person and within-person variability of political content exposure.**

*Note.* Visual illustration of between-person and within-person variability of political content exposure (red) in context of all screen activity (black) in 35 participant-days. Five columns represent five participants; seven rows represent seven consecutive days for each. Participant 43 on the left is an extremely frequent smartphone user and political content consumer, with active screentime spanning whole days (implying interrupted sleep) and political content consumed in bursts and fast instances across several applications at several hours. Other participants in this figure follow a more intuitive distribution of social media usage and political content exposure (i.e., occasional usage during daylight hours), yet the specific sequencing of activity and inactivity in each participant-day is unique.
RQ₂: Political Content Exposure Frequency Across & Within Individuals

RQ₂ examines between-person variation in exposure to political content. In Figure 3, we ranked each participant according to political content exposure, measured in terms of daily seconds (Panel A), unique encounters (Panel B), and percentage of daily screentime (Panel C). Participants are organized according to their ranking in Panel A, and the ranking is held constant in Panels B & C such that a point representing a participant s always appears at the same horizontal level. The five participants that were highlighted shown for illustration in Figure 2 are highlighted in red in Figure 3.

The most prominent finding visualized in Figure 3 is that most political content exposure is experienced by relatively few key participants, with most participants exposed to negligible amounts of political content through their smartphone. Recall that political content exposure comprised 1.89% of all screentime in the sample. In Panel C, we see that Participant 31 lies very close to this group-level average, and many participants are lower, with four participants encountering no political content. However, those with higher percentages range far beyond and exhibit a markedly different experience of politics. Grouping these participants together within single statistic provides an inaccurate portrayal of the phenomenon. Research that relies on averaging political content exposure across this distribution may assign political content into diets that contain virtually none, and thus can overestimate the level of political interest in a population. We present results that assess the utility of group-level summarization in this sample in the Supplementary Materials out of space considerations.
Figure 3. Between-person variation in daily political exposure. Note. Between-person variation in exposure to political content in terms of daily seconds (Panel A), percentage of overall daily screen-time (Panel B), and daily frequency of unique encounters (Panel C). Daily values are averages calculated across all participant-days, per participant. In Panel A, participants are ordered according to their x-axis value, with no other meaning contained in the y-axis; in Panels B and C, participants maintain their ordering from Panel A. Note the logarithmic scales on Panels A and C, and the corresponding participants shown from Figure 2.

RQ3: Durations of Political Content Segments on the Smartphone

To answer RQ3, we analyzed how long political content segments typically last on the smartphone. Drawing on prior work (Goldstein et al., 2011), we give particular focus
to segments lasting just five seconds: the minimum detectable duration with our measurement framework.

To begin, we tabulated all political content segments in the sample and plotted their distribution (Figure 4). Political content segment durations follow a power law distribution, much like durations of sessions of smartphone usage (Muise, Hosseinmardi, et al., 2022) and segments of app usage (Rula et al., 2015). In bold black is the pooled distribution of segment durations for all segments. In low opacity are the distributions of segment durations for each of the 111 participants who encountered political content. As with prior figures, we highlighted five participants for continuity and clarity. In the entire sample, 44% percent of political content segments lasted just five seconds, and 61% lasted ten seconds or less. Only 16.97% lasted longer than thirty seconds. As a robustness check on this result, we note that the 26,238 unique segments detected in the sample occurred across 12,753 unique sessions. The large number of unique sessions in which segments were detected implies that the shortness of segments does not reflect longer latent segments split up into chains of short segments by the detection strategy.

Figure 4 makes clear that the pace of the smartphone experience is in clear contrast to the news formats that formed the basis of conventional measurement strategies in political communication. An extremely small proportion (< 1%) of political content encounters lead to segments lasting one minute or longer. The wealth of encounters with political content on the smartphone lead to extremely short-lived segments. While the impact of these short segments is not directly measured in these data, they do allow estimation of the impact of overall exposure. Out of the 26,238 political content segments detected in the sample, 11,545 lasted just five seconds (44%). These one-off encounters represent 12% of the 92,988 political screenshots detected, or equivalently, 12% of all time exposed to political content. An additional 17% of 26,238 encounters lasted ten seconds, representing 9.5% of the 92,988 political screenshots in the sample. Estimated recall and recognition rates in sequence are 12% and 14% for screen-based advertising lasting five and ten seconds respectively, measured a few minutes after exposure (Goldstein et al., 2011).
Figure 4. Distribution of political content segments across duration lengths.

Note. Segment length distribution for all segments in the data is shown in bold; each participant’s individual distribution is underlaid in thin lines to show between-person variation. Forty-four percent of all political content segments lasted just five seconds, the minimum detectable duration in the sample, and a duration at which recall and recognition rate is as low as 12% just minutes after exposure (Goldstein et al., 2011).

Thus, we can infer that most political content encounters occurred in a manner that does not lead to significant information retention, comprising more than a fifth of all political content exposure in terms of time on screen. Considering the very low likelihood that these instances of exposure lead to elaboration, processing, or opinion formation (based on information retention rates alone), their contribution to political learning or socialization is unclear apart from the invocation of fast-thinking processes (Kahneman, 2011). Moreover, given the broad temporal spread of segments across sessions, any quick impressions left by fast segments are apparently unlikely to compound by occurring together in bursts.
**RQ:** Political Content Segment Durations Across & Within Individuals

As shown in the background of Figure 4, each participants’ experience of political content on the smartphone is fast-paced. However, there are crucial differences across participants that are meaningful for the study of political communication. In Figure 5, we plot the mean, median, and maximum segment duration for all 115 participants in Panels A, B, and C respectively. Participants in Figure 5 are ordered according to mean segment duration in Panel A. Five participants are highlighted for clarity and continuity with prior figures. Not shown, the minimum segment duration for all but one participant is 5 seconds or none at all. All participants’ mean and median segment duration was under one minute long, and the majority of participants’ median segment duration was five seconds.

For the majority participants, the longest political segment was shorter than three minutes, and for many participants the longest segment was under one minute (Panel C). The longest single segment in the sample lasted 29.7 minutes. This was a midnight viewing of the Joe Rogan Experience via YouTube. That the Joe Rogan Experience drives the longest-lasting political segment in the sample is illustrative of how the smartphone environment differs from the traditional media channels upon which conclusions about political communication are based. Participant 43, the participant who experienced the most encounters with political content and the most political content exposure overall, ranks 26th out of 115 participants in terms of mean segment duration and has a median segment duration of ten seconds. This finding challenges the intuitive assumption that brevity of duration of engagement with political content is a proxy for disinterest in political content. Instead, the finding is in accordance with the ‘snacking’ phenomenon described by Molyneux (2018).
Figure 5. Mean duration (Panel A), median duration (Panel B), and maximum duration (Panel C) of political content segments for 115 participants. Note. Maximum duration is presented on a logarithmic scale. Participants are ordered according to mean segment duration. Five participants are highlighted for clarity. Most participants’ median segment duration was five seconds.

Next, we examined the ergodicity of relationships underlying political content exposure. Ergodicity is a property of a relationship between two variables; specifically, it is the similarity of the correlation when calculated at the between-persons levels and when calculated at the within-person levels. When undetected, non-ergodic relationships between variables can lead to inaccurate interpretation of results. For example, a correlation between individual differences in social media use and individual differences
in political engagement in a population does not mean that the prototypical individual’s political engagement levels fluctuate with their own social media use, day-on-day. In Figure 6, we plot the pairwise Pearson correlations of segment duration, encounter frequency, and exposure time both between-person and within-person. To do so, we first calculated each variable’s value for each participant-day, and then calculated the within-person correlation of these day-level variables for each participant, producing a single correlation value for each variable pair. Separately, we calculated each participant’s three average variable values through their entire data collection period and then calculate the between-person correlation between those averages. The result of this process is two bivariate correlations for each variable pair: the average within-person correlation and the between-person correlation of average behavior.

The results are displayed in Figure 6, along with identical ergodicity analyses conducted at the hourly level rather than the daily level. Days or hours in which participants encountered no political content are treated as missing values. All variables are standardized (z-scored) to mean 0 and standard deviation 1 immediately prior to correlation calculation, between-person or within-person.

In all three cases, and at both temporal granularities, the relationship between each variable pair was ergodic. This is clear from the identical polarity and similar magnitudes of correlations calculated between-person and within-person. In Figure 6 (Panel A), we show that the correlation between average segment duration and daily encounter frequency was 0.22 both between-person and within-person at the daily level, and 0.18 at the hourly level. This means individuals with longer average segment durations tended to have more daily encounters, but also that for the prototypical participant, on days (and hours) with more encounters than usual, the participant tended to experience longer segment durations than usual. This relationship was weak (only 0.22) but striking in its robustness to level of analysis.

In Panel B, we compared average segment duration with daily exposure time. Between-person, the correlation between these two variables (as subject averages) was 0.35. Within-person, however, the correlation was 0.57 at the daily level, and 0.60 at the hourly level. While this relationship was ergodic, it was clearer within-person than
between-person. This is intuitive: on days in which the prototypical participant experienced a higher average segment duration than usual, they also experienced a higher-than-average amount of exposure time. At the hourly level, single segments comprised a larger share of overall exposure time within the hour; at faster timescales, this within-person correlation value is expected to increase further.

Figure 6. Between-person (Interindividual) Versus Within-person (Intraindividual) Pairwise Correlations. Note. Between-person correlations are calculated from the participants’ average (mean level) behavior in the data collection period. Within-person values are the average of intra-subject correlations between variables across days. Correlations are Pearson method, with standardization immediately prior to estimation. Correlations with opposing polarity represent non-ergodic pairwise relationships; all relationships display ergodicity. This may be due in part to the bounded nature of the variables (i.e., durations cannot be negative).

In Panel C, we show the relationship between daily encounter frequency and daily exposure. Between-person, these variables were almost perfectly correlated: participants with the highest daily encounter frequency had the most overall exposure. For the prototypical participant, on days (or hours) they encountered political content more
frequently than usual, exposure time was also higher than usual. This within-person relationship weakened with decreases in granularity, as fewer encounters can fit into increasingly small timespans. Altogether, the between-person relationships found in each pairwise comparison confirmed intuition, and the within-person relationships were extremely similar to their between-person counterparts, suggesting that the core components of political content exposure in the smartphone environment were ergodic in time. That is, to the extent that cross-sectional relationships are used to describe or predict within-person change, the results also appear to describe the within-person exposure dynamics. Finally, we examined how participants’ within-person distribution of political content segment durations compared with their within-person distribution of session durations. A positive correspondence between these two variables would suggest participants’ temporal engagement with political content is an artifact of an individual’s more general temporal relationship to their smartphone.

To conduct this analysis, we applied an approach from earlier research (Muise, et al., 2022), using what those authors referred to as a rapidity score. A rapidity score is the exponential degree which minimizes mean squared error (MSE) when fitting a negative exponential distribution to the distribution of session [segment] durations. This value summarizes the relationship between session [segment] duration and frequency. The higher the rapidity score, the greater the propensity for shorter durations relative to longer ones. As rapidity scores are dependent on duration distribution range, we fix the range for both segments and sessions to [0, 20 minutes] discretized into five-second buckets.

In Figure 7, we compared the segment rapidity scores to session rapidity scores for all 111 participants who encountered any political content. Along the x-axis, segment rapidity scores ranged from 0.05 to 10 along a logarithmic scale; along the y-axis, session rapidity scores ranged from 0.05 to 2.1, also along a logarithmic scale. Five familiar participants are highlighted in color for clarity and continuity. Note, rapidity scores at low integer values imply well-known distributions. A rapidity score of zero (not found in any subject) represents a white noise pattern in which all session durations [segment durations] occur equally frequently. In particular, a rapidity score of one means that sessions [segments] follow a pink noise pattern, common in film scene durations (Cutting et al.,
2010) and underlying how humans allocate attention (Van Orden et al., 2003).

Figure 7. Participant-level scatter plot of rapidity scores of segment durations against rapidity scores of political content segment durations. Note. A rapidity score is an MSE-minimizing degree describing the relationship between the frequency (commonness) of an event and the duration of the event ($duration = (freq)^\text{degree}$). Higher rapidity scores imply greater tendency toward faster events versus longer ones. If a correlation exists between session rapidity and segment rapidity, then the lengths of participants’ segments might reflect underlying tendency toward longer durations of screen activity, or longer attention spans. However, there is no correlation. Participants with no political content encounters are not included in this plot.

In this sample, sessions of smartphone use were largely distributed in a manner that resembles pink noise (rapidity score $\approx 1$). That is, for most participants, the likelihood of a smartphone session lasting for duration $f$ was approximately inversely proportional to $f$. In contrast, segment rapidity scores tended to be higher. This was expected given that 44% of all segments lasted for the minimum duration in the data. The three extreme outliers along
the x-axis, with segment rapidity score = 10, represent participants with very few encounters with political content, and thus the overwhelming majority of their segment durations were just five seconds long. This phenomenon is also reflected in the segment rapidity score of Participant 46, shown in blue, who encountered the least political content out of the five highlighted participants. Ultimately, the two sets of subject-level rapidity scores were not meaningfully correlated ($\rho = 0.045$). This lack of correlation suggests that the temporal attention that participants gave to political content on screen was not related to the way they temporally attended to the smartphone screen, in general. Rather, as Participant 46 demonstrates, a smartphone user who distributes their smartphone use across sessions of varied durations may consistently restrict their attention to political content to extremely short windows. Participant 43 shows something of the opposite behavior. The core takeaway is that duration of attention to political content on the smartphone screen is more than a manifestation or artifact of overall screen behavior.

**RQ5: Political Content Exposure Across Application Categories**

We lastly examined how political content encounters were distributed across all application categories, including applications categorized as *news* and *social media*. To do so, we first provided context by examining participants’ overall allocation of time across application categories, political or not. Applications in the sample are categorized primarily according to their designations on the Google Play Store, with some manual adjustments to accommodate the specifics of the present task. Foreground application data were not collected from six participants due to technical problems with the data collection application.

In the four panels of Figure 8, we decomposed participants’ screen use and political content encounters across 23 application categories. In all four panels, the within-person distribution of screen activity from 109 participants is shown as a series of 23 boxplots with underlying subject-level data shown in one-dimensional low-opacity scatterplots. In all four panels, five familiar participants are highlighted in color for clarity and continuity. In Panel A, we show the proportion of screenshots (i.e., overall screen activity) from each
application category. For example, the top row of Panel A shows that Participant 31 (in cyan) allocated just over 30% of their screentime to communication applications, and across 109 participants, the median percentage of screenshots captured from communication applications was 16%. In all four panels, application categories are arranged on the y-axis in descending order of participants’ median time allocation, and the x-axis shows percentage on a log scale. In Panel B, we show the intra-category percent of screenshots that contain political content. In Panel C, we show the percent of political screenshots that came from each category. In Panel D, we show the percent of political content encounters that came from each category.

Panel A makes clear that participants had highly dissimilar within-person distributions of screentime across application categories. Even in the most common application categories, namely communication, browser, and social media, participants’ within-person allocations ranged from 0% to over 50%. Moreover, every application category hosted two or more outliers in terms of within-person time allocation, emphasizing the idiosyncratic nature of even the simplest metrics of smartphone usage. The category gaming was unique in that the median time allocation to gaming was just 1%, yet one-third of the sample allocated up to the plurality or even majority of their time to mobile gaming. The Screenomics application category refers solely to the data collection application used in this study. Notably, news applications were highly uncommon, such that only outlier participants allocated any screentime to news applications at all. Subject 43, in red, is one such outlier, spending 10% of screentime on news applications.

Panel B shows the percentage of screenshots within each category that contain political content, with boxplots again summarizing the underlying subject-level datapoints. Communication, browser, and social media applications are not only the most used applications overall, but also contained the highest percentage of political content in terms of median subject-level time allocation. Even still, median time allocation to political content within any category was under 1%, and for most categories, the median subject spent 0% of screen time exposure to political content. Apart from three outlier participants, no application category was used mostly for political content consumption by any participant. And for the outlier participants, the application category in question was not
For all participants who used news applications, the majority of screenshots captured while news applications were on screen were not classified as containing political content. Moreover, the frequency of political content exposure within the news category was similar to that of application categories not ordinarily associated with political content, such as entertainment and lifestyle. In the smartphone environment, these categories were all similarly likely to provide access to political content, notwithstanding wide variation across participants’ individual-level behavior. Subject 43, the subject who was exposed to the most political content in the study, spent 27% their news application screentime on political content. However, they were also an outlier in several other application categories, implying that political content was a common feature of a wide array of formats and spaces in their personal smartphone environment.

In Panel C, we show how each participant allocated political content exposure across application categories. Here, participants’ within-person distributions were widely varied. For the median subject, 8% of political content exposure occurred via communication applications, another 8% via browser applications, 3% via social media, and negligible amounts from all other application categories. While browser applications may host any type of content available via URL, communication applications provide political content via incoming campaign emails, political newsletters, text conversations, memes in group chats, and so on. In stark contrast, the median participant sourced 0% of their political content exposure from news applications, though several outliers such as Participant 43 sourced the plurality from news applications. In the entire sample, only two participants sourced the majority of their political content exposure from news applications.

Some participants source most of their political content exposure from applications that are not traditionally associated with political content, including cash/survey applications but also games and tools wherein political content may arise in gameplay, advertisements, or even notifications. For Participant 43, roughly half of political content exposure came from browser applications, with the other half split between news applications and entertainment. Lastly, Panel D shows how each participant allocated their political content encounters across application categories. This view illuminates the unique
instances and opportunities that participants have to engage with political content, regardless of exposure time. The similarity between Panels C and D in part reflects that 44% of political content encounters did not lead to segments longer than five seconds. The only exception is within the communication category, in which the median subject sourced 12% of their political content encounters versus 8% of their political content exposure. This difference implies that a sizable share of political content encounters occurring via communication channels did not result in engagement, as proxied by exposure lasting more than one screenshot.

Finally, we provide a between-person breakdown of screenshots containing political content (or not) with screenshots containing a news application (or not). While there was variation across participants, for the mean participant, only 1.13% of screenshots came from news applications, and just 0.15% of screenshots contained political content via news applications. Rather, 1.68% of screenshots contained political content accessed via other application categories, meaning only 9% of exposure to political content occurred while a news application was on-screen, for the average subject. To the extent that news applications are the successor of traditional news formats, the study of politics via news in the smartphone environment is focused very narrowly on the 0.15% of screenshots in the average subject in this sample. This finding strongly suggests that news applications are an inappropriate tool for measuring individuals’ exposure to political content on smartphones.
Figure 8. Decomposition of screen exposure and political content exposure.

Note. Decomposition of overall screen exposure and exposure to political content across 23 application categories for 109 participants with foreground application data. Each participant is represented as a single point 23 times in each panel, with boxplots summarizing participants’ values within each application category. X-axes are explained in panel titles and shown in log scale; y-axis values are 23 categories in descending order.
Discussion

We aim to provide a missing description of political content exposure in the smartphone environment. Here, we summarize our various key descriptive findings as they pertain to each of the three assumptions investigated in this paper.

Regarding the first assumption, we observed that the conventional approach of treating unique instances of exposure as commensurable units of analysis is not appropriate for the study of modern (smartphone) media (Lazer, 2020; Tewksbury et al., 2001). A key assumption in this line of work is that each encounter leads to a similarly meaningful segment of exposure, where similarly meaningful, however construed for a particular research goal, largely implies durations of similar magnitude. The results in the current work suggest that this assumption is untenable. Most exposure to political content is driven by a small share of participants, suggesting that group-level means aiming to describe a single typical individual’s experience fail to capture any given individuals’ actual exposure levels. Moreover, participants’ within-person variation in political content exposure, measured as unique encounters or as percent of screentime, is on par with between-person variation. On average, an average measure of a subject’s daily political content exposure is inaccurate by a factor of 4 in estimating any particular hour of the subject’s experience. Thus, even within-person estimates of political content exposure are increasingly inaccurate when applied at shorter time spans, tracing the value of timescale explication in measurement.

Importantly, the durations of political content segments can plausibly co-vary with researchers’ outcomes of interest, implying serious impacts on substantive findings. Consider recent research which measures overall partisan segregation on the Facebook platform by tabulating users’ encounters with copartisan and counterpartisan political news, using an encounter-based segregation index (González-Bailón et al., 2023). By
basing exposure measures on encounters (and not segments), the core analysis depends on an implicit assumption that users’ attendance to political content does not meaningfully vary, including variation according to the partisan-congeniality of a given news encounter. It is conceivable that users attend more thoroughly to partisan-supporting news, in which case an encounter-based assessment of a user’s news diet would result an underestimate of the bias laden in a Facebook user’s true content consumption. Such a bias would of course lead to an underestimation of overall partisan platform segregation. This is more than a methodological point; the unit of analysis determined by researchers, or in this case made available to researchers, reflects a theoretical understanding that content exposure is to be measured in terms of human-content interaction rather than media’s relevance to a human’s experience.

The second common assumption challenged by our findings is that segment durations sum linearly to a measure of total exposure either within-person or between-person (as in Allen et al., 2020; Guess, 2021; Muise et al., 2022b). Measurement strategies which aggregate political content exposure without regard for the duration of individual segments risk equating entirely different psychological processes (Lemke, 1998). In this case, a subject’s total exposure time to a particular source of information (e.g., news consumption in a month, or exposure time on one partisan source versus another) is assumed to be a straightforward additive composite of all individual segments. This approach is more rigorous than an encounter-only approach in terms of measuring information dosage. However, given the extremely short durations of most segments in smartphone environment, and the low feasibility of information retention and engagement from these segments, linear summation still poses a bias risk in any exposure estimation. Only a negligible amount of political content segments last longer than one minute, unconditional on application category, and consistent across individual, occurring amidst a context of more rapid exposures. Evidence is not in favor of these briefer encounters’ impacts adding linearly toward political outcomes: retention is shown to grow non-linearly with increased exposure time, and more theoretically, different processes simply occur at different timescales (Goldstein et al., 2011).

Leniently supposing that recall and recognition rates following a segment lasting at
least ten seconds (approximately 14%) is sufficient for a political communication researcher’s outcome measures, then 9.5% of all exposures in this sample were below a threshold for inclusion in linear summation, as they are comprised of five second segments. Supposing slightly less leniently that a political outcomes’ necessary recall and recognition rate is greater 14%, then all ten-second segments are invalid for inclusion, meaning that 17% of all exposure time in this sample is not valid for inclusion in such an analysis. By including such short segments of exposure into aggregate temporal measures, a researcher would be introducing an upward bias, including to particular kinds of content such as partisan-biased content. Researchers using linear summation of political content exposure must consider their own threshold for how long is ‘long enough’ in determining which segments in include or exclude from total counts. This decision is specific to individual research goals.

Muise et al. (2022) approximate this strategy when mapping-out the month-on-month evolution of Americans’ online news diets, particularly their partisan-bias composition. The authors exclude individuals from analysis in a given month if their online news consumption amounted to less than two minutes. This was because the bias composition of such small news diets would artificially fluctuate month-on-month as an artifact of how few news sources could be possibly fit into two minutes. Moreover, those individuals were likely not attending to politics intentionally or meaningfully given such slight news diets. To extend this logic to segments, the authors could have identified which seconds of exposure were derived from segments lasting only a few seconds, versus longer segments. Such a control would ensure that each individual’s measured monthly online news diet more closely represents their own intended, or indeed attended, political content consumption, insofar as that consumption may be retained in memory.

Regarding the third assumption, we find that no single application category is a majority source for political content, news notwithstanding, thus calling into question the use of ‘news’ as a proxy for political content. Between-person, most participants who were exposed to political content never opened a news application. Inversely, the majority of news applications produced screenshots that had no political content. In our sample, just 0.15% of an average participant’s smartphone usage is spent on political content within
news applications. Allocation of political content exposure across application categories is extremely varied across individuals, and longer segments were driven mainly by social media rather than news applications. In general, across participants social media is a much more prominent source of political content than news applications. The large degree of between-person variation in application category repertoires means that any singular application-based approach to measuring whole exposure to political content would be biased in a manner idiosyncratic to each subject.

Recent high-powered experiments done in collaboration with Meta altered the frequency and nature of participants’ political content exposure within the Facebook and Instagram news feeds; each found no statistically significant effect on users’ political attitudes (Guess et al, 2023a, 2023b, Nyhan et al., 2023). Our findings better contextualize these experiments’ surprising null findings: social media apps accounted for roughly 6% of smartphone usage time within our sample, and roughly 4% of political content exposure time. This implies that a Facebook-only field experiment would be manipulating only a single-digit percentage of users’ overall media diets, and overall political content diets. Moreover, we found that while 4% of political content exposure occurred on social media apps, roughly 8% of political encounters occurred on social media apps. Even while noting the extreme variability surrounding these summary values, this implication is that the political content segments which do occur on social media are actually more likely to be shorter than segments occurring elsewhere in the smartphone ecosystem. Lastly, these large experiments faced a restriction of analyzing only social media posts associated with a political news URL, which accounted for only 3.9% of post-views on the platform (González-Bailón et al., 2023). As our data suggest, such a restriction yields a political content detection schema with low recall. The distinction between political content and the vehicles through which it is delivered is crucial for our field to recognize; simultaneously, the differential impacts of these vehicles on the reception of political content is an ever-expanding area of interest.
The size and scope of our convenience sample limits generalization specific estimates and values uncovered here to broader populations, as does our recruitment method, which is in-turn reliant on gig workers whose smartphone use may be unusual. Our participant pool did not include any iPhone users, nor Android users unwilling to participate in a longitudinal screen-recording study for $30. How this selection may be related to smartphone usage habits is not clear. Functionally, iOS and Android devices grant extremely similar experiences of content retrieval, navigation, and display within apps and mobile browsers, which is the focus of the present work. Demographically, iOS users tend to skew younger and wealthier in the United States (Brown, 2020; Williams, 2018), which are both factors weakly correlated with political interest in countervailing directions. From the demographic information available about our sample, we do not see a substantively large potential bias present for each demographic factor. The core concept we study in the present work — exposure to political content — is not assumed to be sensitive enough to each user that they would endeavor to hide or promote such usage because of concern an observer. If participants were self-censoring non-political behavior, then this would only meaningfully impact estimates of political content exposure as a percentage of overall screen-time, which is already quite small.

There are various opportunities to expand on this paper’s method in future research. Most obvious is an expansion into greater temporal granularity, in order to capture and document sub-second-long natural encounters that cannot be properly observed with our 5-second resolution. Conversely, expanding the temporal breadth of data collection, i.e., running data collection for longer periods, would allow for the capture and analysis of longer cycles and variations in contexts and content exposure behavior, especially within-person. Expansion into other media platforms, especially other kinds of screens, would bring researchers closer to an ultimate accounting of all political content exposure experienced by an individual. As for future directions within our substantive analysis, researchers could build larger and more dedicated samples to solidify findings and inferences made here. Our sample, though large enough for our purposes in this paper, was
not large enough to extrapolate specific findings to whole populations, and was undoubtedly influenced by its ‘convenience sample’ origins. Nor did our sample have a critical mass of individuals with a dedicated interest in political content, meaning we could not focus on that interesting and meaningful subset.

Finally, using the Screenomics approach, we saw video titles and captions during such relevant sessions, but we were unable to capture actual audio and video (e.g., if text representing such media did not appear on the screen, it would be missed). This is a noteworthy concern (see Gitomer et al., 2023), one that future approaches and studies should seek to address.

**Conclusion**

Modern communication technology affords incredibly fast and varied exposure to political content. Thus, measurement of exposure must be adapted to the new media environment. Political content exposure is best measured in seconds, meaning that audiences’ habits for understanding the political world is butting up against the lower temporal bound of genuine cognitive engagement, and straddling faster types of processing with timescales all their own. An encounter with political content can still lead to a thirty-minute viewing session on television, or an hour-long radio program, but to move forward in the field, we must incorporate the flash of a meme, or a comment scrolled-by in a feed. Each of these instances means that media effects are likely determined as much or more by temporal scale of message exposures as by message content.

**Acknowledgements**

The authors thank the Knight Foundation and the Stanford University Cyber Social Initiative for funding this work. For assistance in research operations, we thank MJ Cho, Jihye Lee, and Hanseul Jun, as well as the many research assistants who worked to describe screenshot content. For a wide range of shared efforts, we thank all affiliates of the Human Screenome Project at Stanford University. For advising throughout, we thank Jennifer Pan.
For their feedback and comments, we thank Ronald Robertson, members of the Polarization and Social Change Lab at Stanford University, attendees of the 2017 Computation + Journalism conference, and attendees of the 2022 ICA Conference. At the time of publication, lead author Daniel Muise is the CEO of Screenlake, Inc, a research company based on granular behavioral data collection and analysis from smartphones. The research in this work was conducted prior to and independently of Screenlake, Inc.

References


Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y.,


Van Damme, K., Martens, M., Van Leuven, S., Vanden Abeele, M., & De Marez, L. (2019). Mapping the Mobile DNA of News. Understanding Incidental and


Supplementary Materials for
(Mis)measurement of Political Content Exposure within the Smartphone Ecosystem: Investigating Common Assumptions

DANIEL MUISE
Stanford University, USA

DAVID M. MARKOWITZ
Michigan State University, USA

BYRON REEVES
Stanford University, USA

NILAM RAM

THOMAS N. ROBINSON

Table of Contents
Privacy and Data Access Commentary.................................................................2
Additional Details on Automated Classification..................................................3
Additional Details on Manual Tagging.................................................................7
Results Related to Group-Level Summarization................................................9
App Categorization Schema...............................................................................11
References.........................................................................................................12

1Daniel Muise: dmuise@alumni.stanford.edu
Date submitted: 2024-03-03

At the time of publication, author’s affiliation is Screenlake, Inc.
Privacy and Data Access Commentary

Regardless of what application was in the foreground, the Screenomics software captured all content and information that was visually rendered on the screen, instantaneously every 5 seconds. Participants were made fully aware of this over and over, were able to opt-out of the study at any time via a large “STOP” button in the app on their device or via removing the app entirely, and participants could also request deletion of their data. When active, the data collection app would display an icon along the top of a participants’ smartphone screen, similar in size and location to the WiFi or battery icons, indicating that the app was taking screenshots.

When human-coding these data, we used a clean-room setup where coders were only allowed to be in a password-protected environment when accessing the data. Each coder had medical research human-subjects and HIPAA privacy training and certification. All data entry was done directly in the primary institution’s secure and HIPAA-compliant Medicine Box installation with two-factor authentication, meaning only specifically approved research personnel were allowed to log in, and only from devices that had been installed with the primary institution’s comprehensive security and activity logging software.

---

2 Some apps, most notably those maintained by Snapchat and Netflix, are designed to redact screen content in any screenshot. E.g., a screenshot captured from a smartphone displaying a Netflix video would accurately reproduce closed captions and video player controls that may be on-screen at the time, but would display black in place of any video content. Other applications redact the screen entirely.
Additional Details on Automated Classification

To select the best classification method, we compared four approaches applied to the ground truth data: (1) bag-of-words based on 153 initial manually chosen stems related to political content (BoW); (2) a manually improved bag-of-words approach based on expansion and restriction of the initial stems (yielding 168 stems); (3) supervised machine learning based on combinations of presence or absence of 168 stems in screenshots; and (4) supervised machine learning based on pre-trained text embedding applied to all text extracted from screenshots (Rudkowsky et al., 2018). Initial word stems underlying approaches (1) through (3) were based on a manual review of top political topics being discussed in late 2019 (the period of data collection), as accessed through Google searches of top news articles. The initial bag of words approaches (approaches 1 & 2) employed a hybrid identification method that combined intuitive word selection with repeated manual auditing, similar to the approach used by Munger et al. (2022).

Approach (1): Intended as a starting point, the appearance of at least one of our initial 153 political stems would result in a screenshot being flagged as containing political content. The results were then compared to the ground truth set to detect initial accuracy metrics. The $F$ metric is a more effective performance metric than accuracy in highly unbalanced classification tasks (Weiss, 2004). In Approach (1), $F = 0.53$.

To move from the initial stem set of Approach (1) to the improved set of Approach (2), we manually examined content of screenshots in the training data which were incorrectly flagged as containing political content (false positives), as precision of Approach (1) proved to be very low ($P=0.39$). We incrementally removed or down-weighted stems from the stem set that disproportionately drove the false positive rate. Some stems, such as rally and debate, were accurate in detecting political content only if in combination with at least one other stem, hence such stems were down-weighted to require the presence of another stem in the same screenshot. The final set of 168 stems used are shown below in Table S1.

We additionally enforced that applications that fell into the category of game were considered as never containing political screenshots. Many mobile games use war terminology on-screen (e.g., discussion of siege campaigns and fantasy scenarios far removed from political reality). Via manual inspection, we determined that no screenshot in the training data displayed a mobile game that included genuine political content. This classification method yielded $F = 0.74$, representing a large improvement over Approach (1).
Table S1. Word stems incorporated used in automated classification of screenshots as containing political content or not.

<table>
<thead>
<tr>
<th>alex.jones</th>
<th>chuck.schumer</th>
<th>fiorina</th>
<th>kamala</th>
<th>o.rourke</th>
<th>quid.quo.pro</th>
<th>syria</th>
</tr>
</thead>
<tbody>
<tr>
<td>americans</td>
<td>climate.chang</td>
<td>fox.and.friends</td>
<td>kellyanne</td>
<td>obama</td>
<td>rachel.madow</td>
<td>Taliban</td>
</tr>
<tr>
<td>amendment</td>
<td>clinton</td>
<td>G7</td>
<td>keystone</td>
<td>obamarac</td>
<td>racis</td>
<td>tax</td>
</tr>
<tr>
<td>andrew.yang</td>
<td>congress</td>
<td>gabbard</td>
<td>klobuchar</td>
<td>ocasio.cortez</td>
<td>rally</td>
<td>terroris</td>
</tr>
<tr>
<td>asylum</td>
<td>conservativ</td>
<td>gavin.newsom</td>
<td>law.maker</td>
<td>ouse.majority</td>
<td>rand.paul</td>
<td>tillerson</td>
</tr>
<tr>
<td>ballot</td>
<td>constitution</td>
<td>george.zimmerman</td>
<td>lawmaker</td>
<td>partisan</td>
<td>recession</td>
<td>transphobic</td>
</tr>
<tr>
<td>barack</td>
<td>covfefe</td>
<td>gerrymand</td>
<td>legaliz</td>
<td>pelos</td>
<td>reform.</td>
<td>trayvon.martin</td>
</tr>
<tr>
<td>battleground.stat</td>
<td>debate</td>
<td>ginsburg</td>
<td>legisla</td>
<td>pence</td>
<td>representative</td>
<td>treason</td>
</tr>
<tr>
<td>beto</td>
<td>deep.state</td>
<td>giuliani</td>
<td>lewandowski</td>
<td>polariz</td>
<td>republican</td>
<td>trevor.nouh</td>
</tr>
<tr>
<td>biden</td>
<td>deepstate</td>
<td>govern</td>
<td>liberal</td>
<td>politic</td>
<td>RNC</td>
<td>trump</td>
</tr>
<tr>
<td>bigly</td>
<td>democrat</td>
<td>gun.control</td>
<td>limbaugh</td>
<td>pompeo</td>
<td>roger.stone</td>
<td>tucker.carlson</td>
</tr>
<tr>
<td>bigot</td>
<td>deplorabl</td>
<td>hassan.minhaj</td>
<td>lindsey.graham</td>
<td>populis</td>
<td>sanctions</td>
<td>united.nation</td>
</tr>
<tr>
<td>black.lives.matter</td>
<td>deport</td>
<td>hate.crime</td>
<td>lobbyis</td>
<td>president</td>
<td>sanders</td>
<td>warren</td>
</tr>
<tr>
<td>blm</td>
<td>DHS</td>
<td>hate.speech</td>
<td>locker.room.talk</td>
<td>primary.electio</td>
<td>schiff</td>
<td>white.house</td>
</tr>
<tr>
<td>booker</td>
<td>DNC</td>
<td>house.minority</td>
<td>maga.</td>
<td>pro.choice</td>
<td>senator</td>
<td>white.nationali</td>
</tr>
<tr>
<td>border.wall</td>
<td>dogwhistl</td>
<td>human.right</td>
<td>manafort</td>
<td>pro.life</td>
<td>shooting</td>
<td>white.supreme</td>
</tr>
<tr>
<td>breitbart</td>
<td>donald.trump</td>
<td>immigration</td>
<td>mandate</td>
<td>prochoice</td>
<td>shutdown</td>
<td>.CIA.</td>
</tr>
<tr>
<td>brexit</td>
<td>economy</td>
<td>impeach</td>
<td>medicaid</td>
<td>prolife</td>
<td>socialis</td>
<td>.DOJ.</td>
</tr>
<tr>
<td>buttigieg</td>
<td>elected.officia</td>
<td>impeachment</td>
<td>melania</td>
<td>prosecut</td>
<td>sondland</td>
<td>.election</td>
</tr>
<tr>
<td>campaign</td>
<td>exit.poll</td>
<td>incel</td>
<td>merkel</td>
<td>protest.</td>
<td>stephen.colber</td>
<td>.facis</td>
</tr>
<tr>
<td>capitol</td>
<td>extradit</td>
<td>incumbent</td>
<td>migrant</td>
<td>protests</td>
<td>supreme.court</td>
<td>.GOP.</td>
</tr>
<tr>
<td>castro</td>
<td>FBI</td>
<td>jeff.sessions</td>
<td>minorities</td>
<td>proud.boys</td>
<td>susan.collins</td>
<td></td>
</tr>
</tbody>
</table>

Note. A list of 168 political word stems used in classification approaches (2) and (3). In approach (2), the presence of these word stems in screenshots identified screenshots as containing political content. In approach (3), the presence or absence of each stem from a screenshot’s text was used as an individual feature with which to predict whether or not a screenshot contains political content, based on a random forest model trained on the ground truth set. This set was created based on political news events occurring in late 2019, and manually updated based on classification performance in a basic bag of words approach within the original ground truth set. In the table, the presence of a period (.) indicates a space between characters.
The manual stem removal and down-weighting used in Approach (2) risked over-fitting to the training sample. Thus, the bag of words approaches (1) and (2) are used only as a benchmark against Approach (3), a random forest based approach ultimately chosen. To that end, the potentially overfitted bag of words approach provides a conservative estimate of the added value of using the random forest approach. Manual omission or down-weighting of stems was not used in the random forest model used for classification.

The random forest model used in Approach (3) was fully non-parametric its selection of stems to use as features, and in its assignment of the value of those stems in classification, thus mitigating overfit concerns when paired with proper training. Within the model, we treated the presence or absence of each of the 168 stems as a single binary feature upon which to predict political content in screenshots. The random forest model was tuned across various values of $m$, or number of features (stems) out of 168, ranging up to 29 per tree. Six decision trees were applied per-forest, and intra-model forest selection was trained under three-fold cross-validation and three repeats, optimized on the $F$ metric. To determine performance metrics and optimal $m$ of the model, we used five-fold cross validation of the entire model within the ground-truth training set. After determining the $F$-optimizing $m$ value of 25, we then trained the model on the entire ground truth set. As the output of the random forest model is probability of political content within screenshots ($p \in [0,1]$), we then determined the optimal threshold for binarized output based on the $F_1$ score, resulting in an optimal cutoff value of $p \geq 0.085$ for positive cases as shown in Figure S1. Random forest based on word stems was the optimal classification method based on our analyses.
Figure S1. Performance of our binary classification model across strictness of output binarization probability thresholds. Note. Four accuracy metrics describing the performance of my chosen keyword-based random forest model in the training data across variation in classification probability threshold. The chosen probability threshold, 0.85, was selected as it is the mean value of the F1-optimizing probability range.

As shown, the optimal threshold value was the lowest non-zero probability value estimated by the model. As the stepwise probability output provided by the model resulted in the lowest step ranging $p \in (0, 0.17]$, we used the mean value of this range as a threshold probability for the entire dataset.

The classification confusion matrix for Approach (3) is shown below in Table S2. As the model was developed using five-fold cross-fold validation rather than a using a single OOB holdout set, this confusion matrix approximates the application of the final classifier on the entire training set.

**Table S2. Confusion matrix of selected classifier.**

<table>
<thead>
<tr>
<th></th>
<th>Predicted Negative Case</th>
<th>Predicted Positive Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative Case</td>
<td>TN: 95.4%</td>
<td>FP: 0.6%</td>
</tr>
<tr>
<td>Positive Case</td>
<td>FN: 1.4%</td>
<td>TP: 3.6%</td>
</tr>
</tbody>
</table>

While we do not detail Approach (4) which made use of text embedding in combination with a random forest design, it is included in Table S3 here for reference.

**Table S3. Overall performance of four classifications models.**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>First-pass bag of words</th>
<th>Improved bag of words</th>
<th>R.F. classification w/ political words</th>
<th>R.F. classification w/ pretrained embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameters</strong></td>
<td>Single stem occurrence</td>
<td>Targeted stem restriction</td>
<td>$m = 25$ stems, threshold = 0.085</td>
<td>$m = 19$ P.C.s, threshold = 0.4175</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.97</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.39</td>
<td>0.75</td>
<td>0.86</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.812</td>
<td>0.726</td>
<td>0.72</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>0.53</td>
<td>0.74</td>
<td>0.78</td>
<td>0.50</td>
</tr>
</tbody>
</table>
Note. Accuracy measures of four methods for identifying political content in screenshots: basic bag of words using intuitive stems, manually improved bag of words, random forest supervised machine learning (R.F.) using random selections of the improved stem set, and random forest supervised machine learning (R.F.) using pre-trained text embedding with principal component analysis, using the GloVe 400k vocab x 300 dimension dictionary trained on Wikipedia and Gigaword (Pennington et al., 2014). Both R.F classification methods utilized five-fold cross-validation for performance metric calculation, and 3-fold cross validation for tuning the number of randomly selected variables, m. Resultant probability values were binarized according to the $F_1$-minimizing threshold. F is the most useful accuracy measure for heavily unbalanced sets.
Additional Details on Manual Tagging

Our primary method for selecting screenshot files for hand-coding relied on stratification by participant, by day, and by time of day in 6 hour chunks (e.g., 12am to 6am was a possible chunk, 6 to noon, etc.), in order to (a) capture variation across the underlying rhythms of life that may impact smartphone usage and (b) to keep chronologically-continuous content in its temporal context to improve accuracy of human coding. Human-coders were assigned to work through as much content as possible in the allotted hours they had available to work. As the content for human-coding was kept securely in a folder and file structure that mirrored the manner in which sampling was conducted, and folders were ordered according to participants’ randomly assigned numeric identifiers. Human coders working through the folder system did not arrive at folders comprised of screenshot content belonging to participants with higher-value randomly-assigned IDs.

The scale of our training set is a reflection of the low incidence rate of political content versus non-political content in our corpus and in reality, i.e., imbalance in our target binary classification. To achieve a training set of this scale, tagging was conducted in two large efforts across 2020 and 2021, with over one hundred hours of total involvement from multiple undergraduate research assistants and one author who oversaw this effort. Temporal contextualization came with the added benefit of expediting content evaluation greatly --- e.g., a continuous session of mobile gaming or driving/maps usage could be identified fairly easily, and would quickly yield large batches of negative-case screenshot content classifications. Complete-graph pairwise IRR calculations between all coders was not available for calculation due to disruptions owed primarily to the COVID-19 pandemic, which resulted in loss of access to our dedicated secured human-coding space, and relatedly, coders not completing their review of all content assigned to them individually. To overcome this, in the paper, we report aggregate reliability by simultaneously evaluating the classifications of each screenshot file coded by more than coder, resulting in a reliability measure of 0.59 as reported. Given the extreme imbalance in our data, which downwardly biases the output value of a reliability measure, we are confident in this value in combination with ample human-auditing and review of hand-coded screenshots’ content conducted thereafter by the overseeing author.
As for classification instructions, coders were provided the instructions referenced in the manuscript, formatted as follows:

We define ‘political topic’ as anything directly related to
- the presidential administration,
- elections and election campaigns (at any level from national to local)
- policy debates and/or decisions (economic policy, foreign policy, social policy, other policy)
- political satire and political satirists
- politicians, discussion of politicians, election discussion of elected and appointed government officials
- discussion of social or partisan groups, including political parties, partisan think tanks
- discussion of political ideology and/or partisanship e.g., liberal, conservative, communist, socialist, progressive
- discussion of hate speech, hate speech
- racial politics, identity politics
- terrorism, current American wars, US foreign policy, foreign affairs
- political analysis/reactions to political current events in the U.S. or abroad.

News organizations like ‘Wall Street Journal’ or ‘Fox’ do not qualify, unless there is discussion of politics.

…with coders selected and hired based on their performance in an initial test. The majority of coders were recruited from our political science department, and engaged in hours of joint discussion with the overseeing author regarding project goals, as well as back-and-forth determination and precedent-setting using what they found to be ambiguous cases early-on in the effort.
Results Related to Group-Level Summarization

To statistically gauge the utility of group-level summarization in this sample, we use the coefficient of variation (or CoV) (Bedeian & Mossholder, 2000). The CoV of a random variable is equal to its standard deviation divided by its mean, thus providing a standardized measure of how incorrect a group mean is for estimating the value of a single point within the group (i.e., a standardized measure of the magnitude of the ecological fallacy if enacted on a given sample). Across the 115 participants, the average daily number of political content encounters (Panel B in main text Figure 3) is 18, and the CoV is 1.7. This implies that if the sample mean of 18 were used to estimate the exposure of any single participant, the estimate would be off by a factor of 1.7. Turning to Panel C in Figure 3: the CoV of political screentime percentage is 1.4; this suggests that, in expectation, a sample-level estimate of time dedicated to political content is off by a factor of 1.4 from any one participant’s lived experience.

Next, we discuss intra-individual variation across units of time. Our prior evidence suggests participants’ behavior is not static nor uniform over time and thus, not well captured by any one value estimate. This is increasingly apparent at smaller and smaller timescales. To summarize this point statistically, we calculated the within-subject CoV of two core measures of political content exposure: the frequency of political content encounters and the percent of screentime spent on political content. The results of this analysis are shown in Figure S2. In both panels, the y-axis shows the CoV of the measure variables, and the x-axis is divided into four temporal levels of aggregation. At the origin of the figure, we estimated the CoV of variable estimates using each participant’s entire data collection period; by construction, this is a single point estimate for each participant with value 0.

At the remaining three temporal levels, we estimated each participant’s intraindividual CoV at the corresponding temporal level by bucketing the participant’s data timeline into a vector of discrete temporal categories. Five participants are highlighted in colors to provide continuity from previous figures; all 115 participants are shown in the background in gray, and sample averages are calculated at each point in bold black.

By calculating across time, the y-axis values indicate how inaccurate a participant's average behavior would be in predicting the behavior of the participant at a given point in time. For example, if a researcher calculated participant 46’s hourly number of encounters with political content (Panel A), that estimate would be off by a factor of 3.72 relative any single hour in the sample, in expectation. Across all participants in this sample, the average CoV of political content encounter frequency is 1 at the daily level, 1.8 at the level of 6-hours, and greater than three at the level of hours. For political screentime percentage, these numbers are even higher.
and, similarly, increasingly with increased temporal granularity. For some participants, the *CoV* for either value ranges as high as 10, meaning that an hourly estimate would be expected to be off by an order of magnitude in expectation.

Figure S2 shows that intraindividual variation is consistent with interindividual variation. A cross-sectional average of daily political content encounters is incorrect by a factor of 1.7 when applied to individuals in expectation (from preceding subsection). At the level of six-hour blocks, interindividual variation and intraindividual variation are nearly equal, and at finer granularity, intraindividual variation is greater than inter-individual variation. More broadly, Figure S1 helps validate the role of timescale explication in measurement. A single core variable, measured in a single participant, displays monotonically increasing variability at successive levels of aggregation over time.

**Figure S2. Political content exposure measurement variability across temporal aggregation levels.** *Note.* Coefficient of Variation (CoV) of daily frequency of political content encounters (Panel A) and the percentage of screen time during which political content was on-screen (Panel B). The y-axis shows CoV values ranging from 0 to 5, and the x-axis shows temporal levels of aggregation ranging from a participant’s entire study duration to hour-long blocks at the most granular. All 115 participants are shown in gray, with five participants highlighted for reference, and sample-wide averages shown in bold black squares. On average across the sample, estimates of hourly politics encounter frequency misrepresent a randomly selected hour by a factor of 3.1, in expectation. On average across the sample, estimates of political screentime percentage misrepresent a randomly selected hour by a factor of 4, in expectation.
Regarding RQ4 in the main text: To investigate within-person variation, we built on the preceding findings that quantified the structure of within-person variation in political content segment durations. First, we applied the CoV measurement strategy to compare within-person and between-person political segment duration measurements. Second, we examined the relationship between segment durations and encounter frequency, plus segment durations and overall exposure, and the ergodicity of these bivariate relationships. Third, we examined whether within-person political content segment durations were related to overall session durations (i.e., Does temporal screen activity correspond to temporal segment patterns?).

To compare within-person and between-person variation in segment durations, we first calculated the CoV of three between-person measures of subject-level segment duration tendency: mean segment duration, median segment duration, and maximum segment duration, as shown in Panels A, B, and C of Figure 5 respectively. The CoV of mean segment duration was 0.52, the CoV of median segment duration was 0.66, and the CoV of maximum segment duration was 1.37. Therefore, the average participant-level mean segment duration was off by a factor of 0.52 in describing any individual participant’s mean segment duration, with analogous interpretations for the median and maximum CoV.

We calculated the CoV for each participant’s within-person distribution of segment durations and took the average of this value across the sample. The resulting average within-person CoV was 1.2, more than double the between-person CoV for the most comparable metric, mean segment duration. In essence, a typical participant’s average experience of political content segment duration was more consistent with that of other individuals than with their own within-person experience. This finding demonstrates that summary values of individuals may provide a deceptively flawed view of political audiences.

**App Categorization Schema**

The smartphone application categorization schema can be accessed through this link.


