Agenda Setting and Market Power in Online News*

Gregory J. Martin† Andrey Simonov‡ Shoshana Vasserman§

October 2026

Abstract

How much power do news outlets have to "set the agenda" — to determine what events their audience hears about and which topics it thinks are important? The answer is essential to understanding the political impacts of changes in media ownership and editorial direction. Using a large panel of online news readership and extensive data on article-level news content for over 3000 online news outlets, we decompose demand for news into two components: stable interest in certain topics, and trust in favored outlets to determine which topics are important each day. Our estimates reveal substantial heterogeneity across outlets and across readers. Most outlets have effectively zero influence over their audience's information diet, but a small number of influential outlets exercise substantial agenda setting power. We use our model estimates to conduct a variety of counterfactual exercises that illuminate the consequences of changes to the supply side of the news market on the compositions of outlets and news topics read by consumers.

JEL codes: L82, D12, D83

Keywords: Media Economics, Online News, Agenda Setting.

*We are grateful for the exemplary research assitance provided by Federico Llarena and Daniil Mikhailov, as well as for the computational support provided by the Data, Analytics, and Research Computing (DARC) group at the Stanford GSB (RRID:SCR 022938). We are also grateful for the many helpful comments and feedback provided by our colleagues and the seminar audiences at Columbia, Harvard, USC, and Sciences Po.

†Stanford GSB. Email: gjmartin@stanford.edu.

‡Columbia GSB. Email: as5443@gsb.columbia.edu

§Stanford GSB and the NBER. Email: svass@stanford.edu.

1 Introduction

When two conservative media entrepreneurs purchased the *Baltimore Sun* in early 2024, one of the new owners told the public what the paper's readers would soon be getting more of: "You're [going to be] seeing more about crime. You're [going to be] seeing more about the mayor and City Hall."[1]

The *Sun* is far from the only prominent American media outlet to undergo a dramatic shakeup in recent years. In January 2026 the *Washington Post* shed roughly a third of its newsroom, eliminating most of its international desk and refocusing its coverage on national politics. CBS News, after the acquisition of its parent company by Skydance Media in August 2025, installed a new editor in chief, cut staff and reoriented its programming.[2] At the local level, more than 200 local TV stations were acquired by three large conglomerate owners in the decade leading up to 2020, with measurable consequences for those stations' news coverage (Martin et al., 2026a).

In each case, changes in ownership have reshaped the mix of coverage that the outlet places before its readers. But the consequences of these changes are not clear. Outlets might command trust among their audiences, allowing them to steer their readers' attention toward the topics they choose to feature. An increase in the coverage of crime or city government would then lead the outlet's readers to learn more about these topics, and less about others. But US media outlets operate in a highly competitive environment with a large number of differentiated outlets to which readers have easy digital access. If readers dislike the mix of reporting they get from a given outlet or wish to learn about specific topics that don't receive adequate coverage, they can easily switch to another. The consequences of a change in ownership therefore depend not only on the loyalty of the outlet's audience but also the alignment between the topical preferences of the outlet's audience and its new owners.

In this paper, we ask how these competing forces shape the news diet of US news readers. How much of what people read online reflects readers' stable preferences for specific topics, and how much reflects the agenda-setting power (McCombs and Shaw, 1972) of the outlets they trust to decide for them what is worth reading? To answer this question empirically, we assemble a dataset that links the news content that readers consume to the coverage of topics available across all English-language news outlets on the internet.

On the demand side, we draw on a year of desktop browsing data from the Comscore panel, for the roughly 85K panelists who read any news online in 2022. For each reader, we observe the full sequence of web pages that they visited—news and non-news alike—and the time that they spent on each page. On the supply side, we draw on hourly records of article placement on the home pages of more than 3,000 news outlets from the GDELT project. We assign news articles to topics using a structural topic model (Roberts et al., 2013) estimated on a large corpus of article text and bylines provided by the vendor Peakmetrics. Linking these sources together allows us to measure, at the level of the outlet, topic, and day, both how prominently each outlet features each topic and how much time each reader spends reading each outlet-topic combination.

[1] Armstrong Williams, co-owner, in an interview with NPR following the sale of the *Baltimore Sun*. Transcript: https://www.npr.org/transcripts/1233856063.

[2] See https://www.npr.org/2026/02/04/nx-s1-5699328/washington-post-layoffs-jobs-bezos and https://www.npr.org/2026/01/27/nx-s1-5689849/after-rocky-start-bari-weiss-to-cut-staff-add-commentators-at-cbs-news.

We begin our study with several stylized facts. First, although all outlets respond to major newsworthy events, many outlets are *specialized* in specific topics. On days with newsworthy events on a given topic, outlets that specialize in that topic cover the event disproportionately more than do non-specialists. Second, although much of outlets' specialization is persistent (e.g., ESPN always covers sports topics extensively) outlets' coverage is driven, in part, by the experience of the journalists that they staff. Hiring a journalist who had written heavily on a given topic at their previous job increases the coverage of that topic at the new outlet. Together, these facts suggest that the relative coverage of topics across outlets varies over time as the mix of newsworthy events changes and newsroom staffing evolves.

Next, we examine how the occurrence of newsworthy events in different topics, and the degree to which different outlets cover, them shape readers' news diets. We find ample evidence that readers are responsive to the stories that their preferred outlets choose to feature more prominently. An article that moves from the bottom to the top of an outlet's front page sees a roughly 50% increase in its hourly readership. However, we also find evidence that readers seek out coverage of specific topics when they are especially relevant. On days with high overall coverage of a given topic—reflecting the occurrence of a notable newsworthy event in that topic—readers increase their consumption of outlets specialized in that topic on both the extensive and intensive margins. These facts suggest that both individual interests in specific topics and the coverage decisions of outlets play a role in shaping the mix of news topics that readers consume.

Motivated by these facts, we specify a model of time-allocation for news consumption, adapting the frameworks of Crawford and Yurukoglu (2012) and Martin and Yurukoglu (2017). Each day, readers divide a fixed budget of time into reading online news in each outlet–topic combination, non-news online activity, and time spent offline. Readers have baseline preferences over outlets that drive their news consumption on a typical day—for instance, a reader with a strong preference for the New York Times will read a typical mix of topics covered by the New York Times front page on average. But readers also respond to day-to-day changes in the mixture of topics available to read on different outlets. These responses are governed by individual preference parameters which can vary according to both the topic being covered (*intrinsic topic preferences*) and the identity of the outlet covering it (*trust in favored outlets to set the agenda*).

Based on the relative magnitudes of these two forces, our model predicts the extent to which a reader substitutes across outlets as the mix of newsworthy topics and the extent to which they are covered by different outlets changes over the year. A reader with strong intrinsic preference for some topic will shift attention towards outlets that heavily cover that topic when it is newsworthy, even if the reader does not typically read those outlets. A reader with strong trust in some outlet will follow that outlet's coverage mix, even when it deviates from what other outlets are covering or what the reader typically reads in other outlets.

To take this model to the data, we estimate the parameters governing each reader's topic- and outlet-level preferences using the control-function method introduced by Blundell and Powell (2007). This method can accommodate two important features of our data. First, attention to news is heavily censored: most readers don't read any news most of the time. Our estimation method is semi-

parametric, and handles censoring without imposing a distributional assumption on unobserved tastes. Second, outlets' topic coverage is endogenous: outlets may cover events that are of particular interest to their audiences. Our method handles endogeneity by specifying instruments and imposing a distributional exclusion restriction.

To identify the preference parameters in our model, we rely on variation in coverage from the quasi-random timing of newsworthy events, interacted with each outlet's degree of specialization across topics. When a newsworthy event in a given topic occurs, outlets that specialize in that topic disproportionately increase their coverage of it. The extent to which readers shift to or away from these outlets in response informs the model about their relative preferences. We estimate preferences separately for each of 1,180 clusters of readers across 59 geographic regions, allowing the distribution of preferences to differ flexibly across places and across readers with different combinations of outlets in their consideration sets.

Our central finding is that intrinsic preference and agenda setting drive consumption in strikingly different proportions across topics and outlets. To gauge the importance of each of these components across topics, we simulate a large news shock[3] to each topic, and ask how much reading of that topic increases in response. We conduct this exercise under three conditions: under our parameter estimates, under a counterfactual in which intrinsic topic preferences are set to zero, and under a counterfactual in which agenda setting is set to zero.

The topic we label "Local Events" has the largest response of any topic in the baseline: reading time increases by over 40% in response to the simulated news shock. This response is almost entirely due to intrinsic interest: it falls below 5% when topic preferences are removed, but only dips to about 35% when agenda setting is removed. The "US Politics and Elections" and "American Sports" topics behave similarly. The pattern reverses for fast-moving international stories: for the Russia–Ukraine war and for COVID-19, the predicted reader response to a news shock is smaller, and agenda setting accounts for 70 to 80% of it. These patterns suggest that readers seek out stories about local news, sports, and elections largely on their own, but they read about the war in Ukraine and the pandemic largely because the outlets they trust place those stories before them.

To gauge each outlet's ability to steer its readers to certain topics, we consider a different counterfactual in which coverage decisions are fixed at their observed levels in 2022, but individual outlets' agenda setting power is shut down, one outlet at a time. We find that the vast majority of outlets have essentially no meaningful influence on what readers read. For all but a small number of prominent national outlets—among them the New York Times, Fox News, and CNN—shutting down agenda setting does little to shift the topics that their audiences read about. Although there is some heterogeneity across places, national outlets generally rank highest in influence both at the national level and within local media markets.[4] While multi-homing by readers is the norm, and a long tail of outlets attract sizable readership, agenda-setting influence is concentrated in a handful of national brands.

Nationalization of agenda setting power, however, does not imply that local preferences for news topics have homogenized. While some topics like crime and sports draw intrinsic interest across the

[3] We simulate an event that drives average coverage levels two standard deviations above the mean for each topic.

[4] Minneapolis, for instance, has three local news outlets among its top 5 most influential outlets, while Atlanta has one and Boston has none.

board, other topics, like energy and the environment, exhibit substantial heterogeneity across space. The median preference parameter for stories about the environment in San Francisco, for instance, is three times that of the median reader in Chicago. Furthermore, readers value the bundle of coverage that local outlets provide. Our model predicts that local newspapers from Boston to Tampa would all lose readers if their overall coverage were reshaped to resemble that of a generic national outlet.

Motivated by the editorial changes outlined by the new owners of the Baltimore Sun, we consider a counterfactual in which the Baltimore Sun increases its 2022 coverage of crime by 25% and decreases its coverage of all other topics proportionally. We find that baseline Baltimore Sun readers would *increase* their time spent reading the Sun on net as a result, driven by a small number of readers who increase their time on the Sun substantially, while the plurality of readers decreases their time on the Sun very slightly. Moreover, all but one of ten cities we consider—San Francisco—would see a net increase in time spent on their local paper as a result of a similar increase in crime coverage.

We also consider case studies inspired by the recent shakeups at the Washington Post and CBS News. Using the relationship between staffing experience and topic coverage estimated in our descriptive section, we simulate a counterfactual in which the Washington Post fired all journalists whose bylines were not primarily on topics related to politics in 2022. Our simulation projects that the Washington Post's coverage of topics like the Russia-Ukraine conflict, US elections, and social issues like abortion would increase substantially as a result. On net, readers of the Washington Post would respond negatively to this change: net readership would decrease by 3.5%, shifting mostly to the outside option, rather than other online outlets. This counterfactual focuses only on short-run changes to the mix of coverage and does not allow for changes in the outlet's brand—whose influence on reading cannot be unbundled from an outlet's typical topic mix in our model. To consider the influence of a shift in brand perception—and capture CBS's perceived rightward shift under its new ownership—we simulate a counterfactual in which the outlet preference that each reader in our panel has for CBS is replaced by the outlet preference that the same reader has for Fox News. Our model suggests that this change would result in polarizing effects across readers: the top and bottom deciles of responses are similar in magnitude and opposite in sign. However, net readership of CBS would increase by 6% overall, drawn in part from competing outlets like Fox News itself.

In the remainder of the paper, we describe how our analysis connects to existing work, describe the datasets and their connection to the research question, and discuss how we use that data to estimate a model of news time allocation and apply it to make predictions about changes to the news market.

2 Related Literature

Our work connects several literatures. First, we contribute to the literature on the demand for news. A long-standing question is whether tastes for news are innate or cultivated (Lippmann, 1925; Dewey, 1927), and what readers value in it (Hamilton, 2004): hard versus soft news (Baum, 2003; Martin et al., 2026b), entertaining content (Durante et al., 2019), like-minded political coverage (Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2010), or any news at all when alternatives are at hand (Chen and Yang, 2019; Levy, 2021; Aridor et al., 2026). We take this question to a large panel of

individual online browsing, building on empirical work on the consumption and production of online news (Flaxman et al., 2016; Sen and Yildirim, 2015; Calzada and Gil, 2020; Athey et al., 2021; Cagé et al., 2020; Durante and Knight, 2012; Knight and Tribin, 2019) and estimate a model of discrete-continuous choice similar to Crawford and Yurukoglu (2012) and Martin and Yurukoglu (2017), in the spirit of other empirical models of news markets (Strömberg, 2004; Gentzkow and Shapiro, 2015; Bang and Sieg, 2026). Our contribution is to estimate readers' preferences for news topics separately from the influence of outlets—using individual consumption data that span the full range of topics and thousands of outlets—whereas existing work has typically isolated a single content dimension such as slant, or studied a single outlet or market.

The most closely related study to ours is Simonov and Rao (2022), who ask whether Russian readers consume state media because they genuinely demand it or because the government controls media supply, and find that demand reflects a taste for a bundle of apolitical content together with government propaganda. We ask a different question: to what degree do outlets throughout the United States news market drive the mix of topics that their readers consume, and how does this agenda-setting power vary across outlets and media markets?

Second, our findings inform a literature on competition and product differentiation in media markets. Competition shapes the ideological diversity of coverage (Gentzkow et al., 2014); it can reduce the quantity and quality of news each outlet provides (Cagé, 2020); and the composition of a market's readers, together with its size, shapes the content and quality that outlets supply (George and Waldfogel, 2003; Berry and Waldfogel, 2010). Whether competition leads outlets to differentiate or to converge depends on whether readers substitute across outlets on the basis of content or of outlet loyalty—a demand-side primitive that is hard to recover without disaggregated consumption data. Our contribution is to estimate that parameter market by market and to use the resulting model to predict how repositioning, entry and exit, and consolidation would shift readership and reallocate attention across outlets.

Third, we speak to a large literature on the influence of media on politics, which documents the media's power to shape what issues the public attends to (McCombs and Shaw, 1972; King et al., 2017; Garz and Martin, 2021), to shape electoral participation and accountability (Snyder and Strömberg, 2010; Gentzkow et al., 2011), and to move votes through persuasion (DellaVigna and Gentzkow, 2010; DellaVigna and Kaplan, 2007; Enikolopov et al., 2011; DellaVigna et al., 2014). Influence over what the audience knows and believes to be true about the world is the core ingredient of media power (Prat, 2018). Where most of this work documents that influence exists or estimates its effect on attitudes and votes, our contribution is to estimate agenda-setting power structurally and directly from consumption behavior—how far each outlet moves the topical composition of its readers' attention—which, to our knowledge, has not been done before. We then use counterfactual simulations to rank outlets by this influence and to trace where readers' attention flows when an outlet changes course or disappears.

Fourth, we bring new evidence to debates over the decline of local news and the policies meant to slow it. National competitors—broadcast television (Angelucci et al., 2020) and the national press (George and Waldfogel, 2006)—and, more recently, online platforms (Djourelova et al., 2025) have eroded local outlets' audiences and the advertising revenue that funds original reporting (Angelucci

and Cagé, 2019), with measurable consequences for local political life (Hayes and Lawless, 2017, 2021). A central question is whether this decline reflects weak demand for local content, or local outlets' inability to compete with national rivals on quality and brand—two diagnoses with opposite policy implications. Our contribution is to distinguish these two explanations: our counterfactuals quantify the latent demand for local content and trace how repositioning a local outlet's coverage would change its readership.

3 Data

Our analysis rests on two key datasets: a consumer-level panel of online news reading and an outlet-level panel of front page article placement. To interpret the content of the articles in these panels, we estimate a structural topic model on a large subsample of article text. To measure the role of staffing in topic specialization, we construct a panel of journalist employment affiliations. In this section we detail each of these datasets and how they are used in our analysis.

3.1 Browsing Demand Data

Our primary demand-side data come from Comscore, a market-research firm that tracks internet browsing through a professionally recruited panel of US households. We use one year of desktop browsing records covering January 1 through December 31, 2022. The records log every URL visit with a time stamp. Our analysis sample consists of approximately 85,000 panelists with at least five minutes of total news consumption over the year. For each panelist, we observe demographic characteristics including gender, age, presence of children, household income, household size, race and ethnicity, and state of residence. See Appendix B.1 for details on the sample-construction process.

We categorize URLs into news and non-news based on a set of pre-specified news domains, and aggregate all non-news URLs into a single "other" category. Because news articles are typically denoted by a unique URL "slug", we process URLs on news outlets into articles (based on slugs) and home pages. We link each article URL to a topic using the classifier introduced in Section 3.4 and aggregate reading across articles in the same topic-outlet pair.[5] For time spent browsing an outlet's home page—which cannot be attributed to a single article—we allocate the browsing duration across topics in proportion to the outlet's aggregate production shares (see Appendix A.4) on that day.

3.2 News Supply Data

Our supply-side data combine two sources: GDELT, which tracks the position of articles on outlet home pages, and Peakmetrics, which archives article text. Together these data allow us to measure what each outlet published on each day of 2022 and how prominently each article was displayed.

GDELT The open-source GDELT project logs, at hourly frequency, every article URL that appears on the home page of over 40,000 US and global news outlets. We collect the full 2022 panel. Two

[5] See Appendix A.3 for the URL-to-topic matching procedure.

features of these data matter for our analysis. First, coverage is uncensored: we see every URL an outlet placed on its home page. Second, the hourly frequency lets us exploit within-day variation in front-page placement as a local supply-side shock to news consumption. GDELT has two limitations: it does not cover articles that never reached the home page, and it does not include article text.

Peakmetrics Peakmetrics complements GDELT by providing article text. The service archives URLs, titles, time stamps, and full text for articles published by roughly 20,000 US and European news outlets. Coverage is broad but not exhaustive—individual articles can be missing for idiosyncratic reasons. As such, we do not treat Peakmetrics as a census of published content. Rather, we use article text from Peakmetrics to train the topic-prediction model introduced in Section 3.4, and then apply that model to URLs observed across GDELT and the browsing records.

3.3 Staffing by topic

We construct a monthly panel of outlet-topic staffing based on the bylines on articles in our PeakMetrics dataset and employment histories from LinkedIn, as provided by Revelio Labs. For each article author denoted in the bylines of a Peakmetrics article in our dataset, we attribute a monthly topic specialization based on the share of articles whose bylines they appeared in during the preceding month. We then merge each author to their LinkedIn employment history to identify which outlets they were employed at in each month.[6] Finally, we construct outlet-topic monthly counts of topic specialization by summing up the number of authors employed at the outlet (based on LinkedIn) who are specialized in the topic (based on Peakmetrics). See Appendix A.7 for details.

3.4 Topic classification

We classify news articles into topics using a Structural Topic Model (STM) (Roberts et al., 2013) trained on the Peakmetrics article text. The STM lets topic prevalence vary by outlet and by day, capturing the fact that outlets emphasize different topics and that topic salience shifts with current events. Training details and model-selection diagnostics for the STM are in Appendix A.2.

We choose a 35-topic model, of which six topics capture non-substantive content (advertisements, navigation text, and other website artifacts) and are merged into a single "junk" category. The remaining 29 substantive topics span the range of news coverage:

- **Hard news:** Elections and U.S. National Politics; Crime and Law Enforcement; Legal Proceedings and Criminal Justice; Ukraine–Russia Conflict; COVID-19 and Public Health
- **Business and finance:** Financial Markets and Economic Trends; Corporate Finance; Business Development and Technology; Market Analysis
- **Sports:** American Sports (e.g., football, basketball); European Sports (e.g., soccer); Canadian Ice Hockey
- **Lifestyle and entertainment:** Entertainment and Celebrity Culture; Lifestyle and Food; Family Life; Nature and Travel

[6] It is possible for authors to be employed by more than one outlet in the same month.

- **Local and regional:** Local Community Events; Education and Local Politics; Urban Development; Weather and Forecasts
- **Other:** Medical News; Energy and Environment; Abortion and Social Issues; Indian Politics; Miscellaneous

Figure 1 plots the average prevalence of each topic across outlets, measured as the share of front-page coverage devoted to each topic. No single topic dominates: the most frequent topic, American Sports, accounts for only 6% of front-page coverage on average, and the top five topics—American Sports, Crime & Law, Entertainment, Business Development & Technology, and Local Events—together account for roughly 23%. The remaining 24 topics are spread fairly evenly, each contributing between 0.6% and 3.4%.

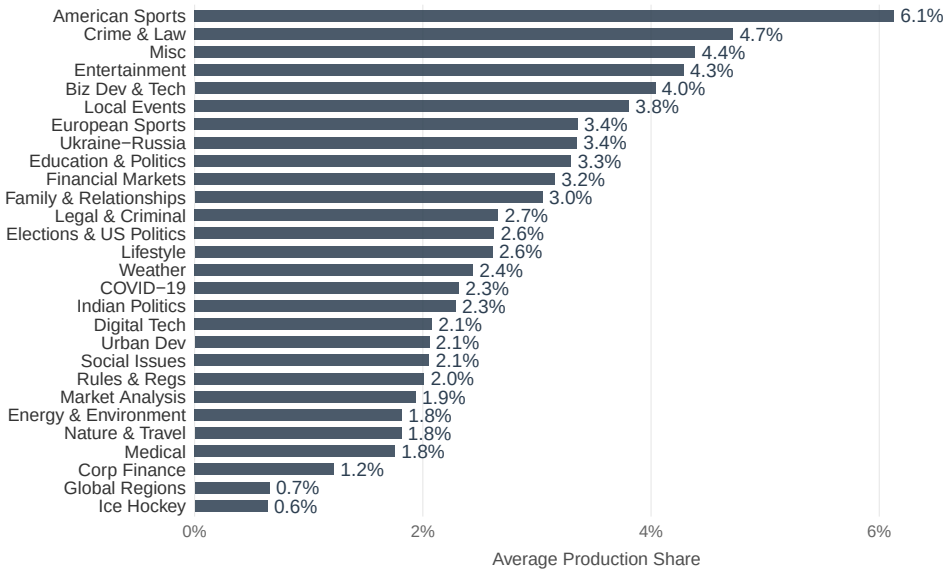


Figure 1: Average production share by topic across all outlets in 2022.

Notes: Each bar shows the average share of front-page coverage devoted to a topic, computed across all 3,244 outlets in our GDELT panel. Six junk topics (advertisements, navigation text, etc.) are excluded. Production shares are weighted by article position on the page and time spent on the front page; see Appendix A.4.

The usefulness of our topic-based description of news coverage depends on the degree to which variation in the mixture of coverage across our 29 substantive topics captures the important dimensions of differentiation between news outlets. As we show in the next section, there is indeed substantial heterogeneity across outlets in their topical emphasis. Topic weights capture quite sharply the difference between a generalist outlet like the *Washington Post* and an entertainment-focused outlet like the *Hollywood Reporter*.

But the topic model can also partially capture other important forms of differentiation, such as political slant. At least some portion of political slant manifests in choices of which stories to cover, rather than how to frame the same stories; see for example Broockman and Kalla (2025), who

document large differences in the coverage of specific events between the Fox News Channel and CNN. In principle, a sufficiently fine partition of topics could capture most or all of what we think of as slant. For practical reasons, our partition is more coarse, and will capture slant only partially. The remaining component of slant—different emphasis on specific stories within the same topic, or different framing of the same story—will load in our demand model onto the baseline outlet preference that each consumer has for each outlet, regardless of what topics they cover.

4 Descriptive Analysis

In this section, we present several stylized facts about the supply and demand of online news. We first show that our topic-level production shares track real-world events and vary differentially across outlets by specialization. Changes in outlet staffing affect the distribution of topics that are covered. We then show that consumers differ in their topic consumption patterns and that outlets' production decisions causally shift what their readers consume. Together, these facts motivate the structural demand model in Section 5.

4.1 Supply

The front page of a newspaper—whether digital or print—is generally the first part of the newspaper that consumers see, and often the part that is most likely to be read. As such, the front page offers a vantage point through which to view the range of articles that different newspapers present to readers over time. In Table 1 we present summary statistics of the topic coverage across the front pages of outlets in our GDELT panel. Our panel covers 3,244 outlets across 365 days between January 1, 2022 and December 31, 2022.

For each outlet-day, we compute the *production share* of each topic: the fraction of front-page articles devoted to each topic, weighted by each article's position on the page and time spent on the front page (see Appendix A.4 for the precise definition). This measure captures the relative prominence of different topics in an outlet's coverage on a given day.

The median outlet displays 41 unique articles per day on its front page, each of which stays on the front page for an average of 6 hours. However, there is substantial variation across outlets: the 25th percentile outlet displays 20 articles per day, while the 75th percentile outlet displays 93. Most outlets spread coverage across many topics: the median outlet devotes 26% of its front page to its single most-covered topic, and 50% to the three most-covered topics. The median normalized entropy is 0.76.[7] A non-trivial tail of outlets is more specialized: at the 90th percentile outlet, the top topic accounts for nearly half of all front-page coverage.

Production shares vary over time as newsworthy events shift the composition of coverage. Figure 2 plots the average production share of two example topics, "American Sports" and "Weather and Forecasts", across all outlets for each day in 2022. In each case, we highlight major events related to the relevant topic in vertical dashed lines. As the figures show, many of the major events—such as the Super Bowl, the NFL draft, and large storms like Hurricane Ian and Winter Storm Elliot—correspond

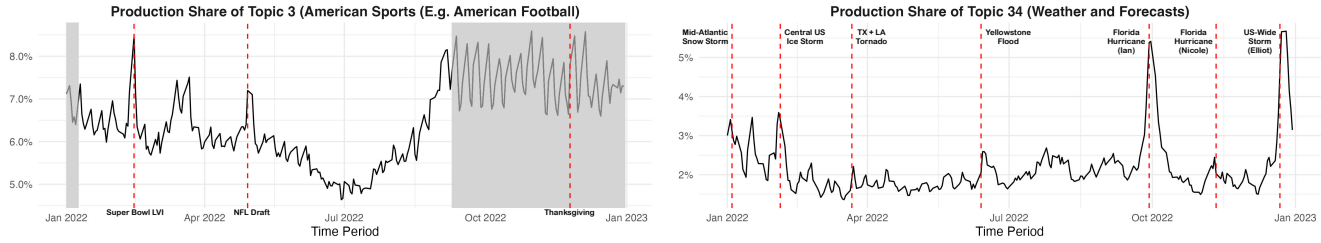
[7] A normalized entropy of 1.00 would indicate uniform coverage across all topics.

	10%	25%	Median	75%	90%	Mean	SD
Number of Articles per Day	8.05	20.03	40.93	93.28	174.58	72.44	90.33
Avg Hours per Article per Day	2.10	3.88	5.70	7.41	9.01	5.67	2.60
% Coverage of Top Topic	0.15	0.19	0.26	0.37	0.47	0.29	0.14
% Coverage of Top 3 Topics	0.36	0.42	0.50	0.60	0.71	0.52	0.14
Entropy of Coverage Across Topics	0.57	0.68	0.76	0.81	0.85	0.73	0.12

Table 1: Distribution of front-page news production across outlets (2022).

Notes: Each row reports the distribution across the 3,244 outlets in our GDELT panel. "% Coverage of Top Topic" is the production share of an outlet's single most-covered topic; "Top 3 Topics" is the combined production share of its three most-covered topics. Entropy is normalized to be between 0 and 1 such that 1 indicates uniform coverage across all topics.

to substantial spikes in the aggregate share of front-page coverage devoted to the relevant topic. These patterns suggest that daily variation in aggregate production shares tracks the occurrence of real-world events, making it a useful proxy for the volume of newsworthy events.



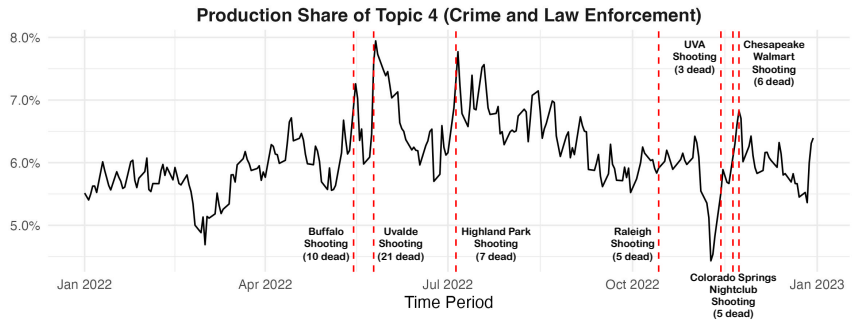
(a) American Sports *(b) Weather and Forecasts*

Figure 2: Daily production shares track real-world events.

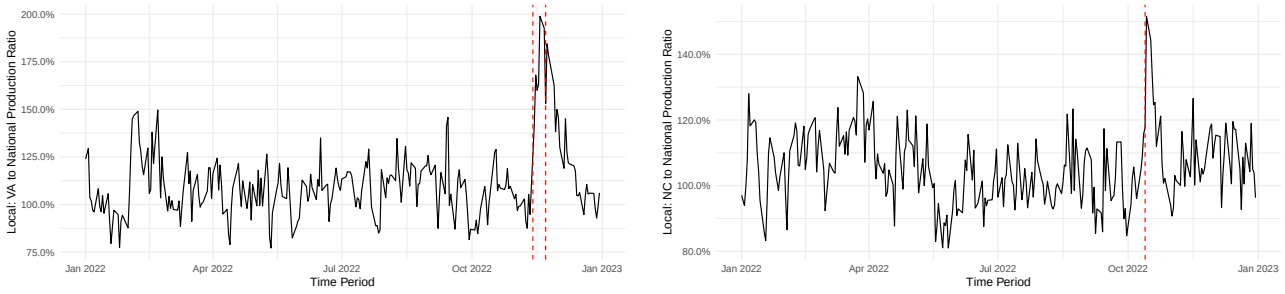
Notes: Each panel plots the average daily production share of a topic across all outlets in 2022. Vertical dashed lines mark major events (e.g., the Super Bowl, Hurricane Ian).

Figure 3a shows a similar plot for the "Crime and Law Enforcement" topic. Here, the vertical dashed lines mark the incidence of mass shooting events throughout the United States in 2022. Although the largest events, such as those in Buffalo, NY, Uvalde, TX and Highland Park, IL correspond to noticeable spikes in aggregate coverage, other mass shooting events such as those in Raleigh, NC and the University of Virginia (UVA) do not. However, when we zoom in on local outlets in North Carolina and Virginia in Figure 3b and Figure 3c, we see substantial spikes in coverage of the crime tpic at the time of the shooting events. These are stylized examples for which notable events are easy to identify, but they highlight the role of specialization across outlets. When events with more specialized interests—in this case, local audiences in North Carolina and Virginia—occur, outlets specialized in the relevant topics increase their coverage sharply even if the national aggregate does not move.

Panel (a) of Figure 5 generalizes this differential response across all topics. For each topic k and



(a) All outlets



(b) Virginia local outlets

(c) North Carolina local outlets

Figure 3: Local outlets respond more to local events than the national aggregate.

Notes: Topic is "Crime and Law Enforcement" in 2022. Panel (a): average daily production share across all outlets. Panels (b) and (c): ratio of local outlets' production share to the national average for Virginia and North Carolina. Vertical dashed lines mark large mass shooting events in each state.

day d, the x-axis plots the aggregate volume of coverage of k on d relative to the average coverage of k across all days. The y-axis plots the difference in coverage between outlets specialized in topic k and all other outlets. We define an outlet as specialized in a particular topic if its average production share of that topic ranks in the top 100 outlets in our sample.[8] Each point represents the average of a bin of topic-day pairs. By construction, outlets that specialize in a topic publish substantially more on it, devoting about 6 times as much coverage to their topic as the average outlet on an average day. However, this gap *widens* on days with especially high aggregate coverage. That is, when newsworthy events drive spikes in production shares across the news industry, the production shares of outlets specialized in the relevant topic increase disproportionately more.

Outlet specialization embeds two dimensions of heterogeneity: a persistent outlet-wide expertise and a potentially time-varying journalist-led expertise. At the extreme, these two dimensions go together—the Athletic, for instance, primarily employs sports journalists and primarily publishes articles about sports. However, for many outlets and topics, specialization varies over time. To demonstrate this, we link each outlet's topic coverage to the composition of its newsroom. Using our

[8]This continuous, production-share-based measure of topic specialization is distinct from the categorical outlet-type label (e.g., "Sports," "Politics") described in Appendix A.1. The production-share ranking is topic-specific and data-driven, while the categorical label is assigned once per outlet across all topics.

Dependent variable:	$p_{j,k,d}$
Variables	
$p^{\text{l1o}}_{j,k,d}$	15.67***
	(0.1188)
num journalists$_{j,k,m(d)}$	1.243***
	(0.0079)
Fixed effects	
Outlet × topic ($\gamma_{j,k}$)	Yes
Outlet (μ_j)	Yes
Topic (ν_k)	Yes
Fit statistics	
Observations	17,331,367
Pseudo R^2	0.17675

Table 2: Production shares rise with lagged topic-specific staffing.

Notes: Poisson regression estimates of outlet-topic production shares on staffing. The dependent variable $p_{j,k,d}$ is outlet j's production share of topic k on day d; $p^{\text{l1o}}_{j,k,d}$ is the leave-one-out average production share of topic k across all other outlets on day d; and journalists$_{j,k,m(d)}$ is the number of journalists employed in month $m(d)$ who specialized in topic k at outlet j in the preceding month. All estimates include a composite outlet-by-topic fixed effect ($\gamma_{j,k}$) together with outlet (μ_j) and topic (ν_k) fixed effects. ***: $p < 0.001$; **: $p < 0.01$; *: $p < 0.05$.

panel of monthly outlet-topic staffing counts, we estimate a Poisson regression of each outlet's daily production of each topic on the count of journalists who had recently published heavily in that topic and were employed by the outlet at the time. We control for time-varying newsworthiness at the topic level with a leave-one-out average production share, and for persistent outlet-wide specialization with outlet-by-topic fixed effects. The remaining variation—generated by job switches among journalists who had specialized in a topic at one outlet and moved to a new position elsewhere—identifies the extent to which specialized staffing drives outlet-level topic coverage. We find a large, statistically significant effect, in which one additional specialized journalist corresponds to 8% of the added coverage from moving the leave-one-out average production share from zero to one. This suggests that while outlets likely publish on news stories of interest to their audiences, the extent of their coverage constrained by their slow-moving on-staff expertise.

4.2 Demand

Table 3 summarizes news consumption across the 83,028 panelists who spend at least five minutes on news websites during 2022, out of roughly 470,000 total panelists in the Comscore data. Three facts stand out. First, most panelists read very little news. The median panelist spends 18 minutes on news websites over the entire year, and the 90th percentile spends under three hours; the mean of 105 minutes is pulled up by a thin upper tail.[9] News is also a small part of online life: the median panelist spends about 39 hours on non-news websites, so news accounts for roughly one percent of

[9]Note that these statistics are computed among our estimation sample, who by definition spend at least 5 minutes on news websites over the year. The median panelist in the entire Comscore sample spends 0 minutes on news websites.

their online time. The top decile of panelists by reading time—those with more than 2.8 hours over the year—account for the majority of news reading in the sample.

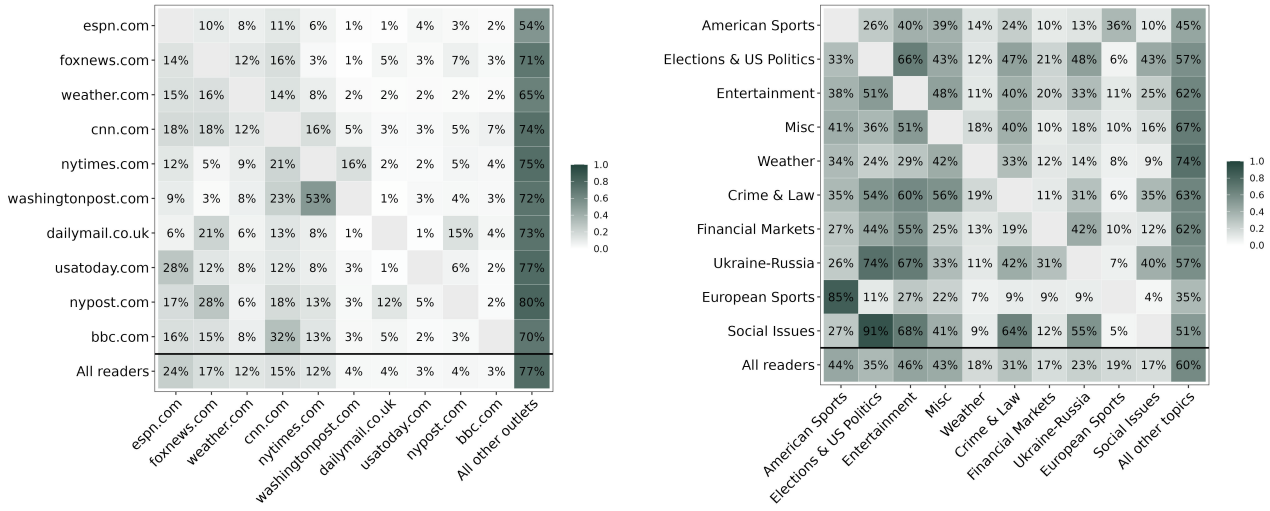
Second, news reading is only somewhat concentrated within each panelist. The median panelist visits 7 distinct outlets and reads 9 distinct topics over the year, devoting 35% of reading time to their most visited topic and 72% to the top three topics.

Third, what panelists read differs widely across panelists. Panel B of Table 3 reports, for the ten most-read topics, the share of panelists with any reading time on the topic and the distribution of the topic's share of reading time across all panelists. No topic reaches more than 70% of panelists: Entertainment (69%), Digital Tech (61%), and American Sports (57%) come closest, while the hard-news topics—Elections and US Politics, Crime and Law, Ukraine–Russia—are read by only 35–43%. For six of the ten most-read topics, the median share is zero. Among those who do read a topic, it often occupies a large share of their time: the 90th percentile of the share of reading time spent on American Sports is 50%; for Entertainment, the 90th percentile is 37%. Topic choices are not a matter of everyone reading the same mix in different amounts; most panelists skip most topics, and so their reading time is split among the few topics they consider at all.

	Any (%)	10%	25%	Median	75%	90%	Mean	SD
Panel A: Volume and concentration of news reading								
Minutes Reading News		6.18	8.77	17.63	48.25	156.90	105.47	545.24
Minutes Online, Non-News		409.51	1,011.27	2,367.27	5,202.25	10,292.12	4,326.38	5,894.33
Number of Outlet-Day Visits		3.00	5.00	10.00	26.00	73.00	36.82	106.24
Number of Unique Outlets Read		2.00	4.00	7.00	13.00	25.00	11.61	16.47
Number of Unique Topics Read		4.00	5.00	9.00	15.00	22.00	10.79	6.92
% Consumption of Top Topic		20.76	26.32	35.36	48.87	64.05	39.29	17.29
% Consumption of Top 3 Topics		47.89	57.79	71.86	86.47	97.75	71.93	17.75
Panel B: Share of reading time on each topic, in percent (top 10 topics)								
American Sports	57.2	0.00	0.00	1.78	15.76	50.30	13.28	22.37
Entertainment	68.5	0.00	0.00	5.47	18.29	36.51	12.59	17.36
Elections & US Politics	37.3	0.00	0.00	0.00	2.88	11.96	3.59	8.57
Misc	59.2	0.00	0.00	2.30	9.04	19.35	6.74	11.07
Weather	35.6	0.00	0.00	0.00	2.26	13.79	4.28	11.08
Crime & Law	42.5	0.00	0.00	0.00	4.15	12.94	4.00	8.72
Financial Markets	38.5	0.00	0.00	0.00	2.76	12.91	4.06	10.16
Digital Tech	60.7	0.00	0.00	2.17	11.74	27.79	8.82	14.36
European Sports	42.0	0.00	0.00	0.00	3.11	10.45	3.49	8.46
Ukraine-Russia	34.6	0.00	0.00	0.00	1.78	8.56	2.77	7.40

Table 3: Distribution of news consumption across panelists (2022): All news consumers.

Notes: Panel A reports the distribution across the 83,028 news consumers in the Comscore panel over 2022 (panelists with at least five minutes of news reading over the year). "% Consumption of Top Topic" is the share of a panelist's news reading time on their single most-read topic; "Top 3 Topics" is the combined share of their three most-read topics. Panel B covers the 82,888 panelists with time on substantive topics; shares are in percent of a panelist's reading time excluding the junk/sections bucket. "Any" is the share of panelists with any time on the topic. Mean and SD are taken across all panelists, counting those with no time on the topic as zeros; the percentiles also include those zeros.



(a) Outlets *(b) Topics*

Figure 4: Co-reading across outlets and across topics.

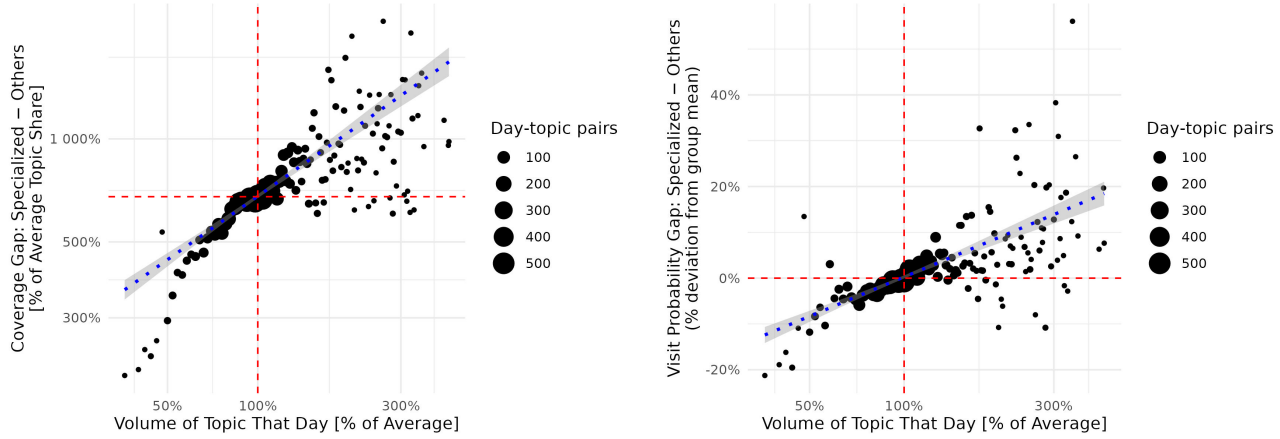
Notes: Each cell is the share of readers of the row outlet (left) or row topic (right) who also read the column outlet or topic during 2022, among frequent consumers. A panelist counts as a reader of an outlet or topic if it accounts for at least 5% of their total news reading time over the year. Each panel shows the 10 most-read outlets or topics by total reading time, ordered from the top, with all remaining outlets or topics pooled into the last row and column; a panelist reads the pooled group if any outlet or topic in it clears the 5% threshold. Diagonal cells are omitted.

Figure 4 shows how readers combine outlets and topics. Among the ten most-read outlets, co-reading is relatively rare: for most pairs, fewer than one in five readers of one outlet devote 5% or more of their time to the other. The exceptions trace political and brand affinities. More than half of Washington Post readers also read the New York Times, and a fifth read CNN, while only 3% read Fox News; conversely, 16% of Fox News readers read CNN but only 3% read the Times and 1% the Post. However, readers of every top outlet are far more likely to have a substantial share on some outlet outside the top ten (54–80%) than on any single other top outlet. This suggests that the long tail of outlets with disparate audiences, rather than the top brands, are the drivers of multihoming in online news consumption.

Overall topic co-consumption follows two main blocks. In the aggregate over the course of 2022, hard-news topics are often read together: 91% of readers of Social Issues also read Elections and US Politics, as do 74% of readers of the Ukraine–Russia war, with Crime and Law close behind. Sports forms its own block: 85% of European Sports readers also read American Sports, but only 11% read Elections. Entertainment is co-read with nearly everything, while Weather is not systematically co-read with anything other than small local content topics.

However, these aggregate co-consumption patterns smooth over the periodic shocks to specific topics as newsworthy events pertaining to them occur. As panel (b) of Figure 5 illustrates, when overall interest in a topic spikes, readers respond by disproportionately visiting outlets that are specialized in that topic—and which, as panel (a) shows, cover stories in that topic more heavily. In fact, visits

to specialized outlets increase even among readers who had never visited those outlets before in our panel.[10] This suggests that readers have intrinsic preferences for topics, and that they use are willing to switch outlets in order to express those preferences.



(a) Coverage *(b) Consumption*

Figure 5: Specialized outlets produce more coverage and are more likely to be visited on high-event days.

Notes: Binscatters pooling all topic × day pairs. The x-axis is the aggregate production share of a topic on a given day, relative to that topic's average. The y-axis is the gap between outlets specialized in the topic and other outlets in production share (left) or the probability of a visit (right). An outlet is specialized in a topic if its average production share ranks in the top 100. The probability of a visit is the share of panelist–outlet pairs in the group with positive reading time on the outlet that day, on any topic; it is normalized as the percentage deviation from the group's own time-series average, computed separately for specialized and other outlets, and the gap is the difference in these normalized values. Dashed red lines mark the average-day reference point. Point size reflects the number of topic-day pairs in each bin.

4.3 Evidence of Agenda Setting

The evidence above suggests that consumers read more on specialized outlets on days with newsworthy events relevant to their topics of specialization. This suggests that consumers pursue outlets that cover topics of interest to them. We now ask the reverse question: do outlets' decisions of what to cover change what their readers consume?

Article-level evidence. We begin at the article level. Outlets adjust their front pages multiple times a day: articles appear, move up or down the page, and eventually drop off. We exploit this within-article variation to test whether front-page placement causally drives readership. For each article a observed across hours h on a given day, we estimate:

$$y_{a,h} = \rho \cdot \mathbf{1}[\text{front page}]_{a,h} + f(\text{age}_{a,h}) + \phi_a + \zeta_h + \varepsilon_{a,h} \tag{1}$$

[10] See Appendix figure 21 in Appendix F.

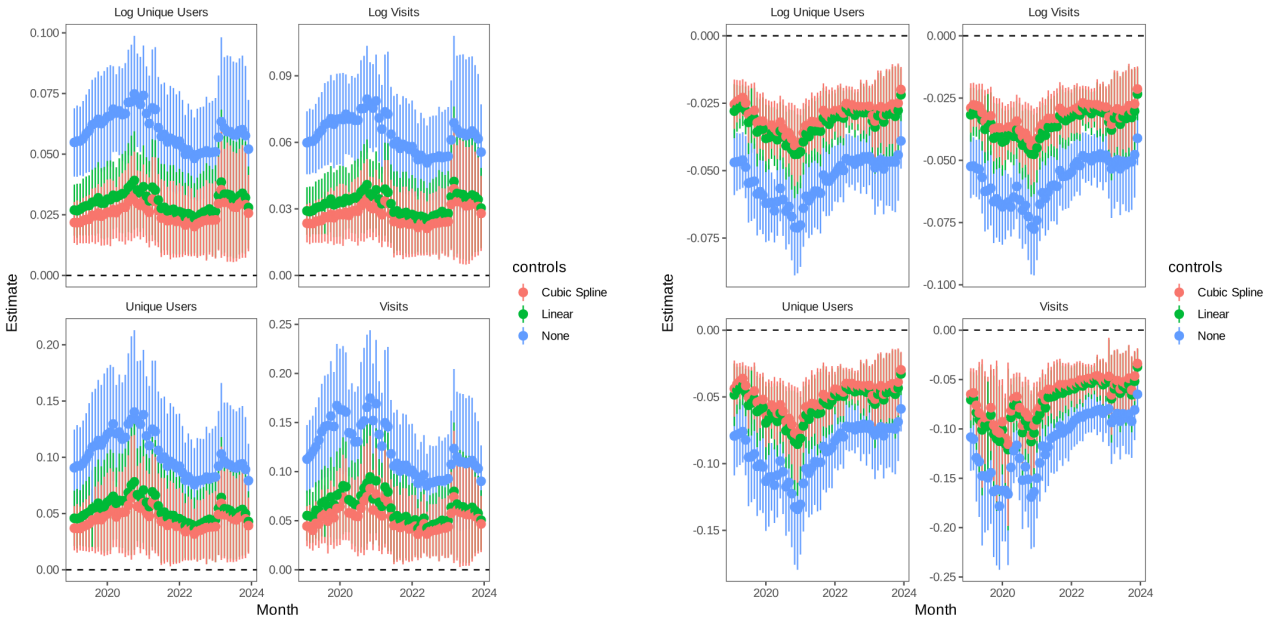
where $y_{a,h}$ is a measure of readership of article a in hour h (either unique visitors or total visits, in levels or logs), $\mathbf{1}[\text{front page}]_{a,h}$ indicates whether the article appears on the outlet's front page that hour, $f(\text{age}_{a,h})$ is a flexible control for article age (hours since first appearance, with linear and cubic-spline specifications), ϕ_a are article fixed effects, and ζ_h are hour-of-day fixed effects. The article fixed effects absorb all time-invariant differences across stories, so the coefficient ρ is identified from within-article variation in placement over time, relative to the expected rate decay of interest in a story as it ages $f(age_{a,h})$. The identifying assumption is that the shape of this decay in reader interest is smooth with respect to time and common across articles. Under this assumption, discontinuous changes in front-page positioning identify the positioning effect.

Panel (a) of Figure 6 plots the estimated coefficient ρ by month.[11] An article that is on the front page of an outlet has approximately 0.05 more unique Comscore readers than an article that is not on the front page—a difference equivalent to about 50% of the overall mean. This effect is stable across months and robust to the functional form of the age controls. Omitting the age control yields an estimate roughly twice as large, confirming that front-page placement is, as expected, positively correlated with newsworthiness. Panel (b) estimates a complementary specification restricted to articles currently on the front page, regressing readership on the article's relative position (0 = top of page, 1 = bottom). Placement towards the top of the homepage further increases readership: moving an article from the bottom to the top of the front page has an effect comparable in magnitude to placing it on the front page in the first place. Together, these results confirm that editorial placement decisions—not just the intrinsic interest of stories—drive what readers consume.

Appendix C reports estimates from an additional specification of Figure 6 that adds event-by-calendar-hour fixed effects. These fixed effects absorb changes in interest in the underlying news event that occur simultaneously across outlets. In this sense, they go beyond our baseline control for article-specific interest. Because classifying articles into events requires full text, we can only run this specification on the subset of articles that appear in the PeakMetrics article corpus. However, the qualitative results are similar to those presented in Figure 6.

Topic-level evidence. The article-level results establish that placement matters for individual stories. We now test whether changes in the *topic composition* of an outlet's front page shift its readers' consumption at the topic level. A useful source of identifying variation comes from the supply side. When a major event occurs, many outlets increase coverage of the relevant topic simultaneously, but as we showed in Figure 5a, outlets that specialize in the topic respond more strongly than others. We construct an instrument that captures this interaction: for each outlet j, topic k, and day d, we fit a Poisson regression of j's production share on the leave-one-out average production share of all *other* outlets on that topic-day, with outlet×topic fixed effects (see Appendix A.5). The fitted value $z_{j,k,d} = \exp(\hat{\alpha}_{jk} + \hat{\beta}\,\bar{p}_{-j,k,d})$ serves as our instrument. The exponential form gives the instrument a shift-share structure: the same aggregate news shock produces a larger predicted shift for outlets that specialize in the topic. The quasi-random timing of differentially newsworthy events across topics then

[11] We estimate the regression with one month of data at a time for computational feasibility.



(a) On front page (indicator)

(b) Front-page position

Figure 6: Front-page placement causally drives article-level readership.

Notes: Monthly estimates of front-page placement effects on article readership. Panel (a): coefficient ρ from equation (1), the effect of being on the front page. Panel (b): coefficient on front-page position (0 = top, 1 = bottom) among articles on the front page. Each panel shows results under three age-control specifications (none, linear, cubic spline) and for multiple outcome measures (unique users, visits, in levels and logs). All regressions include article and hour fixed effects. Regressions are estimated separately by month.

drives exogenous variation in how much each outlet covers each topic, over its typical topic share.[12]

We estimate the following specification at the outlet-topic-day level:

$$c_{j,k,d} = \theta \cdot p_{j,k,d} + \mu_{j,k} + \eta_{k,d} + \varepsilon_{j,k,d}, \tag{2}$$

where $c_{j,k,d}$ is the consumption of topic k on outlet j on day d (measured as hours of reading time), $p_{j,k,d}$ is the outlet's production share of that topic, $\mu_{j,k}$ are outlet-by-topic fixed effects that absorb time-invariant differences in topic coverage and readership across outlets, and $\eta_{k,d}$ are topic-by-day fixed effects that absorb aggregate shocks to topic salience. The coefficient θ captures the within-outlet, within-topic-day effect of increased coverage on consumption. OLS estimates of θ may be biased if outlets adjust their coverage in response to anticipated reader demand, or if measurement error in production shares attenuates the coefficient toward zero. To address these concerns, we instrument

[12] See Appendix A.5 for the formal derivation. To first order, $z_{j,k,d} \approx \bar{p}_{j,k} \cdot [1 + \hat{\beta}(\bar{p}_{-j,k,d} - \bar{\bar{p}}_{-j,k})]$, where $\bar{p}_{j,k}$ is the outlet's long-run average share of the topic (the "share") and $\bar{p}_{-j,k,d} - \bar{\bar{p}}_{-j,k}$ is the deviation of other outlets' coverage from its mean on day d (the "shift"). We assume that each outlet is individually small, so that $\bar{p}_{-j,k,d}$ is a proxy for aggregate newsworthiness of topic k on day d, uncorrelated with the idiosyncratic demand of typical consumers of outlet j. Appendix A.6 tests this assumption directly by rebuilding the instrument excluding peer outlets that share either j's specialization or its audience.

for $p_{j,k,d}$ using $z_{j,k,d}$ described above.

	National	Local	International
$\hat{\theta}$	1.285	0.171	2.961
	(0.169)	(0.249)	(0.604)
First-stage F	580.9	475.4	514.6
N	3,427,788	17,368,148	645,456

Table 4: Effect of outlet production share on consumption (IV).

Notes: IV estimates of θ from equation (2), pooling across all topics. Each column restricts to outlets of the indicated geographic scope. The instrument is the fitted value from a Poisson regression of the outlet's production share on the leave-one-out average, with outlet×topic fixed effects (see Appendix A.5). Standard errors are clustered by time period.

Table 4 reports the results separately for national, local, and international outlets. The first-stage F-statistics are well above conventional thresholds for all three outlet scopes. The IV estimates of θ are positive and significant for national and international outlets: a one-percentage-point increase in production share increases consumption by 1.3 and 3.0 percentage points, respectively. The IV estimate for local outlets is smaller and imprecisely estimated. A similar exercise split by topics in Appendix D shows that while the effect is positive for the majority of topics, the magnitude varies substantially—higher coverage of some topics elicits a much stronger consumption response than others, consistent with heterogeneous topic preferences among readers. The estimates for national and international outlets are also robust to how the leave-one-out average is constructed: building it only from outlets outside j's specialization, or weighting peer outlets by how little of j's audience they share, yields estimates of θ within 0.1 of those in Table 4 (Appendix A.6).

Taking Stock Together, these exercises suggest that both agenda setting and topic-specific preferences play a role in shaping the news diet, but they cannot quantify the extent to which each channel would contribute to the demand response to coverage changes by a given outlet. To do so, in Section 5 we develop a model of news time allocation as a function of outlet-specific and topic-specific preferences. Using variation in the incidence of newsworthy events throughout 2022 and the differential response of outlets with different specialties discussed in this section, we estimate the distribution of consumer preferences from observed shifts in the mix of reading time across outlets and topics.

5 A Model of News Consumption

We posit a time-allocation model adapting that in Crawford and Yurukoglu (2012) and Martin and Yurukoglu (2017), in which, on every day d, consumer i derives utility that is proportional to the logarithm of one plus the time (in hours) allocated to reading news from outlet j about some topic k,

$T_{i,j,k,d}$. That is:

$$u_{i,d} = \sum_{j=1}^{J}\sum_{k=1}^{K} v_{i,j,k,d}\log(1+T_{i,j,k,d}) + v_{i,O,d}\log(1+T_{i,O,d}) + v_{i,F,d}\log(1+T_{i,F,d}).$$

The set of $v_{i,j,k,d}$ across outlets and topics determine the consumer's marginal utility from allocating additional time to reading news from different outlets on different topics each day. $v_{i,O,d}$ is the marginal value of online time devoted to non-news-reading activities (e.g., browsing social media) and $v_{i,F,d}$ is the marginal value of offline time. We choose the normalization $v_{i,F,d} = 1 \ \ \forall i, d$, such that all other values are relative to the value of offline time for the same person on the same day.

Each day, the consumer's problem is then to maximize utility subject to a time budget constraint:

$$\begin{aligned} \max_{\mathbf{T}_{i,d}} & \sum_{j=1}^{J}\sum_{k=1}^{K} v_{i,j,k,d}\log(1+T_{i,j,k,d}) + v_{i,O,d}\log(1+T_{i,O,d}) + \log(1+T_{i,F,d}) \\ \text{s.t.} & \sum_{j,k} T_{i,j,k,d} + T_{i,O,d} + T_{i,F,d} = 24, \\ & \mathbf{T}_{i,d} \geq 0, \end{aligned} \tag{3}$$

where the inequality constraint in the third line holds for every element of the vector $\mathbf{T}_{i,d}$. The solution to this problem has a simple characterization. For all j, k with $T_{i,j,k,d} > 0$, and so long as time is not fully allocated to online activity such that $T_{i,F,d} > 0$:[13]

$$v_{i,j,k,d} = \frac{1+T_{i,j,k,d}}{1+T_{i,F,d}}. \tag{4}$$

The derivation of this relationship is straightforward: convert the constrained maximization problem above to Lagrangian form, with multipliers $\lambda^B_{i,d}$ on the time budget constraint and $\lambda^Z_{i,j,k,d}$ on the nonnegativity constraints. By complementary slackness, $\lambda^Z_{i,j,k,d} = 0$ for all nonzero times $T_{i,j,k,d}$, and the first-order condition thus yields:

$$v_{i,j,k,d} = \lambda^B_{i,d}(1+T_{i,j,k,d}).$$

This condition also holds for the normalized outside-option (offline) time, implying

$$\lambda^B_{i,d} = \frac{1}{1+T_{i,F,d}} \tag{5}$$

which gives the desired result.

Equation (4) implies that, for combinations of j and k to which consumer i devotes non-zero time, there is a linear relationship between the consumer's marginal value of reading topic k on outlet j, and the normalized share of time that the consumer i allocates to the combination j, k relative to the outside option of spending time offline. For the combinations for which $T_{i,j,k,d}$ is exactly zero, $v_{i,j,k,d}$

[13] We take the time budget to be the full 24 hours in a day; though not impossible, a consumer spending all 24 hours of a day online is unlikely.

is not point-identified, but is bounded above by $\lambda^B_{i,d}$.

We specify a parsimonious form of the $v_{i,j,k,d}$'s which allows for consumers to have tastes both for topics and for outlets. Specifically,

$$v_{i,j,k,d} = \exp\left(\alpha_{i,j} + (\beta_{i,k} + \gamma_{i,j})\dot{p}_{j,k,d} + \delta^N_{i,w(d),m(d)} + \varepsilon_{i,j,k,d}\right) \tag{6}$$

$$v_{i,O,d} = \exp\left(\delta^O_{i,w(d),m(d)} + \varepsilon_{i,O,d}\right), \tag{7}$$

where, letting $p_{j,k,d}$ be the share of production that outlet j devotes to topic k on day d, $\dot{p}_{j,k,d} \equiv p_{j,k,d} - \frac{1}{D}\sum_d p_{j,k,d}$ is outlet j's deviation, on day d, from its long-term average production share of topic k.

The outlet fixed effects $\alpha_{i,j}$ capture consumers' average preference for visiting outlet j. They also absorb consumer preferences for the *average* mixture of topics that outlet j produces. The $\beta_{i,k}$ coefficients capture active preferences for topics by controlling how much more a consumer wants to read any outlet that publishes more than it usually does on topic k. The $\gamma_{i,j}$ coefficients allow for agenda setting, controlling how much more consumers read of topic k when an outlet j that they trust publishes more about that topic.

We impose that β and γ are non-negative to satisfy the logical condition that if an outlet publishes more of some topic, readers should not spend less time reading *that topic on that outlet*. Since β and γ multiply the production share, they represent marginal utilities of spending time reading a given topic on a given outlet when that outlet allocates a larger share of its production to that topic. The non-negativity constraint imposes that these marginal utilities are weakly positive, which is sensible given that higher production share means that the outlet either produces more distinct articles on the topic or emphasizes them more on the front page, making them easier for readers to find. To avoid collinearity in the utility specification, we must drop one of the outlet or topic interactions. We choose $\beta_{i,0} = 0$ as our excluded topic interaction.

The time fixed effects $\delta_{i,w(d),m(d)}$ allow for interest in reading news and doing non-news activities on the internet to vary over time; we use day-of-week ($w(d)$) and month ($m(d)$) fixed effects in our baseline specification. As will become clear shortly, only the difference $\delta^N_{i,w(d),m(d)} - \delta^O_{i,w(d),m(d)}$ is identified by our method. We therefore set $\delta^N_{i,w(d),m(d)} = 0$ and estimate only $\delta^O_{i,w(d),m(d)}$. The residuals ε are stochastic error terms that allow for random variation in reading.

We plug equations (6) and (7) into (4), take logs of both sides, and difference to arrive at our estimating (in)equality:

$$\log(1 + T_{i,j,k,d}) - \log(1 + T_{i,O,d}) \geq (\beta_{i,k} + \gamma_{i,j})\dot{p}_{j,k,d} + \delta^N_{i,w(d),m(d)} - \delta^O_{i,w(d),m(d)} + \varepsilon_{i,j,k,d} - \varepsilon_{i,O,d} \tag{8}$$
$$\forall (i,d) : T_{i,O,d} > 0.$$

This inequality holds whenever $T_{i,O,d} > 0$, i.e. on days on which the consumer had some non-news online time.[14] In practice nearly all[15] consumer-days with positive time on any news site also have

[14] We use the fact that (4) holds with equality for non-news online time whenever $T_{i,O,d} > 0$ in deriving the relation in (8); otherwise the direction of the comparison is indeterminate.

[15] More than 99%.

positive time on non-news sites, and so limiting to days where $T_{i,O,d} > 0$ is not a significant restriction on the data.

Inequality (8) holds with equality whenever $T_{i,j,k,d} > 0$. Hence, the time-allocation structure and our utility specification together yield a censored regression model with differences in log time on the left-hand side and (de-meaned) production shares on the right-hand side.

5.1 Reader heterogeneity

The model generates a set of equality and inequality restrictions that we can use to estimate the preference parameters $\alpha, \beta, \gamma, \delta$. In principle, we could estimate these parameters separately for each individual consumer i in our data. In practice, however, the computational demands of estimation are large enough to make this approach infeasible. Instead, we group consumers together into clusters, based on their news reading habits and geographic proximity, and assume that preferences are common among consumers in the same cluster. See Appendix B.3 for details of our clustering method.

The result of this procedure is 1180 distinct clusters: 20 clusters each from 59 distinct geographic regions, namely the top 50 US DMAs by population plus users residing in non-top-50 DMAs in each of the nine US Census sub-regions. We assume that all preference parameters are shared among users within a cluster, but place no restrictions on heterogeneity in preferences across clusters. We are therefore able to capture rich variation in preferences across different kinds of news consumers and different geographic markets. While our estimation will be at the cluster level, using data from one cluster at a time, we are interested in the population distribution of these parameters rather than their values for any given cluster.

For each cluster, we estimate the set of demand model parameters for the subset of outlets and topics which *some* member of the cluster visits during the year.[16] We additionally impose a minimum of 25 visits by members of the cluster to estimate an outlet- or topic-specific parameter. Outlets visited less than 25 times are given a common parameter. E.g., if users in cluster i visit outlets j and j' more than 0 but less than 25 times, we impose $\alpha_{ij} = \alpha_{ij'}$ and $\gamma_{ij} = \gamma_{ij'}$. Topics visited less than 25 times are pooled with the excluded topic (0), which, per the discussion above, is normalized to have $\beta_{i0} = 0$.

5.2 Estimation

Our estimation problem is a censored regression model (where censoring occurs from below) with a potentially endogenous regressor, $\dot{p}_{j,k,d}$. The source of endogeneity here is that if outlet j knows that its readers may be especially interested in some topic k on some day d (i.e., ε_{ijkd} is unusually high for readers of j), it may choose to publish more than average on that topic that day. This would lead on average to positive correlation between $\dot{p}_{jkd}$ and ε_{ijkd} and hence over-estimation of the agenda setting parameters γ. We therefore apply the method of Blundell and Powell (2007), which allows for consistent estimation of the parameters $\alpha, \beta, \gamma, \delta$ with censored outcome data and endogenous regressors, and which does not require specification of the distribution of ε.

[16] Parameters for outlets or topics which are never visited are not point-identified, only bounded above.

The method relies on a control function approach. The key identifying assumption is a quantile exclusion restriction: that conditional quantiles of the error term ε are functions only of the control variable $v \equiv \dot{p} - \pi(z)$, where z is an excluded instrument and π is a function that will be estimated nonparametrically. v captures residual variation in $\dot{p}$ given z; the identifying assumption is that once this residual is controlled for, remaining variation in ε is independent of $\dot{p}$.

Our instrument z is the fitted value from a Poisson regression of production shares on the average of all other outlets' production share of that topic that day, plus outlet $\times$ topic fixed effects. The instrument, as discussed in the Descriptive Analysis section and in more detail in Appendix A.5, follows a shift-share logic. It captures the interaction between the outlet's specialization in a topic (the outlet $\times$ topic fixed effect) and the amount of newsworthy events occurring on that topic that day (measured by the leave-one-out production share average). The Poisson model is multiplicative, implying that specialized outlets respond more to event occurrence in their area of specialization, which matches the empirical pattern documented in Figure 5a. Exogeneity of the instrument relies on the random occurrence of events on different topics at different times.

The method proceeds in three steps:

1. First, we estimate a univariate nonparametric kernel regression of production shares $\dot{p}$ on our instrument z. We extract the residual from this nonparametric regression, $\hat{v}_{i,j,k,d} = \dot{p}_{j,k,d} - \hat{\pi}(z_{j,k,d})$. Because both the regressor and the instrument do not vary across consumers, the nonparametric regression step can be run only once and applied to all clusters.

2. Second, we estimate conditional quantiles of the log attention share outcome $y_{i,j,k,d} \equiv \log(1 + T_{i,j,k,d})$ using a local-linear quantile regression,[17] as described in Khan and Powell (2001) and applied by Blundell and Powell (2007). Our implementation of this method estimates the conditional 99th percentile[18] of y given $\dot{p}$, $y_{i,O,d} \equiv \log(1 + T_{i,O,d})$, z, and binary outlet and topic indicators $\mathbf{d}$. We denote the estimated conditional quantile $\hat{q}(y \mid \dot{p}, z, y_O, \mathbf{d})$.

3. Third, for observations that are predicted to be uncensored at the 99th percentile given the quantile regression in the previous step, we estimate a kernel-weighted regression of pairwise differences in $\hat{q}$ on pairwise differences in $\dot{p}$, in the outlet and topic indicators, and their interactions as specified in equations 6 and 7 above. The weights use an Epanechnikov kernel applied to pairwise differences in the estimated control variable $\hat{v}$. See equation (2.15) of Blundell and Powell (2007) for the form of this regression. We estimate this regression using a non-negative least squares (NNLS) algorithm to satisfy the nonnegativity constraints on β and γ.

The logic of the estimation procedure is that, for a pair of observations which are both uncensored at the chosen quantile ($\hat{q}(y \mid \dot{p}, z, y_O, \mathbf{d}) > 0$) and which have similar levels of the control variable v,

[17] The local-linear method is a standard linear quantile regression restricted to a subset of nearby observations. We use a rectangular kernel with bandwidth suggested by Khan and Powell (2001) on the production share, leave-one-out instrument, and $\log(1 + T_{i,O,d})$ and exact match on outlet and topic identity to determine nearby observations.

[18] The estimand is the same regardless of the choice of quantile. The only difference is where the quantile exclusion restriction is imposed. The necessary identifying assumption in our case is that the conditional 99th quantile of the errors ε is a function only of the control variable v; see Blundell and Powell (2007). We choose the 99th percentile because the censoring in our data is high enough that lower quantiles are usually or always censored for many consumers.

the difference in their conditional quantiles depends only on the difference in the data (production share and outlet / topic indicators) between them. The method uses local-linear quantile regression to estimate q, nonparametric regression to estimate v, and kernel weighting to select pairs of observations which are close enough in v.

Intuitively, the method relies on the deviation from the conditional expectation of $\dot{p}$ given the leave-one-out prediction z to determine whether pairs of observations are appropriate to compare or not. If we observe that a given outlet's production of articles on topic k is much higher than average, on some day where other outlets are not covering k heavily, this likely implies that the outlet knows something about its audience's idiosyncratic interest in k that day: i.e., that $\varepsilon_{i,j,k,d}$ is unusually high for its readers. Comparing readership of j, k on a day like this to more typical days would bias upward our estimate of j's agenda-setting parameter.

The Blundell-Powell method works because an observation like this will have a positive value of ν. The kernel-weighted pairwise regression (step 3 in the estimation method) will down-weight or eliminate comparisons between this observation and others with close to zero or negative ν (where the outlet produces at or below what we expect given its specialization and what other outlets are covering that day). The leave-one-out prediction z allows us to control for observed (market-wide) variation in interest in topic k, and conditioning on the residual $\hat{\nu}$ controls for unobserved (outlet-specific) variation in interest. When the quantile exclusion restriction holds, the pairwise comparisons that inform our estimator are drawn only from pairs of observations with the same distribution of the error term.

5.3 Estimation Results

Our estimation procedure generates predictions of topic preference ($\beta_{i,k}$) and outlet agenda-setting ($\gamma_{i,j}$) coefficients for each market i. For many markets, we do not observe enough readership for a given topic k or outlet j to estimate a non-zero coefficient. The left panel of Figure 7 plots the 50th, 75th, 90th and 95th percentile of topic preferences and agenda-setting for a selection of topics and outlets across markets. The median preference for some topics, like COVID-19, is zero, suggesting that reading articles about COVID-19 was entirely driven by trusted outlets publishing about it rather than by intrinsic interest in at least 50% of markets in our 2022 sample. Other topics like Local Events and Legal & Criminal news have positive topic preferences at the 50th percentile market and coefficients surpassing 0.5 at the tail.

Note however, that the ranking of topics is not constant across preference percentiles. For instance, while the median topic preference for Urban Development is about half that of Local Events, the 95th percentile preference for Urban Development is significantly higher. This captures the sense in which some topics draw niche interest groups: the median reader may not read any urban development articles if they are not front and center, but readers with particular interests in urban development seek out articles about it and read a high share of what is available. Furthermore, niche interests are not uniform across geographies. As the left panel of Appendix figure 22 shows, the median preference for Energy & Environment is more than twice as high in Boston and San Francisco than in Chicago, and is near zero in Atlanta, Baltimore and Tampa. By contrast, American Sports has more similar medians across media markets.

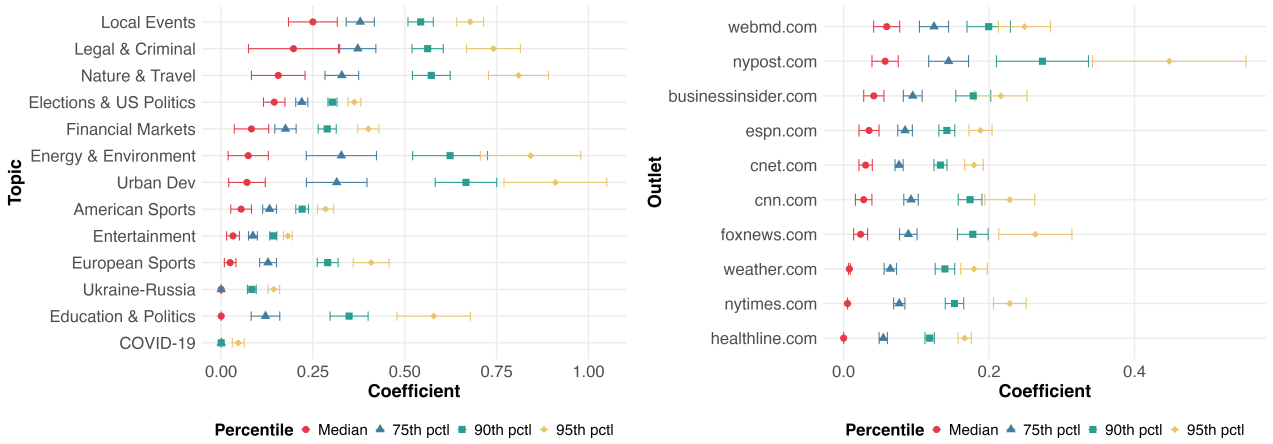


Figure 7: Percentiles of estimated $\beta_{i,k}$ and $\gamma_{i,j}$ coefficients across markets

A similar pattern holds for agenda setting parameters. As the right panel of Figure 7 shows, the ranking of the median outlet agenda-setting coefficient is different from the ranking of the 90th or 95th percentile. While the median $\gamma_{i,j}$ parameter for the New York Post is just a hair higher than that of the New York Times, the 95th percentile is nearly twice as high. Furthermore, the distribution of agenda setting power is not evenly distributed across media markets. As the right panel of Appendix figure 22 shows, for instance, the median agenda setting power across all of the outlets in Figure 7 is higher in Baltimore than in Boston and Atlanta. These differences map to different impacts that editorial decisions might have on readership across outlets and topics in different media markets. All else held equal, editorial shifts by outlets with more agenda setting power would shift their readers' attention more to the topics they prioritize. Shifts by outlets whose audience has strong topic preferences would largely induce their readers to switch to alternatives. How much a given outlet can direct attention on net depends on the distribution of these parameters across its audience, and the mix of news content covered by its competitors.

6 Counterfactuals

Our model allows us to generate predictions for how readership would change under a variety of counterfactual scenarios. In this section, we present a series of exercises that illustrate how sensitive readership of different topics and outlets is to changes in what news is covered and by whom, and how recent changes in editorial direction might impact readership in major outlets on the local and national stage.

In each counterfactual exercise, we take a set of estimated parameters $(\hat{\alpha}_i, \hat{\beta}_i, \hat{\gamma}_i, \hat{\delta}_i)$ and a set of counterfactual residualized production shares $\dot{p}^c_{j,k,d}$ and translate them into simulated readership $\hat{\mathbf{T}}_{i,d}$.[19] Given $\hat{v}_{i,j,k,d}$, the time allocation problem (4) can be solved very quickly with an iterative procedure. However, while the key model parameters can be estimated consistently without any

[19] Following Equation (6), the marginal utility term uses the demeaned counterfactual share $\dot{p}^c_{j,k,d} \equiv p^c_{j,k,d} - \overline{p}_{j,k}$, where $\overline{p}_{j,k} \equiv D^{-1} \sum_d p_{j,k,d}$ is the observed long-run average production share.

assumptions on the distribution function of ε, generating simulated values of $\hat{v}_{i,j,k,d}$ requires additional assumptions on how the error terms are drawn in counterfactuals.

To resolve this we adopt an approach based on resampling with replacement from the empirical distribution. Given a set of estimates $(\hat{\alpha}_i, \hat{\beta}_i, \hat{\gamma}_i, \hat{\delta}_i)$ for some reader cluster i, we construct $N_i \times D$ empirical error vectors $\hat{\varepsilon}_{i,d}$, each of length $J \times K + 1$. These take the form:

$$\begin{aligned}
\hat{\varepsilon}_{i,O,d} &\equiv \log(1 + T_{i,O,d}) - \log(1 + T_{i,F,d}) - \hat{\delta}^O_{i,d} \\
\hat{\varepsilon}_{i,j,k,d} &\equiv \begin{cases} \log(1 + T_{i,j,k,d}) - \log(1 + T_{i,F,d}) - \left(\hat{\alpha}_{i,j} + (\hat{\beta}_{i,k} + \hat{\gamma}_{i,j})\dot{p}_{j,k,d}\right) & T_{i,j,k,d} > 0 \\ -\infty & \text{o.w.} \end{cases}
\end{aligned} \tag{9}$$

To generate simulated readership, we resample with replacement from the observed sample of $\hat{\varepsilon}_{i,d}$. These vectors are treated as IID across users within the cluster and across days. Each time we resample a vector, we also reshuffle its finite elements (i.e., those originating from an observation of positive time allocation in the data) across the subset of (j,k) tuples for which $\hat{q}_{i,j,k,d} > 0$.[20] We draw a sample of $\hat{\varepsilon}_{i,d}$ 100 times per reader-day, solve the time allocation problem for each one, and report the average of the resulting time allocations across all 100 draws.

This method preserves the correlation between non-news time and news time observed in the data. It also preserves the number and values of the finite elements of each vector $\hat{\varepsilon}$ across resamples, while allowing those elements to be reassigned to different (j,k) pairs than were observed in the data. The limitation to (j,k) pairs for which $\hat{q}_{i,j,k,d} > 0$ effectively restricts simulated readers' choice sets to those combinations of j and k. This prevents the simulator from predicting, for instance, that readers will read international politics stories on ESPN that do not exist.[21] It also means that we will never predict readership going outside the set of outlets (or topics) for which some reader in the cluster had positive observed time at some point in the year.

6.1 How Important are Readers' Tastes for Topics?

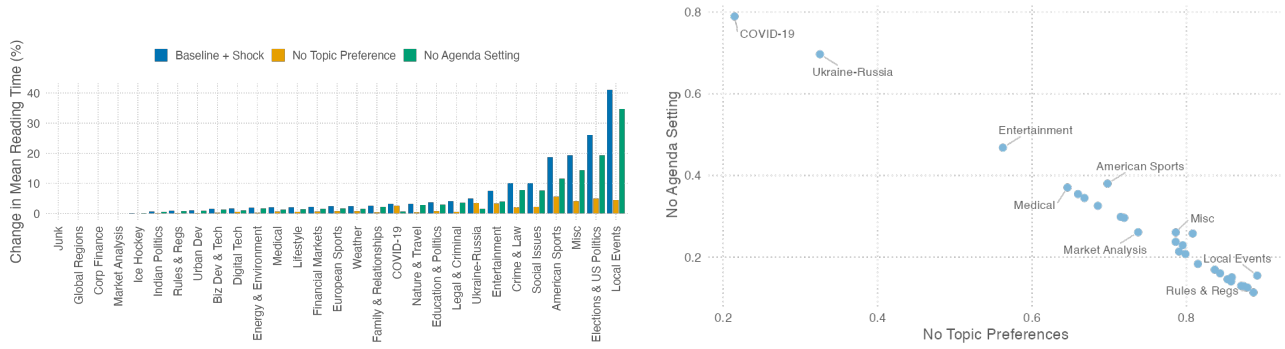
The distribution of topic preference coefficients $\beta_{i,k}$ are informative of the extent to which readers seek out news content about specific topics, but they cannot easily be interpreted on their own. In order to interpret the magnitude of topic preferences, we instead consider the following counterfactual exercise. For each topic k, we consider the demand response to a large event relating to that topic. To do this, we construct a synthetic shock to the average coverage of the topic, equal in magnitude to two standard deviations above the average in our sample. We choose two standard deviations because this corresponds to high-interest events: in the American Sports topic, two standard deviations above the average is roughly the magnitude of the 2022 Super Bowl spike visible in Figure 2a. Using our leave-one-out production model (the first stage for our IV), we predict how the production shares of each outlet would change in response to this shock on a typical day. We then simulate the reading

[20] $\hat{q}_{i,j,k,d}$ is the estimated conditional 99th quantile for observation i,j,k,d given the production share, instrument, outlet and topic indicators, and the non-news time allocation by i on d, which is estimated in step 2 of our estimation procedure. Note that $\hat{q}$ can be positive even for observations (j,k,d) where no reader in the cluster spent positive time.

[21] Note that our utility specification alone does not guarantee $T_{i,j,k,d} = 0$ when $p_{j,k,d} = 0$. For a topic k that outlet j never produces, $\dot{p}_{j,k,d} = 0 \; \forall \; d$ and $\hat{v}_{i,j,k,d} = \exp(\hat{\alpha}_{i,j} + \hat{\delta}^N_{i,d} + \hat{\varepsilon}_{i,j,k,d})$, allowing positive time allocation when $\hat{\varepsilon}_{i,j,k,d} > -\infty$.

time on each topic and outlet for each cluster of consumers i under the model as it is estimated, under a modified version in which $\beta_{i,k}$ is set to zero (no topic preferences), and under a modified version in which $\beta_{i,k}$ is unchanged but $\gamma_{i,j}$ is set to zero (no agenda setting). The $\beta_{i,k} = 0$ scenario models reader response in the absence of genuine interest in the topic, while the $\gamma_{i,j} = 0$ scenario models reader response in the absence of agenda setting by the entire online news industry.

As Figure 8a shows, the demand response to a large event in the Local Events topics is by far the strongest of all topics. Our model predicts that readers would increase their reading time by over 40% on average in response to such a shock. The vast majority of this effect is driven by topic preferences: while the demand response to a Local Events shock would decrease to about 35% without agenda-setting, it would plummet to less than 5% without topic preferences. A similar pattern holds for other topics with strong demand responses such as Elections & US Politics and American Sports. Other topics, like the Russia-Ukraine Conflict and COVID-19 exhibit a different pattern: not only is the overall demand response to a news shock in these topics weaker overall, but the relative importance of topic preferences and agenda setting is reversed. To see this more clearly, we plot the share of the baseline demand response that disappears when topic preferences are shut down versus when agenda setting is shut down for each topic in Figure 8b. Russia-Ukraine Conflict and COVID-19 are in the top-left corner, indicating the agenda-setting accounts for 70-80% of the baseline demand response. The Entertainment and Medical (non COVID-19) topics are in the middle, suggesting that topic preferences and agenda setting play similar roles in driving the demand response to news shocks in these topics. Local Events are in the bottom-right corner, indicating that topic preferences are the primary driver of the demand response to this news shock.



(a) Absolute change in mean reading time (%). *(b) Relative changes in mean reading time: $\beta_{i,k} = 0$ vs $\gamma_{i,j} = 0$.*

Figure 8: Increase in topic-level mean reading time following a large shock to the newsworthiness of each topic. The left panel shows the absolute change in mean reading time for each topic across the population under the baseline shock and under counterfactuals that shut down topic preferences or agenda setting. The right panel plots, for each topic, how much of the baseline response disappears when topic preferences are removed versus when agenda setting is removed.

6.2 How Important are Specific Outlets?

Our exercise in the previous section compared reader responses to a large event with high topical coverage market-wide. A natural question to follow is how important specific outlets are for driving the differences in news consumption across topics. To answer this, we consider a counterfactual exercise

in which we shut down the agenda setting power of one outlet at a time. For each outlet j, we set $\gamma_{i,j} = 0$ for all reader clusters i in j's audience (i.e., the set of clusters who read j in the baseline). We then simulate reading time on all topics and outlets for each cluster i over the course of our sample period under the observed production shares. Finally, we compute the difference in each i's attention to each topic by subtracting the total reading time on each topic (across days and outlets) in the baseline and in the counterfactual with $\gamma_{i,j} = 0$, and sum across cluster i in each outlet j's audience.

Appendix figure 23 plots the Euclidean distance between the vector of topic reading shares in the baseline and under the counterfactual that shuts down each outlet's agenda setting for the clusters in the outlet's audience. The distribution is sharply right-skewed: for most outlets the shift in the audience's news diet is negligible, while a handful of high-traffic national domains lie in the far right tail. Table 5 reports the top five outlets according to this notion of influence in the national sample and in three example markets (Atlanta, Boston, and Minneapolis). Consistent with the overall distribution of $\gamma_{i,j}$ parameters in Figure 7, the most influential outlets nationally are high-traffic domains like the New York Times, Fox News and CNN. Somewhat surprisingly, these national outlets also make up the majority of the top five in each of the local media markets as well; Atlanta has only one local outlet (WSB-TV) in the top five while Boston has none. This speaks to the nationalization of online news consumption. By contrast, ranks 3-5 in Minneapolis are three major local newspapers: the Pioneer Press (in St. Paul, MN), Star Tribune (in Minneapolis, MN) and Post-Bulletin (in Rochester, MN).

Rank	National	Atlanta	Boston	Minneapolis
1	New York Times	Fox News	NPR	New York Times
2	New York Post	CNN	New York Times	ESPN
3	CNN	WSB-TV	CNN	Pioneer Press
4	Fox News	New York Post	USA Today	Star Tribune
5	USA Today	USA Today	New York Post	Post-Bulletin

Table 5: Most influential news outlets, ranked within each market. Outlets are ordered by the magnitude of the impact on an outlet's audience's news diet from shutting down $\gamma_{i,j}$ for that outlet.

6.3 Case Study: Changes in Local Newspapers' Editorial Direction

In an interview with NPR following the sale of the Baltimore Sun, new co-owner Armstrong Williams declared that the Sun would change its editorial direction: "We want to return the paper to localism. You're [going to be] seeing more about crime. You're [going to be] seeing more about the mayor and City Hall." Using our model, we study how the changes in focus described by Williams would impact readership at the Baltimore Sun, and how similar changes might impact readership in other local markets. In each case, we simulate a change in editorial direction by shifting the production shares of a large local newspaper toward a pre-specified set of topics (e.g., crime) at the expense of all other coverage. We then ask how readership throughout the local media market is affected: To what degree would existing readers of the local newspaper view more of the up-weighted content? To what degree would existing readers stop reading the local newspaper at all? To what degree would non-readers begin to read the local newspaper, if at all?

For each counterfactual simulation, we consider the effect of a shift in production by one large regional newspaper on readership by consumers residing in the newspaper's DMA. We consider four types of editorial changes, motivated by recent trends in newspaper ownership: (1) Motivated by the Baltimore Sun purchase, we consider a 25% increase in crime (topic 4) reporting; (2) Motivated by the increase in buyouts by private equity firms (Ewens et al., 2023; Peterson and Dunaway, 2023) we consider a 25% decrease in reporting on urban development (topic 18) and education and local politics (topic 19); Motivated by the recent growth of nonprofit newsrooms (Benson, 2018) we consider a 25% shift toward the publication shares of ProPublica, a well-regarded nonprofit newspaper focused on local reporting; (4) Motivated by recent lay-offs and closures of local papers, we simulate a 25% shift toward the publication shares of USA Today, representing a typical national newspaper that relies heavily on syndicated content.

We begin our analysis by considering an increase in crime reporting by the Baltimore Sun. To do this, we apply the following procedure: for each day d in our sample, we take the observed publication share of crime by the Baltimore Sun $p_{\{\text{j = Baltimore Sun}\},\{k=4\},d}$ and inflate it by 25%. We then renormalize the publication shares of all other topics $p_{\{\text{j = Baltimore Sun}\},\{k\neq 4\},d}$, deflating them by a uniform scaling factor such that the total publication shares still sum to 100%. Holding the publication shares of all other outlets fixed as they are in the data, we then simulate the reading time on each topic and outlet for each cluster of consumers i in the Baltimore DMA under the modified marginal utility term from Equation (6):

$$v^c_{i,j,k,d} = \exp\left(\alpha_{i,j} + (\beta_{i,k} + \gamma_{i,j})\dot{p}^c_{j,k,d} + \delta^N_{i,d} + \varepsilon_{i,j,k,d}\right), \tag{10}$$

where $\dot{p}^c_{j,k,d} \equiv p^c_{j,k,d} - \overline{p}_{j,k}$ accounts for the modified production shares $p^c_{j,k,d}$ by the Baltimore Sun. Notably, we do not modify the overall outlet preference $\alpha_{i,j}$ or the year-long average topic-share $\overline{p}_{j,k}$ when computing counterfactual utility terms. Our reasoning for this is that our data do not allow us to extrapolate how public perception of outlets might change under a substantive change in editorial direction (or its precipitating ownership change). As such, our simulated readership changes can be interpreted as asking: how would readers' mix of content change if publication shares shifted but the overall reputation of the newspaper remained the same?

Figure 9 considers this exercise in each of the ten cities under our consideration: Atlanta (Atlanta Journal-Constitution), Baltimore (Baltimore Sun), Boston (Boston Globe), Chicago (Chicago Tribune), Dallas (Dallas Morning News), Houston (Houston Chronicle), Los Angeles (Los Angeles Times), Minneapolis (Minnesota Star Tribune), San Francisco (San Francisco Chronicle) and Tampa Bay (Tampa Bay Times).[22] For each city, we ask how local readership would respond to a 25% increase in crime reporting by a major local newspaper. To answer this question, we plot the aggregate percent change in time spent on the website of the focal newspaper in each city's local media market (DMA) (e.g., the Baltimore Sun in Baltimore), on the websites of other outlets read by consumers in the DMA, on other websites (not related to news) and on the outside option (offline time). Focusing on the Baltimore Sun, we find that Armstrong Williams has a point: our model predicts that readership

[22] We omit newspapers from with large national circulation like the New York Times and the Washington Post from this exercise as their local constituencies are only a subset of their markets.

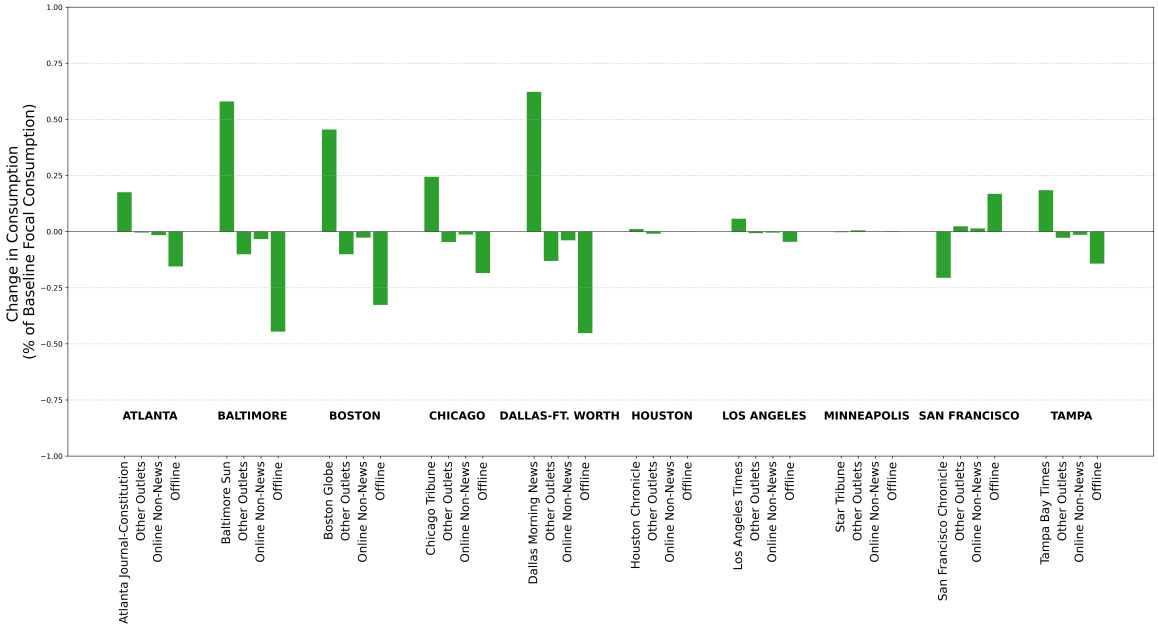


Figure 9: Case Study 1a across cities

of the Baltimore Sun would increase by **0.55%** in response to a 25% increase in crime reporting. Interestingly, the majority of substitution occurs not from other outlets or non-news websites but rather from the outside option.

Appendix figure 19 and Appendix figure 20 zoom into the aggregates in Figure 9 at the consumer level. Our model parameterization and cluster specification are rich enough to capture heterogeneity in substitution across consumers, as well as the specific topics consumers read when they substitute toward or away from the Baltimore Sun.

Coming back to Figure 9, we see that the preferences for crime reporting among Baltimore readers are not unique. Our model predicts that readers in Boston, Chicago, Dallas and to a lesser extent, Atlanta, Tampa and L.A. would respond in similar ways if their focal local newspapers increased crime reporting by 25%. However, we predict that readers in Houston and Minneapolis would have almost no response at all, and readers in San Francisco would respond in the opposite way—shifting to reading less of the SF Chronicle and spending more time on everything else. These differences in preference extend beyond taste for crime reporting. Figure 10 presents the aggregate impacts on readership for each of the other reporting shift counterfactuals in each city. A 25% decrease in reporting on urban development and local politics (as might be directed by private equity acquisitions) would modestly decrease readership of the local newspaper in Boston, L.A. and San Francisco, but modestly increase readership in Atlanta, Houston and Tampa. A 25% shift toward reporting that looks like ProPublica (topic-wise) would modestly increase readership of the local papers in Boston, L.A. and Minneapolis but decrease it everywhere else, with the biggest drop in Dallas. At the same time, all local newspapers would lose readers if they shifted their production shares to resemble a more generic national newspaper, proxied by USA Today. This suggests that while consumers in different media markets have different tastes for particular types of local reporting, they all value the overall bundle

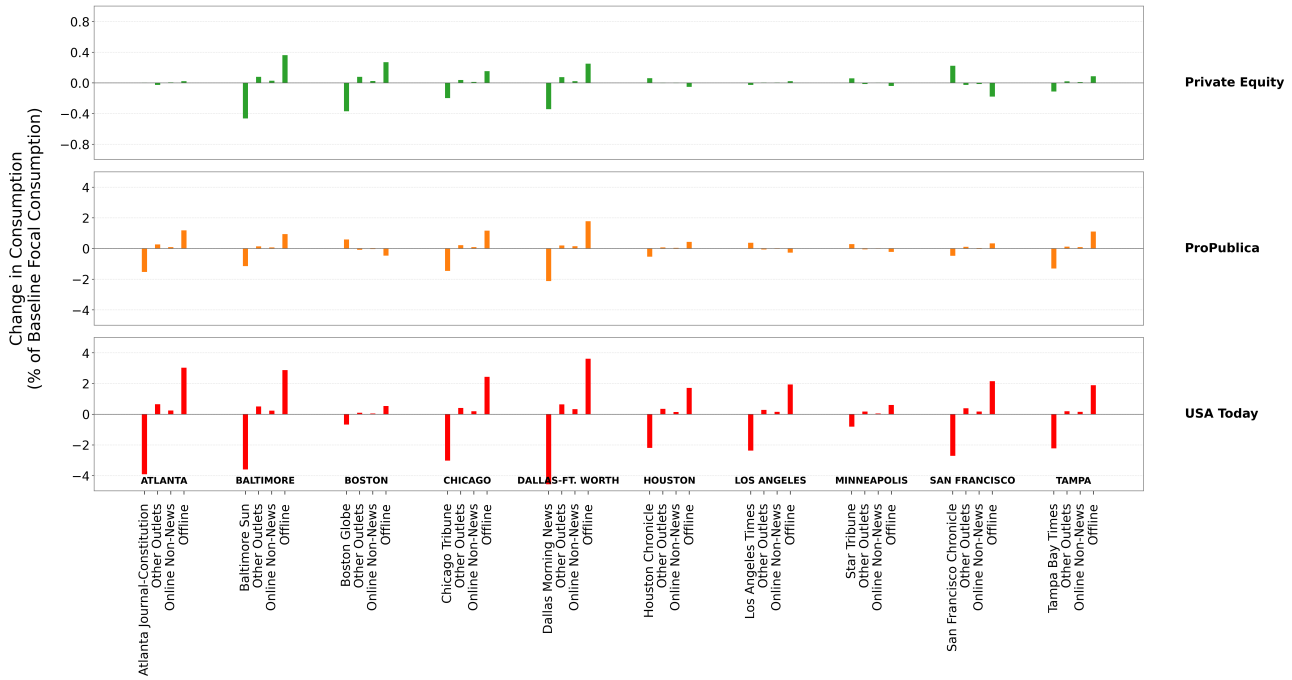


Figure 10: Effects of counterfactual reporting shifts in a major regional paper across DMAs

of topics that constitutes local news.

6.4 Case Study: Washington Post and CBS

Recent ownership changes at large national outlets like the Washington Post and CBS have also led to large shifts in editorial direction—often driven by significant turnover in staffing. In February 2026, *The Washington Post* fired a third of its newsroom staff, "shutter[ing] its sports desk," reducing the Metro section by over 70%, and "shrink[ing] dramatically" its international desk. In October 2025, Paramount Skydance, the parent company of CBS, acquired Bari Weiss's "The Free Press" and appointed Weiss be the editor-in-chief of CBS News. Upon the acquisition, Weiss vowed to remake her division "fit for purpose in the 21st Century" and restructured much of its programming (https://www.npr.org/2026/01/27/nx-s1-5689849/after-rocky-start-bari-weiss-to-cut-staff-add-commentators-at-cbs-news).

To assess the impact of compositional changes to newsroom staffing on readership across the country, we consider two counterfactual exercises. In the first, we use the Poisson regression model of production shares on topic-specialized staffing counts from Table 2 to simulate the impact of the staffing shakeup in the Washington Post. To account for the magnitude of the reported change in newsroom composition, we simulate a counterfactual in which *only* journalists with majority bylines on topics related to politics (that is, any byline in topics 4, 10, 15, 18, 19, 21, 22, 23, 25, 30, 31, 33) are retained (and all other journalists fired). The left panel of Figure 11 reports the changes in production shares that result. As expected, if the Washington Post only employed politics writers, it would publish more about political topics like Russia-Ukraine and national elections. However, the right panel of Figure 11 shows that this change would result in a relatively modest loss in readership: roughly 3.5%

of readership would shift away from the Washington Post, most of which would substitute to offline time.

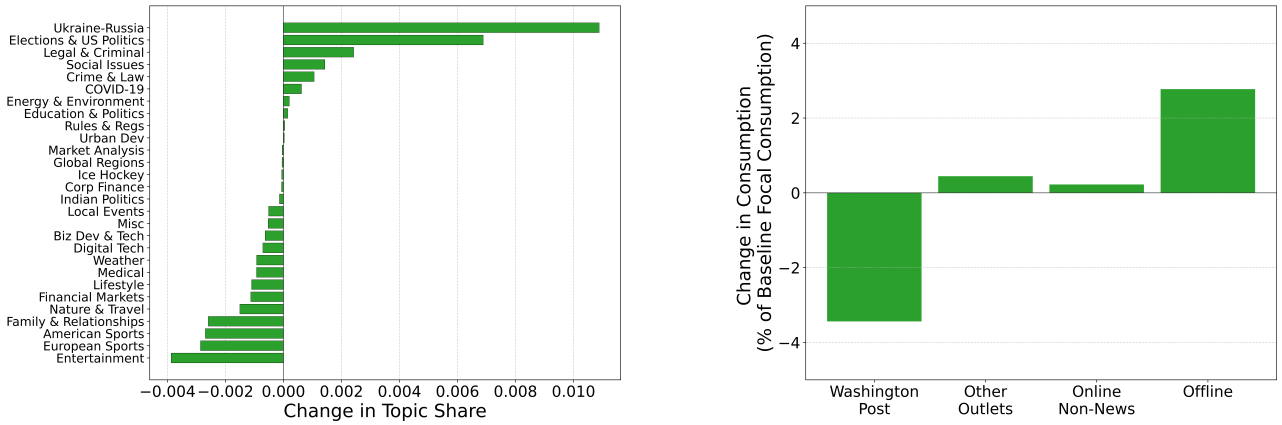


Figure 11: Washington Post: percent change in topic shares (left) and audience response (right) following staffing changes.

In the second counterfactual exercise, we consider the role of the newsroom's persistent perception among its audience. Although the changes in CBS staffing largely spanned the same mix of topic-level specializations as before Weiss was hired, critics argued that the reputation of the newsroom would shift. To assess this criticism, we simulate a counterfactual in which the outlet-level fixed effect $\gamma_{i,j}$ for each cluster of CBS readers i is set to the fixed effect for Fox News in the same cluster. Because some clusters of readers prefer CBS and others prefer Fox News in the data, this counterfactual increases CBS readership for some readers and decreases it for others. The left panel of Figure 12 reports the changes in consumption shares by decile of readers. The bottom decile of readers reduces their consumption of CBS by as much as the top decile creases it, but deciles in between lean more heavily toward increasing consumption. As a result, the aggregate impact is positive: CBS readership increases by over 6%, coming both from other news outlets and from offline time. This suggests that among readers who consume both CBS and Fox News, a plurality would prefer CBS's mix of reporting with Fox News's reputation.

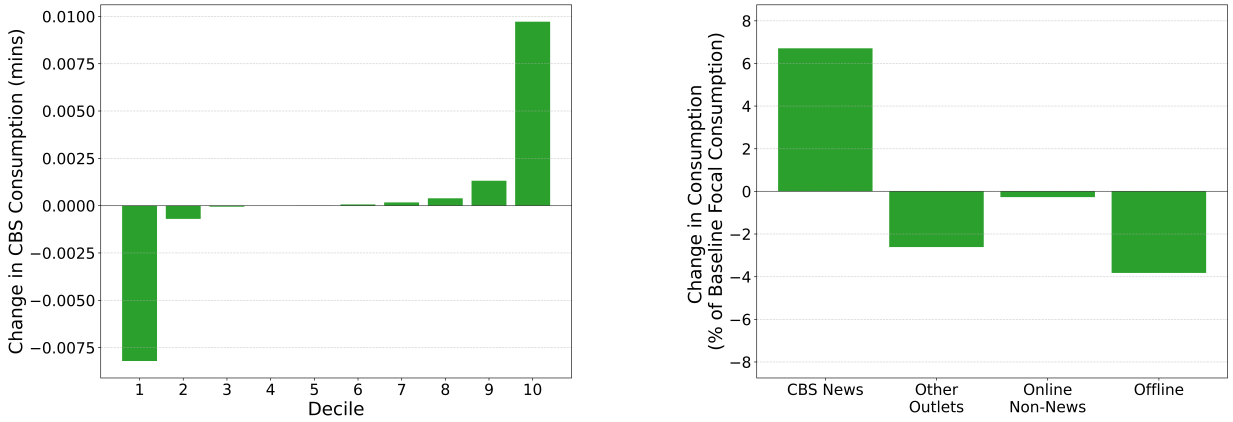


Figure 12: CBS: percent change in topic shares (left) and audience response (right) following staffing changes.

7 Conclusion

The 21st century news consumer looks down the barrel of a firehose of information. The internet makes available an array of sources and topics so vast that no reader could hope to consume even a small fraction of the news published on the internet each day, even if she devoted all her time to it and ignored other enticing digital options like short-form video or streaming TV platforms. In this environment of plenty, there remains a scarce resource: attention (Hayes, 2025). How readers allocate scarce attention across the enormous set of options available to them is the key question for anyone interested in the structure of media markets or in the political power of the few news "barons" (Peterson and Dunaway, 2023) who control increasingly large combinations of media properties.

We show that news outlets collectively have substantial power to steer reader attention. Editorial choices of what to produce and what to emphasize have measurable impact on what readers ultimately attend to. Our page-placement discontinuity design suggests that the typical story's readership can be doubled through placement choices. And our structural estimates show that absent such steering ability, most readers would have had essentially no interest in several topics that made up sizable shares of aggregate news coverage in 2022.

This power, however, is not evenly distributed. It is highly concentrated in a few prominent outlets, most of which are national in focus. The owners of these outlets have real power over the topics that the public attends to. Our studies of editorial changes at these outlets show that sizable shifts in topical focus are possible with only small loss of readership and minimal substitution. At the local level, agenda-setting power is more limited, and local news outlets are closer to attention-takers than attention-setters.

A policy implication of these findings is that regulators need to consider media power (Prat, 2018) when setting competition policy for the media sector. Our method provides a data-driven approach to estimating such power of the outlet level.

References

Allan, J., J. G. Carbonell, G. Doddington, J. Yamron, and Y. Yang (1998): "Topic detection and tracking pilot study final report," .

Angelucci, C. and J. Cagé (2019): "Newspapers in times of low advertising revenues," *American Economic Journal: Microeconomics*, 11, 319–64.

Angelucci, C., J. Cagé, and M. Sinkinson (2020): "Media competition and news diets," Tech. rep., National Bureau of Economic Research.

Aridor, G., T. Dekel, R. Jiménez Durán, R. Levy, and L. Song (2026): "Digital News Consumption: Evidence from Smartphone Content in the 2024 US Elections," .

Athey, S., M. Mobius, and J. Pal (2021): "The impact of aggregators on internet news consumption," Tech. rep., National Bureau of Economic Research.

Bang, M. and H. Sieg (2026): "Time Allocations and News Acquisition," .

Baum, M. A. (2003): "Soft news and political knowledge: Evidence of absence or absence of evidence?" *Political Communication*, 20, 173–190.

Benson, R. (2018): "Can foundations solve the journalism crisis?" *Journalism*, 19, 1059–1077.

Berry, S. and J. Waldfogel (2010): "Product quality and market size," *The Journal of Industrial Economics*, 58, 1–31.

Blundell, R. and J. L. Powell (2007): "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141, 65–83.

Broockman, D. E. and J. L. Kalla (2025): "Consuming Cross-Cutting Media Causes Learning and Moderates Attitudes: A Field Experiment with Fox News Viewers," *The Journal of Politics*, 87, 246–261.

Cagé, J. (2020): "Media competition, information provision and political participation: Evidence from French local newspapers and elections, 1944–2014," *Journal of Public Economics*, 185, 104077.

Cagé, J., N. Hervé, and M.-L. Viaud (2020): "The production of information in an online world," *The Review of Economic Studies*, 87, 2126–2164.

Calzada, J. and R. Gil (2020): "What do news aggregators do? Evidence from Google News in Spain and Germany," *Marketing Science*, 39, 134–167.

Chen, Y. and D. Y. Yang (2019): "The impact of media censorship: 1984 or brave new world?" *American Economic Review*, 109, 2294–2332.

Clemm von Hohenberg, B., S. Stier, A. S. Cardenal, A. M. Guess, E. Menchen-Trevino, and M. Wojcieszak (2024): "Analysis of Web Browsing Data: A Guide," *Social Science Computer Review*, 42, 1479–1504.

Crawford, G. S. and A. Yurukoglu (2012): "The welfare effects of bundling in multichannel television markets," *American Economic Review*, 102, 643–685.

DellaVigna, S., R. Enikolopov, V. Mironova, M. Petrova, and E. Zhuravskaya (2014): "Cross-border media and nationalism: Evidence from Serbian radio in Croatia," *American Economic Journal: Applied Economics*, 6, 103–132.

DellaVigna, S. and M. Gentzkow (2010): "Persuasion: Empirical Evidence," *Annual Review of Economics*, 2, 643–669.

DellaVigna, S. and E. Kaplan (2007): "The Fox News Effect: Media Bias and Voting," *The Quarterly Journal of Economics*, 122, pp. 1187–1234.

Dewey, J. (1927): "The public and its problems." .

Djourelova, M., R. Durante, and G. J. Martin (2025): "The Impact of Online Competition on Local Newspapers: Evidence from the Introduction of Craigslist," *Review of Economic Studies*, 92, 1738–1772.

Durante, R. and B. Knight (2012): "Partisan control, media bias, and viewer responses: Evidence from Berlusconi's Italy," *Journal of the European Economic Association*, 10, 451–481.

Durante, R., P. Pinotti, and A. Tesei (2019): "The political legacy of entertainment TV," *American Economic Review*, 109, 2497–2530.

Enikolopov, R., M. Petrova, and E. Zhuravskaya (2011): "Media and political persuasion: Evidence from Russia," *The American Economic Review*, 101, 3253–3285.

Ewens, M., A. Gupta, and S. T. Howell (2023): "Local journalism under private equity ownership," Tech. rep., National Bureau of Economic Research.

Flaxman, S., S. Goel, and J. M. Rao (2016): "Filter bubbles, echo chambers, and online news consumption," *Public opinion quarterly*, 80, 298–320.

Garz, M. and G. J. Martin (2021): "Media influence on vote choices: Unemployment news and incumbents' electoral prospects," *American Journal of Political Science*, 65, 278–293.

Gentzkow, M. and J. M. Shapiro (2010): "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78, 35–71.

——— (2015): *Ideology and Online News*, University of Chicago Press.

Gentzkow, M., J. M. Shapiro, and M. Sinkinson (2011): "The Effect of Newspaper Entry and Exit on Electoral Politics," *American Economic Review*, 101, 2980–3018.

——— (2014): "Competition and ideological diversity: Historical evidence from us newspapers," *American Economic Review*, 104, 3073–3114.

George, L. and J. Waldfogel (2003): "Who affects whom in daily newspaper markets?" *Journal of Political Economy*, 111, 765–784.

George, L. M. and J. Waldfogel (2006): "The New York Times and the market for local newspapers," *American Economic Review*, 96, 435–447.

Hamilton, J. (2004): *All the news that's fit to sell: How the market transforms information into news*, Princeton University Press.

Hayes, C. (2025): *The Sirens' Call: How Attention Became the World's Most Endangered Resource*, Random House.

Hayes, D. and J. L. Lawless (2017): "The Decline of Local News and Its Effects: New Evidence from Longitudinal Data," *The Journal of Politics*, 80, 000–000.

——— (2021): *News hole: The demise of local journalism and political engagement*, Cambridge University Press.

Hornik, K., I. Feinerer, M. Kober, and C. Buchta (2012): "Spherical k-Means Clustering," *Journal of Statistical Software*, 50, 1–22.

Khan, S. and J. L. Powell (2001): "Two-step estimation of semiparametric censored regression models," *Journal of Econometrics*, 103, 73–110.

King, G., B. Schneer, and A. White (2017): "How the news media activate public expression and influence national agendas," *Science*, 358, 776–780.

Knight, B. and A. Tribin (2019): "The limits of propaganda: Evidence from chavez's venezuela," *Journal of the European Economic Association*, 17, 567–605.

Levy, R. (2021): "Social media, news consumption, and polarization: Evidence from a field experiment," *American economic review*, 111, 831–70.

Lippmann, W. (1925): *The phantom public*, Transaction Publishers.

Martin, G. J., J. McCrain, N. Mastrorocco, and A. Ornaghi (2026a): "Media Consolidation," .

Martin, G. J., S. Vasserman, and C. Pfiffer (2026b): "What do news readers want?" Tech. Rep. 35289, National Bureau of Economic Research.

Martin, G. J. and A. Yurukoglu (2017): "Bias in Cable News: Persuasion and Polarization," *American Economic Review*, 107, 2565–2599.

McCombs, M. E. and D. L. Shaw (1972): "The Agenda-Setting Function of Mass Media," *Public Opinion Quarterly*, 36, 176–187.

McInnes, L., J. Healy, and J. Melville (2018): "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*.

Mullainathan, S. and A. Shleifer (2005): "The Market for News," *American Economic Review*, 95, 1031–1053.

Peterson, E. and J. Dunaway (2023): "The New News Barons: Investment Ownership Reduces Newspaper Reporting Capacity," *The ANNALS of the American Academy of Political and Social Science*, 707, 74–89.

Prat, A. (2018): "Media power," *Journal of Political Economy*, 126, 1747–1783.

Roberts, M. E., B. M. Stewart, D. Tingley, and E. M. Airoldi (2013): "The structural topic model and applied social science," in *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*, Harrahs and Harveys, Lake Tahoe, vol. 4, 1–20.

Sen, A. and P. Yildirim (2015): "Clicks bias in editorial decisions: How does popularity shape online news coverage?" *Available at SSRN 2619440*.

Simonov, A. and J. Rao (2022): "Demand for Online News under Government Control: Evidence from Russia," *Journal of Political Economy*, 130, 000–000.

Snyder, J. M. and D. Strömberg (2010): "Press Coverage and Political Accountability," *Journal of Political Economy*, 118, 355–408.

Strömberg, D. (2004): "Mass Media Competition, Political Competition, and Public Policy," *The Review of Economic Studies*, 71, 265–284.

Appendices

A Production data processing

A.1 Outlet classification

We classify each outlet in our analysis along two dimensions that we use throughout the paper.

Geographic scope. Each outlet is assigned to one of three categories. *National* outlets target a US-wide audience (e.g., the New York Times, Fox News). *Local* (and regional) outlets primarily serve a particular state or metropolitan area (e.g., the Baltimore Sun); each local outlet also carries an associated US state code. *International* outlets are based outside the United States (e.g., the BBC).

Topic specialization. Each outlet is labeled as either *general-interest*, covering the full range of news topics, or *specialized*, focusing primarily on a named area such as sports, business, politics, health, or entertainment. Specialty labels come from the same whitelist and are checked against each outlet's observed topic production shares in GDELT.

A.2 Classification of News Topics

We classify news articles into topics using the structural topic model (STM) of Roberts et al. (2013), which extends latent Dirichlet allocation by letting each document's topic mixture depend on document-level covariates. We include two prevalence covariates: a factor for the outlet (at the levels defined below) and a factor for the day of the year. The outlet covariate lets topic prevalence vary across outlets that specialize in different kinds of coverage; the day covariate lets it vary as industry-wide events ebb and flow.

Outlet set. We train the model on articles from the Peakmetrics archive (Section 3). Starting from a whitelist of US and international English-language news domains, we drop outlets with less than 50% English-language content or fewer than 100 articles published in 2022.

Outlet labeling. For the prevalence covariate to be informative at the outlet level, each level needs enough observations. We therefore give an outlet its own level only when our consumption panel records at least 30 panelists who spent five or more minutes reading it; this threshold produces 416 "own-label" outlets. Remaining US outlets are pooled by geographic classification and state (see Appendix A.1) into "generic national" and "generic <state>" levels, and remaining international outlets are pooled into a single "generic international" level. International outlets with fewer than ten qualifying panelists are dropped from the training corpus.

Training corpus. For each outlet, we take a uniform random 4% sample of its 2022 articles. We lowercase the article text, strip punctuation, numbers, and newlines, tokenize, remove English stop-words and URL-fragment tokens (e.g., those beginning with `www`), and trim the vocabulary to tokens appearing in between 0.01% and 40% of documents.

Choice of K. We estimate the STM for $K \in \{15, 20, 25, 30, 35, 40, 45, 50\}$ topics and select K using an in-article link validation. Articles often cite other articles through embedded hyperlinks; when the cited URL is itself in our corpus, we can check whether the fitted STM assigns the same topic to the citing and cited articles. A topic model that captures real content should make this coincidence more common than chance. For a sample of 103 outlets and for each candidate K, we compute the

share of within-corpus citation pairs in which the citing and cited articles receive the same topic.[23] As a placebo baseline, we randomly permute the cited articles' topic labels within each outlet and recompute the match rate.

Appendix figure 13 plots the observed and placebo match probabilities across K (left panel), and their difference (right panel). The observed probability exceeds the placebo by a factor of roughly 1.5 to 1.7 at every K, confirming that the model is picking up content-level similarity rather than noise. The observed–placebo gap is relatively flat — between 0.124 and 0.133 — across the candidate values, with its highest value at $K = 35$.

Because the link-validation gap is relatively flat across candidate K, we also examine the fitted topics directly, focusing on two requirements for our demand application: each topic should correspond to an identifiable subject of news so that consumer preference is interpretable, and topics should not have mechanically correlated production shares that would impair identification. Two failure modes distinguish $K = 35$ from neighboring values. At $K \leq 30$, substantively distinct subjects remain merged: COVID-19 and inflation share a single topic at $K = 25$, and weather is bundled with event listings at $K = 30$. At $K \geq 40$, US politics, a single coherent topic at $K = 35$, fragments at $K = 40$ into an electoral-mechanics topic (top FREX words *ballots, voters, gubernatorial, fetterman*) and a presidential-coverage topic (top FREX words *capitol, pelosi, maralago, insurrection*). A separate topic also emerges at $K = 40$ whose top FREX words *dad, mom, shes, ive, knew, friend, im, felt, mum, tears* reflect first-person narrative voice rather than subject matter. We therefore select $K = 35$ as the smallest model in which every major news subject is represented by exactly one coherent topic and the largest model before topics begin to split into very related and along stylistic lines.

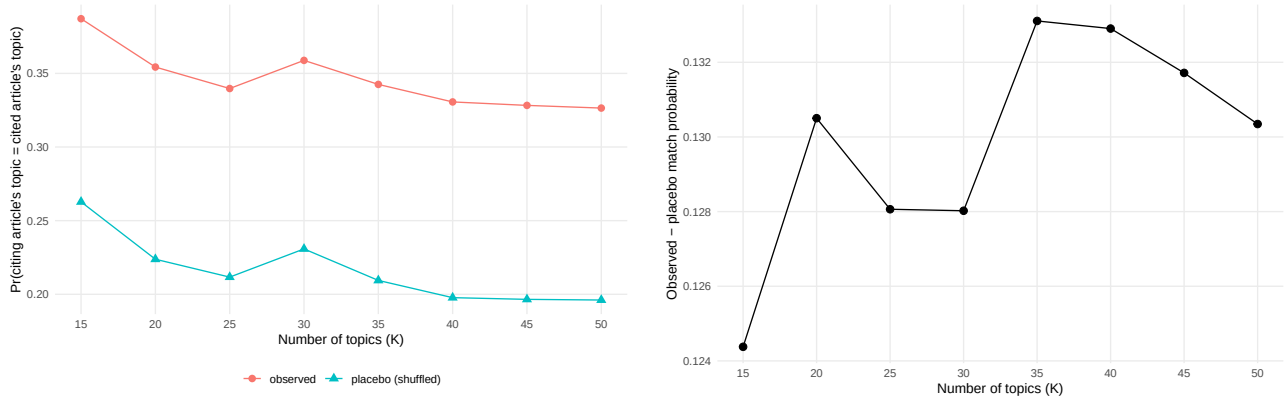
Given $K = 35$, six estimated topics capture non-substantive content (advertisements, navigation text, and other artifacts) and are merged into a single "junk" category. The remaining 29 substantive topics are described in Section 3.4.

A.3 Assigning articles to topics

Our article content data come from two sources: PeakMetrics, which has full text of articles published by our set of news outlets, and GDELT, which reports the position of links on the home page of each of our outlets, at an hourly frequency. We use the union of URLs identified in these two datasets as the starting point to define our universe of articles published.

GDELT tracks *all* links that appear on the homepage, including permanent navigation links, advertisements, and so on, in addition to articles. We filter the GDELT URLs by first removing any URL pointing to a domain other than the outlet domain (in order to remove advertisements and other links to content not published by the outlet). Second, we remove any URL that appears on the home page on any two occasions that are at least 10 days apart. This second criterion removes permanent navigation links and other long-lasting page features that are not articles.

[23] The sample is drawn in two steps. We first select 50 outlets with probability proportional to the number of panelists who spent five or more minutes reading them. We then add one outlet from each "generic" label group that contains multiple outlets, so that every level of the prevalence covariate is represented. After deduplication, the two steps together yield 103 outlets.



(a) Observed vs. placebo match probability

(b) Observed minus placebo

Figure 13: In-article link validation of the STM across candidate values of K. For each K and each of 103 *sampled outlets, we compute the share of within-corpus citation pairs (article → linked article) in which the citing and cited articles receive the same STM topic. The placebo series randomly permutes cited articles' topic labels within outlet. Points are unweighted averages across the* 103 *outlets.*

We then use our fitted topic model to assign each article-URL a to a single topic $k(a)$. We use either the frequencies of words appearing in the full-text (for articles that appear in PeakMetrics) or in the URL "slug" (for articles that appear in GDELT but not in PeakMetrics) to predict the article's topic from our topic model.[24]

The topic model returns a vector of the article's predicted weights for each topic. To assign the article to a single topic we first drop our set of "junk" topics[25] and also any weight lower than 0.1. We randomly assign the article to the surviving topics with probabilities equal to the surviving topics' weights, re-normalized to sum to 1. Any URL with no surviving topics, i.e. no non-junk topic weights that exceed 0.1, is assigned to topic 0, our consolidated missing-topic category.

A.4 Production shares

Given our assignment of articles to topics, we define an aggregate outlet-topic-day level measure p_{jkt}, denoting outlet j's production share of topic k on day t. We compute this daily production share using a time- and position-weighted average that makes use of the high-frequency GDELT data.

The GDELT front-page data reports, for each article URL a appearing on the front page of outlet j on day d and hour h, the link's relative position among all links on the front page at that day and hour, $l_{j,a,d,h}$, expressed as a percentage from 0 to 1. When an article a's URL does not appear on the home page of j on day d and hour h, we set $l_{j,a,d,h} = \infty$.

We aggregate the link-by-hour-level position data into a production share as follows:

[24] The "slug" is the part of the URL excluding the domain and any file extensions or parameters, e.g. `2026/02/13/nyregion/epstein-donations-democrats-hochul-jeffries` in the URL `https://www.nytimes.com/2026/02/13/nyregion/epstein-donations-democrats-hochul-jeffries.html`. In this example the resulting stemmed words extracted from the slug and used in prediction from the topic model would be `nyregion, epstein, donat, hochul, jeffri`.

[25] See Appendix A.2.

$$p_{jkd} \equiv \frac{\sum_{a:\, k(a)=k} \sum_h e^{-l_{j,a,d,h}}}{\sum_a \sum_h e^{-l_{j,a,d,h}}} \tag{11}$$

The production share defined in this way weights articles more heavily when they appear towards the top of the page (l close to 0), with weight that diminishes at an exponential rate. It also up-weights articles that remain on the front page for more hours on day d (because once they drop off, $l_{j,a,d,h} = \infty$ and the weight $e^{-l_{j,a,d,h}}$ is zero for hours in which the article is not on the front page).

Because weekend days have much lower publication volume and web traffic, we collapse weekends together with the preceding Friday to form a single synthetic day observation, with production shares averaged across the three days. This aggregation smooths over lower weekend publishing volumes.

A.5 Instrumenting production share

We construct a shift-share instrument for the topic composition of each outlet's production using a Poisson regression that interacts the outlet's long-term topic specialization (the "shares") with day-to-day variation in what *other* outlets are producing (the "shifts").

We first construct a leave-one-out instrument $\bar{p}_{-j,kd}$: the average production share of topic k on day d across all outlets *other than* j. This captures aggregate news supply shocks—e.g., a major election or natural disaster raising coverage of a topic industry-wide—while excluding outlet j's own editorial choices. We then estimate a Poisson regression that specifies the conditional mean of p_{jkd} as:

$$E\left[p_{jkd} \mid \bar{p}_{-j,kd},\, \alpha_{jk}\right] = \exp\!\big(\alpha_{jk} + \beta\, \bar{p}_{-j,kd}\big), \tag{12}$$

where α_{jk} is an outlet-by-topic fixed effect. The outlet-by-topic fixed effects absorb each outlet's persistent specialization in a topic (e.g., ESPN always devotes most of its reporting to sports). The exponential form means that the model is multiplicative: outlets with greater specialization in a topic (higher α_{jk}) will respond more strongly to increases in aggregate interest in the topic, compared to outlets with lower specialization. The fitted values $\hat{p}_{jkd} = \exp(\hat{\alpha}_{jk} + \hat{\beta}\, \bar{p}_{-j,kd})$ serve as the instrument $z_{j,k,d}$ in the IV regressions in Section 4 and as the basis for the control variable in the demand model in Section 5.

Shift-share interpretation. The Poisson first stage gives the instrument a shift-share interpretation. To see this, take a first-order Taylor expansion of $z_{j,k,d}$ around the time-series mean $\bar{\bar{p}}_{-j,k} \equiv \frac{1}{D}\sum_d \bar{p}_{-j,k,d}$:

$$\begin{aligned} z_{j,k,d} &= \exp(\hat{\alpha}_{jk} + \hat{\beta}\, \bar{p}_{-j,kd}) \\ &\approx \exp(\hat{\alpha}_{jk} + \hat{\beta}\, \bar{\bar{p}}_{-j,k}) \cdot \left[1 + \hat{\beta}\, (\bar{p}_{-j,kd} - \bar{\bar{p}}_{-j,k})\right] \\ &\approx \bar{p}_{j,k} \cdot \left[1 + \hat{\beta}\, (\bar{p}_{-j,kd} - \bar{\bar{p}}_{-j,k})\right], \end{aligned}$$

where the last line uses the fact that the Poisson conditional mean evaluated at the average regressor value approximates outlet j's long-run average production share $\bar{p}_{j,k}$. The time-varying component of

the instrument is therefore $\bar{p}_{j,k} \cdot \hat{\beta} \cdot (\bar{p}_{-j,kd} - \bar{\bar{p}}_{-j,k})$: the outlet's long-run specialization (the "share") multiplied by the deviation in other outlets' coverage on that day relative to its cross-day mean (the "shift"). The exponential link ensures that fitted values are non-negative and that the interaction is inherently multiplicative, so that the same aggregate shock produces a proportionally larger predicted response for more specialized outlets.

Production function interpretation The idea of the leave-one-out instrument is to capture the technological constraint by which outlets specialized in specific topics respond differently to newsworthy events related to their specialty. In order words, the instrument estimates a simple production function that maps a measure of topic-day level newsworthiness to outlet-specific production shares. To formalize this logic, we posit the following model of outlet-topic production.

On each day d, outlet j chooses its production for topic k (measured in log shares) as a function of baseline outlet-topic expertise and the amount of newsworthy content available according to:

$$p_{jkd} = \alpha_{jk} + \beta_{jk}\overline{\rho}_{kd} + u_{jkd}. \tag{13}$$

The outlet-topic coefficients α_{jk} and β_{jk} capture the notion of specialization: a higher α_{jk} indicates that j generally publishes more on topic k, while a higher β_{jk} indicates that j covers k disproportionately more on days in which k is more newsworthy. The residual term u_{jkd} captures idiosyncratic shifts in production that are independent of newsworthiness or other outlets' production.

The extent of newsworthiness $\overline{\rho}_{kd}$ is directly related to demand. For instance, in following with Equation (2), we might define the demand that an individual i has for reading about topic k on outlet j and day d as:

$$c_{ijkd} = \theta p_{jkd} + \gamma \rho_{ikd} + \lambda_{ij} + \mu_{jk} + \eta_{kd} + \varepsilon_{ijkd}, \tag{14}$$

where ρ_{ikd} is a generic consumer-topic-day specific interest shock separate from the outlet-topic-day production share p_{jkd}, λ_{ij} is a consumer-outlet fixed effect and the other parameters match Equation (2). The aggregate measure of newsworthiness $\overline{\rho}_{kd} \equiv \int \rho_{ikd} dF_i$ corresponds to the average consumer-topic interest on a given day in the population.

As Equation (14) makes clear, regressing observed outlet-topic consumption at the individual or population level on production shares would suffer from an endogeneity problem: $\overline{\rho_{kd}}$ enters both supply and demand. However, under the production share model in Equation (13), the leave-one-out production share can serve as a proxy for $\overline{\rho}_{kd}$ that can resolve the endogeneity issue.

To see this, denote the leave-one-out production share:

$$\overline{p}_{-jkd} \equiv \frac{1}{J-1} \sum_{\ell \neq j} p_{\ell kd} = \overline{\alpha}_{-jk} + \overline{\beta}_{-jk}\overline{\rho}_{kd} + \overline{u}_{-jkd}, \tag{15}$$

where $\overline{\alpha}_{-jk}$, $\overline{\beta}_{-jk}$ and $\overline{u}_{-jkd}$ denote averages over outlets $\ell \neq j$. For each outlet j, we can project p_{jkd}

on $\overline{p}_{-jkd}$ using OLS to produce the linear analog to our leave-one-out instrument:

$$z^{\text{ols}}_{jkd} = \hat{a}_{jk} + \hat{b}_{jk}\overline{p}_{-jkd}, \tag{16}$$

where $\hat{a}_{jk}$ and $\hat{b}_{jk}$ are the OLS coefficient estimates from fitting p_{jkd} to a linear function of $\overline{p}_{-jkd}$ within outlet-topic pairs. As u_{jkd} is mean-independent of $\overline{\rho_{kd}}$ and $\overline{u}_{-jkd}$ by assumption, we can then rewrite the linear leave-one-out instrument z^{ols}_{jkd} as follows:

$$z^{\text{ols}}_{jkd} = \underbrace{(\hat{a}_{jk} + \hat{b}_{jk}\overline{\alpha}_{-jk})}_{\text{average specialized production}} + \underbrace{\hat{b}_{jk}\overline{\rho}_{kd}}_{\text{specialized news response}} + \underbrace{\hat{b}_{jk}\overline{u}_{-jkd}.}_{\text{interacted market residual}} \tag{17}$$

It is clear from Equation (17) that z^{ols}_{jkd} captures day-to-day movements in topic newsworthiness. For it to also satisfy the exclusion restriction, it must be the case that $\mathrm{E}[z^{\text{ols}}_{jkd} \cdot c_{ijkd} | i, j, k, \overline{\rho}_{kd}, \alpha, \beta] = 0$. Residualizing out the fixed effects, this condition boils down to:

$$\mathrm{E}[z^{\text{ols}}_{jkd} \cdot (\gamma\rho_{ikd} + \varepsilon_{ijkd}) | \overline{\rho}_{kd}, \alpha, \beta] = 0. \tag{18}$$

If there is all consumers are relevant for all outlets (e.g. the aggregate newsworthiness measure used for all outlets' production $\overline{\rho}_{kd}$ integrates over the full population of consumers) then $\mathrm{E}[\gamma\rho_{ikd} | \overline{\rho}_{kd}, \alpha, \beta] = \overline{\rho}_{kd}$. The exclusion restriction then follows from the definition of ε_{ijkd}. If outlets are instead selective with respect to which consumers they consider in their production, then an additional assumption is necessary for the exclusion restriction to hold. In this case, it must be that a given outlet's specialized news response must be orthogonal from the specialization of its audience.

Alternative outlet sets. The exclusion restriction can fail if the outlets that form $\overline{p}_{-jkd}$ serve the same readers as outlet j, so that their coverage responds to the idiosyncratic interest of j's own audience rather than to aggregate newsworthiness. To assess this, we rebuild the instrument using a weighted leave-one-out average

$$\overline{p}^{w}_{-jkd} \equiv \frac{\sum_{\ell \neq j} w_{j\ell}\, p_{\ell kd}}{\sum_{\ell \neq j} w_{j\ell}}, \tag{19}$$

where the sums run over outlets publishing on day d. The baseline instrument sets $w_{j\ell} = 1$ for every other outlet, so that $\overline{p}^{w}_{-jkd}$ is the simple mean. We consider four alternatives. Let $g(j)$ denote outlet j's specialization group (the outlet types in Appendix A.1, with general-interest outlets split into national, local, and international), and let $n_{j\ell}$ be the number of Comscore panelists who read both outlets j and ℓ at any point in 2022.

(a) *Same specialization group.* Only outlets in j's own group enter the average, each with equal weight: $w_{j\ell} = 1$ if $g(\ell) = g(j)$ and 0 otherwise.

(b) *Other specialization groups.* Only outlets outside j's group enter, each with equal weight: $w_{j\ell} = 1$ if $g(\ell) \neq g(j)$ and 0 otherwise.

(c) *Weighted by audience overlap.* Every other outlet enters, weighted by the number of readers it shares with j: $w_{j\ell} = n_{j\ell}$. Outlets that share no readers with j receive no weight.[26]

(d) *Weighted by inverse audience overlap.* Every other outlet enters, with weight decreasing in the number of shared readers: $w_{j\ell} = 1/(n_{j\ell} + 1)$. Outlets that share no readers with j receive the largest weight of one; an outlet sharing one reader receives one half, and so on.

The audience weights are computed once from the full year of Comscore data, so they are constant over time and do not themselves contribute day-level variation to the instrument. For each outlet set we re-estimate the Poisson projection with outlet-by-topic fixed effects and use its fitted value as the instrument.

Outlet sets (b) and (d) are the conservative constructions: they measure newsworthiness from outlets that do not share j's specialization, or that share little of its audience, so their coverage cannot be responding to the interests of j's readers. Outlet sets (a) and (c) do the opposite, concentrating the average among outlets most likely to serve the same readers as j. Comparing IV estimates across these constructions therefore brackets the extent to which audience-specific demand contaminates the baseline instrument.

A.6 Robustness of the instrument to the choice of outlets

Table 6 reports the projection of an outlet's production share on each version of the leave-one-out average, alongside the correlation between the alternative average and the baseline one. Every outlet set delivers a strong projection: the Poisson coefficient ranges from 10 (same specialization group) to 17.3 (other specialization groups), with t-statistics above 20 throughout. Two features of the table guide the interpretation of the IV results. First, the inverse-overlap average is nearly identical to the baseline (correlation 0.993). With thousands of outlets in the panel, most pairs share no readers at all, so down-weighting the few that do barely moves the average; the baseline instrument is therefore already, in effect, built from outlets that do not share j's audience. Second, the other-group and overlap-weighted averages differ but are also strongly correlated with the baseline (correlations of 0.587 and 0.614). The overlap-weighted construction is the informative diagnostic: it concentrates the shift among the outlets most exposed to j's audience, and an IV estimate under this instrument that matches the baseline indicates that audience-specific demand plays little role in the baseline shift.

[26] Because the weights are normalized in Equation (19), using the count of shared readers is equivalent to using the share of all news-reading panelists who read both outlets.

Peer set for leave-one-out average	Linear	Poisson	Corr. with baseline	N
All other outlets (baseline)	0.969 (0.003)	15.432 (0.740)	1.000	21,509,980
Same specialization group	0.868 (0.009)	10.018 (0.185)	0.803	21,509,980
Other specialization groups	0.694 (0.030)	17.271 (0.523)	0.587	21,509,980
Weighted by audience overlap	0.609 (0.017)	10.580 (0.277)	0.614	21,244,637
Weighted by inverse audience overlap	0.927 (0.008)	14.941 (0.560)	0.993	21,292,341

Table 6: Instrument construction under alternative peer sets.

Notes: Each row projects outlet j's production share of topic k on day d on a leave-one-out average of other outlets' shares, with outlet×topic fixed effects, using a different set of peer outlets to form the average. Linear and Poisson columns report the coefficient on the leave-one-out average (standard errors clustered by day in parentheses). Corr. with baseline is the correlation between the alternative leave-one-out average and the baseline one. N is the Poisson estimation sample; outlets without a specialization group are excluded from the group-based rows, and outlets without Comscore readers from the overlap-weighted rows.

Table 7 re-estimates the pooled specification in Table 4 with each alternative instrument. All rows use the common sample in which every instrument is defined, so the baseline row is re-estimated on that sample; for national and international outlets the sample is unchanged, and for local outlets it is smaller by about 0.3%. The first stage is strong under every construction, with F-statistics between 111 and 1,675, so the comparison across rows is not confounded by weak instruments.

For national and international outlets, the estimates of θ are stable across peer sets. For national outlets, the estimate ranges from 0.99 (same specialization group) to 1.31 (other specialization groups), compared with 1.29 for the baseline; for international outlets, from 2.49 to 2.99, compared with 2.96. In both cases every alternative lies within about one and a half standard errors of the baseline. The two conservative constructions, which build the shift from outlets outside j's specialization or from outlets sharing little of its audience, deliver estimates of 1.31 and 1.23 for national outlets and 2.85 and 2.99 for international outlets, essentially identical to the baseline. The two exposed constructions, which concentrate the average among same-specialty and audience-sharing outlets, give somewhat *smaller* estimates (0.99 and 1.15 for national outlets, 2.49 and 2.84 for international outlets). If the baseline shift were contaminated by demand shocks specific to j's audience, the exposed constructions would be expected to inflate the estimate; the opposite pattern indicates that this contamination is not driving the baseline result.

For local outlets, the baseline and the two conservative constructions all yield small and imprecisely estimated effects (0.18, 0.30, and 0.07), consistent with the main text. The one construction that behaves differently is the audience-overlap weighting, which yields 0.74 (s.e. 0.24). For a local outlet, the peers that share its readers are predominantly other outlets in the same market, so this average tracks local rather than national newsworthiness, and the resulting instrument captures shifts in local coverage that the nationwide average largely misses. This is also the construction most exposed to demand shocks shared within a local audience, so we do not rely on it; we note only that the conservative constructions leave the main-text conclusion for local outlets unchanged. We conclude

that the response of consumption to instrumented coverage for national and international outlets is not driven by peer outlets that share j's audience or its specialization.

Peer set for leave-one-out average	National	Local	International
All other outlets (baseline)	1.285	0.181	2.961
	(0.169)	(0.249)	(0.604)
First-stage F	580.9	475.6	514.6
Same specialization group	0.986	0.171	2.489
	(0.136)	(0.235)	(0.510)
First-stage F	896.9	631.8	426.7
Other specialization groups	1.309	0.299	2.853
	(0.174)	(0.269)	(0.585)
First-stage F	620.4	111.1	443.9
Weighted by audience overlap	1.149	0.741	2.844
	(0.142)	(0.242)	(0.494)
First-stage F	1675.2	339.2	451.1
Weighted by inverse audience overlap	1.228	0.069	2.993
	(0.207)	(0.263)	(0.691)
First-stage F	677.2	496.7	338.2
N	3,427,788	17,309,628	645,456

Table 7: Effect of outlet production share on consumption (IV) under alternative peer sets.

Notes: IV estimates of θ from equation (2), pooling across all topics, with the instrument built from the leave-one-out average over the indicated set of peer outlets (see Appendix A.5). Each column restricts to outlets of the indicated geographic scope. All rows use the common sample in which every instrument is defined, so the baseline row is re-estimated on that sample and may differ from Table 4. Standard errors in parentheses, clustered by time period.

A.7 Newsroom staffing across topics

In order to estimate the impact of beat-specific staffing on production shares in Section 4 and Section 6, we construct a monthly panel of author staffing using PeakMetrics and LinkedIn data.

Journalist specialization panel. We extract author names (from both single- and multi-author fields) from the PeakMetrics metadata of each article and split co-authored articles so that each listed author receives credit. Merging these records with the article-level STM topic assignments, we count, for each author–month pair, the number of articles written on each STM topic, along with a total article count.

Journalist employment panel. We construct a monthly panel of journalist employment by linking PeakMetrics authors to LinkedIn employment histories, identifying each author's outlet affiliation by month. The goal is to observe, for each author and month, which outlet the author was affiliated

with and how article production was distributed across STM topics in the preceding month. We then aggregate these author-level measures to the outlet level.

mple construction. We begin from raw LinkedIn profile data and apply a series of filters to obtain a sample of authors employed at outlets in our Comscore list. We retain accounts located in the United States with both a first and last name recorded, requiring that the combined name length exceed five characters. We restrict attention to employment spells active in 2022 and keep only authors whose full names appear in the PeakMetrics author production panel.

In company pages. To identify each news outlet's LinkedIn presence, we combine three approaches. First, we match outlets to LinkedIn company pages whose registered website equals the outlet URL. Second, we supplement this set using self-reported employment histories in which individuals list both a LinkedIn company page and an outlet URL consistent with one of our outlets. Third, for outlets that remain unmatched, we assign LinkedIn company tags using ChatGPT. The resulting mapping links each outlet in our sample to a LinkedIn company identifier.

blace observations. LinkedIn records employment at the author–employer–spell level. We transform these data into author–month–workplace observations. An author–month pair is not necessarily unique: a single individual may appear in multiple workplaces in the same month. This arises for three reasons: (i) individuals sometimes report employment spells with one month of overlap when changing jobs; (ii) some individuals can hold positions at multiple organizations; and (iii) some profiles omit end dates for prior employment. To limit measurement error from concurrent or incomplete spell reporting, we drop authors with more than 15 author–month–workplace observations in 2022. This restriction removes authors with extensive concurrent employment, incomplete employment histories, or more than three job changes during the year.

n with production. We merge the monthly outlet affiliation panel with the PeakMetrics author specialization panel. For each author–month, we compute the share of articles written on each STM topic. We then lag these author-topic shares by one month, so that an author being counted in month t is attributed to a topic specialization based on their topic shares from month $t-1$. Finally, we aggregate the resulting author–month topic specialization counts to the outlet level.

A.8 Connecting production shares to staffing

This section describes how we process the GDELT production panel and the outlet-topic author panel to estimate the Leave-One-Out Poisson regression.

Calendar Aggregation The outlet-level staff panel, which we construct in Appendix A.7, is defined at the monthly level. We merge the two datasets by the month corresponding to each day, topic, and outlet. The resulting panel contains approximately 18.0 million outlet–day–topic observations across 2,570 outlets. Author counts are missing for 13.6% of the observations, equal to zero for 77.5%, and strictly positive for 8.9%. Among the 2,570 outlets, 1,824 (71.0%) have at least one strictly positive author count.

Leave-One-Out Averages For each production-share measure y and each day–topic pair (d, k), we construct the following leave-one-out peer average:

$$y^{\text{l1o}}_{j,k,d} = \frac{\sum_{j'} y_{j',k,d} - y_{j,k,d}}{N_{k,d} - 1}, \tag{20}$$

where $N_{k,d}$ is the number of outlets with an observation for topic k on day d. This quantity is the average production share devoted to topic k on day d across all outlets other than outlet j. The projection regression below uses the article-count production-share measure.

Poisson Regressions We construct fitted production projections using Poisson regressions with high-dimensional fixed effects. In our main specification, we regress the article-count production share on its leave-one-out peer average and the lagged author count:

$$y_{j,k,d} \sim \text{Poisson}(\lambda_{j,k,d}), \tag{21}$$

where

$$\lambda_{j,k,d} = \exp\Big(\alpha y^{\text{l1o}}_{j,k,d} + \beta \,\text{authors}_{j,k,m(d)-1} + \gamma_{j,k} + \mu_j + \nu_k\Big). \tag{22}$$

Here:

- $y_{j,k,d}$ is the share of articles published by outlet j on day d that cover topic k, and $y^{\text{l1o}}_{j,k,d}$ is the corresponding leave-one-out peer average.
- $\text{authors}_{j,k,m(d)-1}$ is the number of authors covering topic k at outlet j in month $m(d)-1$, that is, in the month preceding day d.
- $\gamma_{j,k}$ is a composite outlet–topic fixed effect. For outlet–topic pairs with variation in the number of authors over time, the corresponding observations share a common fixed-effect category, allowing the coefficient on $\text{authors}_{j,k,m(d)-1}$ to be identified from this variation. For outlet–topic pairs without observed variation in author counts, $\gamma_{j,k}$ is instead defined as a separate outlet–topic fixed effect. Thus, the author-count coefficient is estimated whenever the data contain within-category variation, while a separate outlet–topic fixed effect absorbs persistent differences when such variation is unavailable.
- μ_j and ν_k are outlet and topic fixed effects, respectively.

We estimate Equation (21) using Poisson quasi-maximum likelihood with IID standard errors. Table 2 in Section 4 reports the resulting estimates.

B Consumption data processing

B.1 Comscore data

We start from the raw Comscore URL traffic table, which records time-stamped observations, measured to the nearest second, of specific users' visits to specific URLs. We limit to the set of users with non-

missing demographics data. We compute the time a user spent on each page ("attention") as the difference between the timestamp of the current visit and the next visit by the same machine, capping this duration at 90 seconds per the recommendation in Clemm von Hohenberg et al. (2024). When multiple URLs share the same timestamp for a given machine, we divide the computed duration between this timestamp and the previous distinct timestamp equally among them.

We then join the time-stamped attention data with URL-level topic predictions, described in Appendix A.3. We match on outlet domain, visit date, and URL. This attaches a predicted topic to each page visit where a prediction is available. Time spent on pages other than the home page on recognized news outlets is assigned to the matched topic, where available, or topic 0 otherwise. Time spent on the home page of a news outlet is allocated across topics in proportion to the outlet's production shares on that date (constructed as described in Appendix A.4). Time spent on any domain not in the list of news outlets available in GDELT is aggregated into a single non-news category, with null topic.

Finally, we aggregate to the level of user × topic × outlet × day, summing our attention measure over all visits by the user to the topic × outlet that day. Because weekend days have much lower publication volume and web traffic, we collapse weekends together with the preceding Friday to form a single synthetic day observation, with attention averaged across the three days. This aggregation smooths over lower weekend traffic volumes. We then perform two filtering steps. We first drop any user × topic × outlet × day observation with under 6 total seconds of browsing time. Second, we flag users as likely bots if either (a) more than 20% of the days in which they appear in the panel have more than 6 hours of total browsing, or (b) any single day exceeds 12 hours of total browsing.[27] This identifies 4,034 suspected bot users which we remove from the panel.

The resulting attention dataset has 33,396,377 user × topic × outlet × day observations, from 470,207 distinct users. Of these, 7,499,184 observations are time spent on news domains, by 174,949 distinct users.

B.2 Top outlets

Table 8 lists the 15 most-read outlets in our sample by total reading time. All 15 are national or international; no local outlet appears in this list. Table 9 lists the 15 most-read local outlets. Even the largest local outlet commands a small share of total reading time relative to the top national outlets, reflecting the nationalization of online news consumption.

[27] Note that the sum is over browsing intervals which are individually capped at 90s. Getting to 6 (12) hours of total browsing time measured this way hence requires visiting a minimum of 240 (480) pages in a day.

Outlet	Scope	Type	Share (%)
espn.com	National	Specialized: Sports	13.30
foxnews.com	National	General: National	8.50
weather.com	National	General: National	5.30
cnn.com	National	General: National	5.30
nytimes.com	National	General: National	5.00
washingtonpost.com	National	General: National	1.90
crunchyroll.com	National	Specialized: Entertainment and culture	1.90
dailymail.co.uk	International	General: International	1.30
usatoday.com	National	General: National	1.30
nypost.com	National	General: National	1.20
bbc.com	International	General: International	1.10
breitbart.com	National	Specialized: Politics and social issues	0.90
247sports.com	National	Specialized: Sports	0.90
webmd.com	National	Specialized: Health and wellness	0.70
marketwatch.com	National	Specialized: Finance and business	0.70

Table 8: Top 15 outlets by total reading time.

Notes: Outlets ranked by total reading time across all panelists. Share is the percentage of total news reading time in the sample. Scope indicates geographic classification. Type combines the general/specialized label with the outlet's subcategory.

Outlet	State	Share (%)
ksl.com	UT	0.60
patch.com	MULTIPLE	0.40
wtop.com	DC	0.30
startribune.com	MN	0.30
wqad.com	IL	0.20
al.com	AL	0.20
latimes.com	CA	0.20
nj.com	NJ	0.20
sfgate.com	CA	0.20
chattanoogan.com	TN	0.20
cleveland.com	OH	0.20
stltoday.com	MO	0.20
wral.com	NC	0.20
10tv.com	OH	0.20
mlive.com	MI	0.20

Table 9: Top 15 local outlets by total reading time.

Notes: Local outlets ranked by total reading time across all panelists. Share is the percentage of total news reading time in the sample (across all outlets, not just local). State is the outlet's primary geographic market.

B.3 Clustering consumers

We cluster consumers into groups with similar news consumption patterns using spherical k-means (Hornik et al., 2012), a variant of k-means that uses cosine dissimilarity rather than Euclidean distance.

For each consumer, we construct a consumption vector over outlet–topic pairs, computing total consumption hours across all browsing sessions for each unique combination of news outlet and topic (as classified by our topic model). This yields a sparse consumer-by-(outlet $\times$ topic) matrix, where each row represents a consumer and each column represents a specific outlet–topic pair.

This sparse consumption matrix is the input data to the clustering procedure. We use hard partitions $m = 1$, and 25 random restarts to mitigate sensitivity to initialization. We cluster users separately within each of 59 geographic markets: the 50 largest Designated Market Areas (DMAs) by population, plus non-top-50-DMA areas in each of the 9 US Census sub-regions. Within each market, we apply spherical k-means with 20 clusters, yielding 1,180 total clusters across all markets.

Cluster characteristics. The resulting clusters vary substantially in size and composition. The median cluster contains 95 consumers (mean 145; range 26 to 1,287). Appendix figure 14 visualizes the consumption space using a UMAP projection (McInnes et al., 2018) of a random subsample of 20,000 consumers, colored by each consumer's most-consumed topic. Consumers with similar reading habits are placed nearby in the projection. The plot reveals distinct groups: sports readers (right) form a tight cluster, while technology, weather, and entertainment readers separate into recognizable groups. The center contains consumers with more diffuse reading patterns.

Appendix figure 15 shows how clusters differ in the geographic scope of their reading. Each row is a cluster, ordered by the share of reading time devoted to local outlets. Some clusters read almost exclusively from national outlets, while others devote more than half of their reading to local sources. Among markets, Salt Lake City has the highest average local consumption share (41% averaged across its 20 clusters), with one cluster reading 93% local outlets. This variation across clusters within the same geographic market confirms that our clustering captures meaningful differences in how consumers engage with local versus national news.



Figure 14: Consumers with similar reading habits cluster together in topic space.

Notes: UMAP projection of a random subsample of 20,000 consumers based on their outlet×topic consumption vectors (top 1,000 columns by total consumption). Each point is a consumer, colored by their most-consumed non-junk topic. Proximity reflects similarity in consumption patterns (cosine distance).

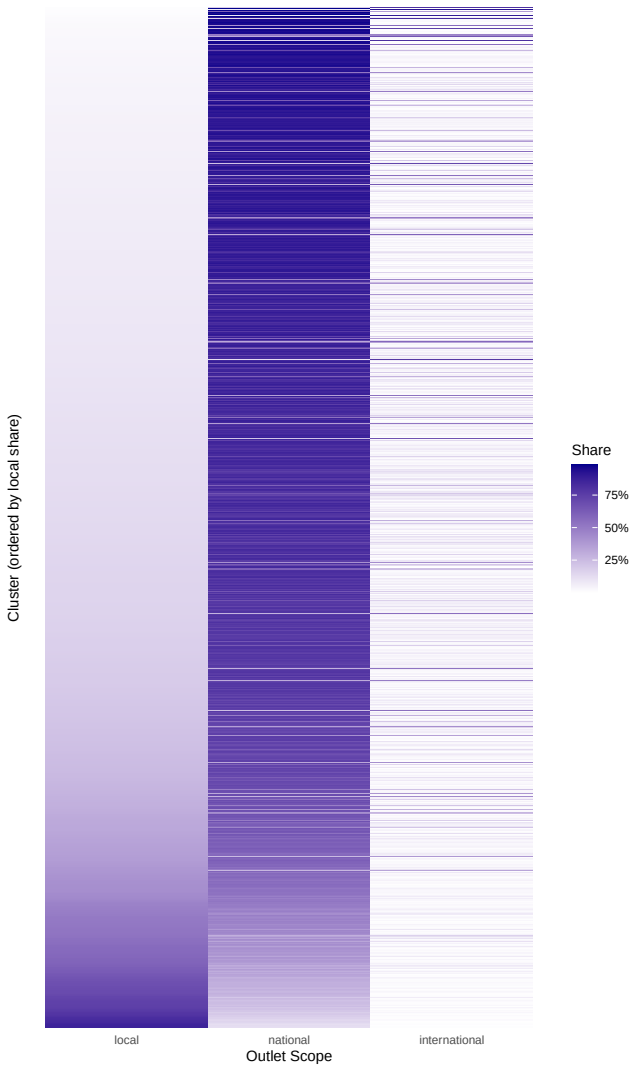


Figure 15: Clusters vary widely in the geographic scope of their news reading.

Notes: Each row is one of the 1,180 consumer clusters, ordered by the share of reading time on local outlets (decreasing from bottom to top). Columns show the share of total cluster reading time devoted to local, national, and international outlets.

C Article-Level Front-Page Placement Robustness

Figure 6 uses within-article changes in front-page placement to estimate how editorial prominence affects readership. Here we report estimates for an event-matched sample and assess whether the results are robust to controlling for changes in reader interest in the underlying news event that are common across outlets.

Event classification. We identify events using the PeakMetrics article corpus. After restricting to English-language articles from outlets in the Comscore sample and removing invalid dates, HTML, and frequently repeated within-outlet boilerplate language[28], we represent each article's combined title and body with a term frequency–inverse document frequency (TF–IDF) vector. Following the Topic Detection and Tracking framework of Allan et al. (1998), we apply DBSCAN with cosine distance within overlapping three-day windows, a radius of 0.25 and the requirement that initial clusters contain at least two articles. We merge clusters across overlapping windows when they share an article, split clusters when consecutive articles are separated by at least 24 hours, and discard candidate events that consist only of near-duplicate syndicated text, include fewer than two outlets, span more than two weeks, or contain predominantly non-English text. The resulting set of clusters denotes events. The intention is that these events capture granular stories that are covered by multiple outlets are are closely related in specific language and timing, rather than a broad topic.

We map these events to GDELT using outlet-specific URL matches. When an exact URL match is unavailable, we compare the informative portion of the GDELT URL with PeakMetrics article titles in the same TF–IDF space, searching within a five-day window and accepting matches with cosine distance no greater than 0.1. Because events are constructed from PeakMetrics text, this exercise restricts the GDELT sample to articles that also appear in the PeakMetrics corpus and receive a valid event assignment.

Event-by-hour fixed effects. Let e_a denote the event assigned to article a and let t index calendar hours. We augment the article-level specification in Equation (1) with event-by-hour fixed effects:

$$y_{a,t} = \rho \cdot \text{placement}_{a,t} + f(\text{age}_{a,t}) + \phi_a + \lambda_{e_a,t} + \varepsilon_{a,t}, \tag{23}$$

where $\text{placement}_{a,t}$ is either an indicator for appearing on the front page or the article's position among links currently on the front page. The fixed effect $\lambda_{e_a,t}$ absorbs event-specific changes in readership at each hour that are shared across outlets—for example, a new development that simultaneously raises interest in every article about the event. The additional event-by-hour control helps confirm that the identifying variation in Eq. (23) is driven by differences in the placement across articles covering the same event in the same hour and not the intrinsic interest of the topic covered by a given article.

Tables 10 and 11 report pooled estimates for the two placement measures. The front-page indicator remains positive, while a lower placement on the page remains associated with less readership. The

[28] Boilerplate language is recurring header or footer text. For example, the text, "1996–2020 The Washington Post Read content from all..." is repeated across full text samples from the Washington Post.

monthly position estimates in Appendix figure 16 are likewise negative across outcomes and age-control specifications, although the event-matched sample and the more demanding fixed effects produce wider confidence intervals than in the main analysis.

Table 10: Front-page indicator regressions with event-by-hour fixed effects

	Total Visits (1)	Unique Users (2)	Log (1 + Visits) (3)	Log (1 + Unique Users) (4)
On Front Page	0.055*	0.054*	0.031**	0.030**
	(0.030)	(0.029)	(0.014)	(0.014)
Age Controls (Cubic Spline)	✓	✓	✓	✓
Event-Hour FEs	✓	✓	✓	✓
Outlet-URL FEs	✓	✓	✓	✓
Observations	1,782,912	1,782,912	1,782,912	1,782,912
R^2	0.422	0.465	0.459	0.458

Notes: Pooled article-hour regressions on the event-matched 2022 sample. Each column includes a cubic-spline control for article age, article (outlet-by-URL) fixed effects, and event-by-calendar-hour fixed effects.

Table 11: Front-page position regressions with event-by-hour fixed effects

	Total Visits (1)	Unique Users (2)	Log (1 + Visits) (3)	Log (1 + Unique Users) (4)
Link Placement (Rank)	-0.0005**	-0.0004**	-0.0002***	-0.0002**
	(0.0002)	(0.0002)	(8.76×10^{-5})	(8.97×10^{-5})
Age Controls (Cubic Spline)	✓	✓	✓	✓
Event-Hour FEs	✓	✓	✓	✓
Outlet-URL FEs	✓	✓	✓	✓
Observations	1,174,812	1,174,812	1,174,812	1,174,812
R^2	0.461	0.532	0.507	0.509

Notes: Pooled article-hour regressions restricted to articles currently on the front page in the event-matched 2022 sample. A higher link-placement rank corresponds to a position farther down the page. Each column includes a cubic-spline control for article age, article (outlet-by-URL) fixed effects, and event-by-calendar-hour fixed effects.

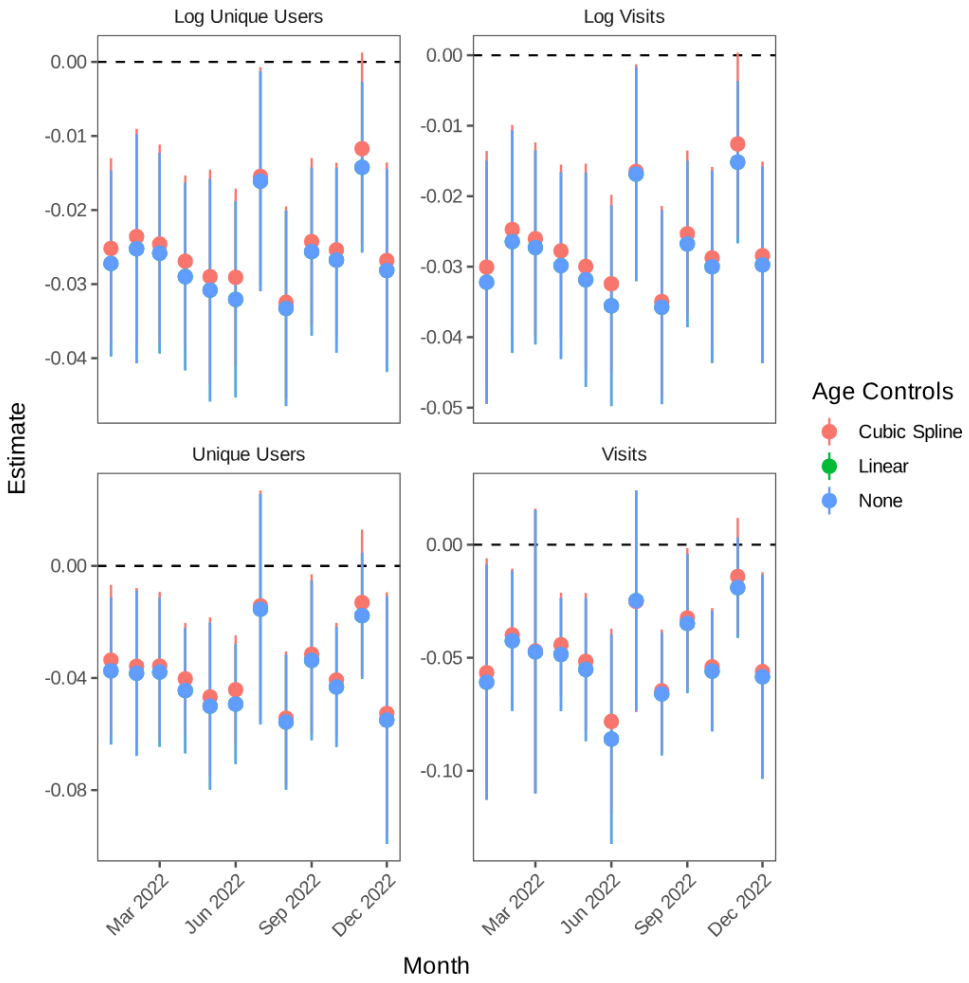


Figure 16: Front-page position effects with event-by-hour fixed effects.

Notes: Monthly estimates of equation (23) for 2022, restricted to articles observed in both GDELT and PeakMetrics and assigned to a valid event. The sample includes articles currently on the front page, and the coefficient is on front-page position, with larger values indicating placement farther down the page. Panels report visits and unique users in levels and logs; colors indicate the article-age controls shown in the legend. All specifications include article and event-by-calendar-hour fixed effects. Vertical bars show 95% confidence intervals.

D Additional Reduced-Form Evidence

The main text reports pooled IV estimates of the effect of production share on consumption (Table 4). Here we present the topic-by-topic results. We estimate equation (2) separately for each topic k, restricting to national specialized outlets, with outlet×topic and topic×day fixed effects, and instrumenting $p_{j,k,d}$ with the Poisson-fitted leave-one-out instrument $z_{j,k,d}$ described in Appendix A.5. Standard errors are clustered by time period.

Appendix figure 17 shows that the instrument is strong ($F > 10$) for 13 of 29 substantive topics. Topics with the strongest first stages—Ukraine–Russia, Financial Markets, American Sports, COVID-19—are those for which industry-wide coverage responds sharply to common events. Appendix figure 18 reports the IV estimates of θ for the topics with strong instruments. The estimates are positive and significant for the majority of these topics, confirming that exogenous increases in an outlet's topic coverage shift its readers' consumption of that topic. The magnitudes range from approximately 0.5 to 2.5 percentage points per one-percentage-point increase in production share, with substantial heterogeneity across topics.

These topic-level coefficients should not be interpreted as measuring consumer preferences for topics. Each θ_k captures the total consumption response to an exogenous coverage increase—a composite of genuine reader interest in topic k and outlet-specific agenda-setting power. A large θ_k could reflect strong demand for the topic, or it could reflect the fact that outlets covering k happen to have loyal audiences who read whatever those outlets publish. Separating these two forces requires the structural model in Section 5, which decomposes the response into a topic preference ($\beta_{i,k}$) and an outlet-specific agenda-setting component ($\gamma_{i,j}$).

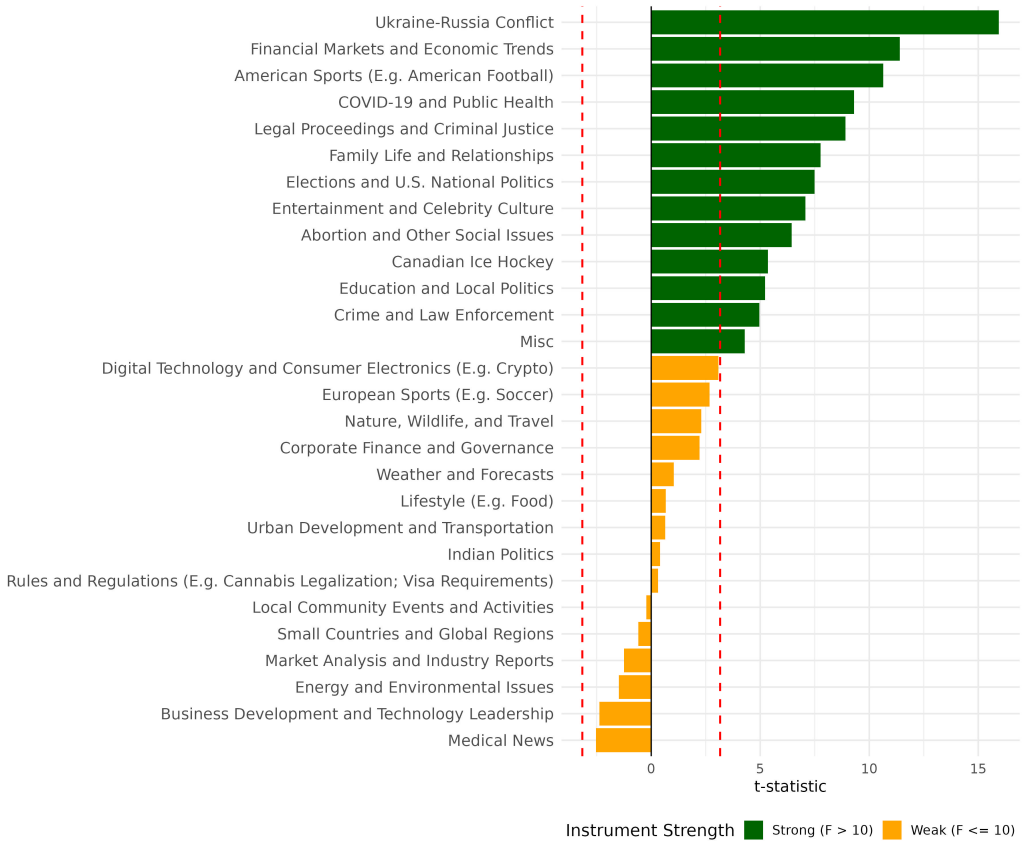


Figure 17: First-stage t-statistics by topic.

Notes: Each bar shows the t-statistic from a first-stage regression of outlet j's production share of topic k on the Poisson-fitted instrument $z_{j,k,d}$, with outlet×topic and topic×day fixed effects. Sample: national specialized outlets. Dashed red lines mark $|t| = 3.16$ ($F \approx 10$). Green bars: $F > 10$; orange bars: $F \leq 10$.

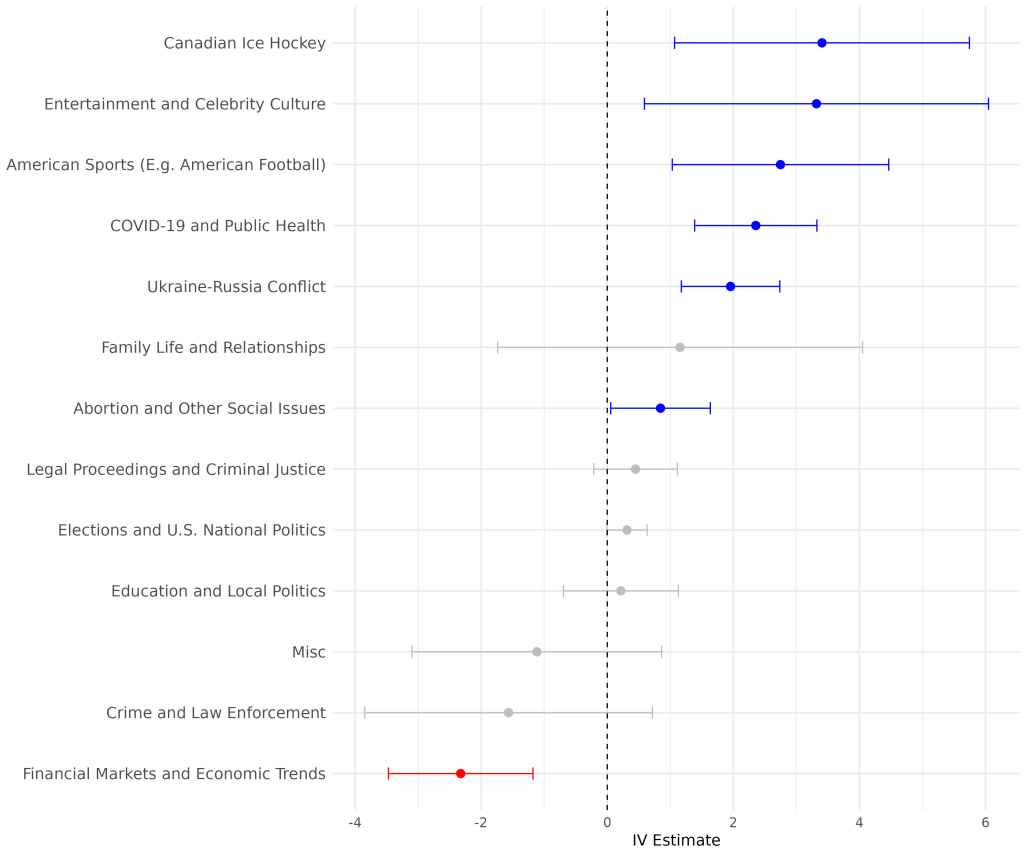


Figure 18: IV estimates of the effect of production share on consumption, by topic.

Notes: IV estimates of θ from equation (2), estimated separately by topic. Sample restricted to national specialized outlets and topics with first-stage $F > 10$. Bars show 95% confidence intervals. Blue: positive and significant at 5%; gray: not significant. The instrument is $z_{j,k,d}$ as defined in Appendix A.5.

E Cross-consumer heterogeneity in counterfactuals

??

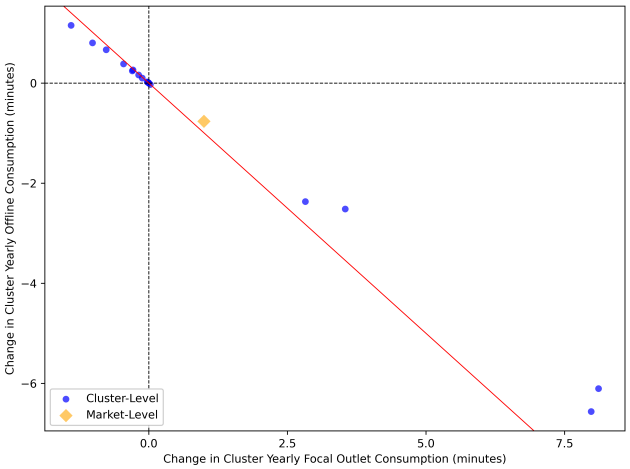


Figure 19: Case Study 1a: Baltimore Sun Crime Reporting

Appendix figure 19 zooms into the aggregates in Figure 9 at the consumer level. Because most of the substitution is between the Baltimore Sun and the outside option ("offline time"), we focus on these two options. On the x-axis, we plot the absolute change in reading time (in minutes) on the Baltimore Sun by each cluster of consumers in the Baltimore DMA. On the y-axis, we plot the absolute change in time spent offline. The red line corresponds to the -45 degree line, indicating that the Baltimore Sun's share of consumption increases by the same amount that the outside option decreases and vice versa. As the plot shows, the majority of clusters slightly decrease their reading time on the Baltimore Sun as a result of the increased crime reporting, largely due to substitution toward the outside option. However, the handful of clusters that increase their reading time on the Baltimore Sun do so by an order of magnitude more and their points on the plot are farther from the red line, suggesting that they substitute toward the Baltimore Sun both from the outside option and from other outlets.

Appendix figure 20 zooms in further to show what consumers read when they substitute toward or away from the Baltimore Sun. The left panel shows that consumers who substituted away from the Baltimore Sun mostly read less about American sports, entertainment and miscellaneous topics—suggesting that their response was driven by the decrease in the share of reporting on these topics that compensated the increase in crime reporting. The right panel shows that consumers who substituted toward the Baltimore Sun largely increased their reading of articles about crime (and slightly decreased their reading of all other topics).

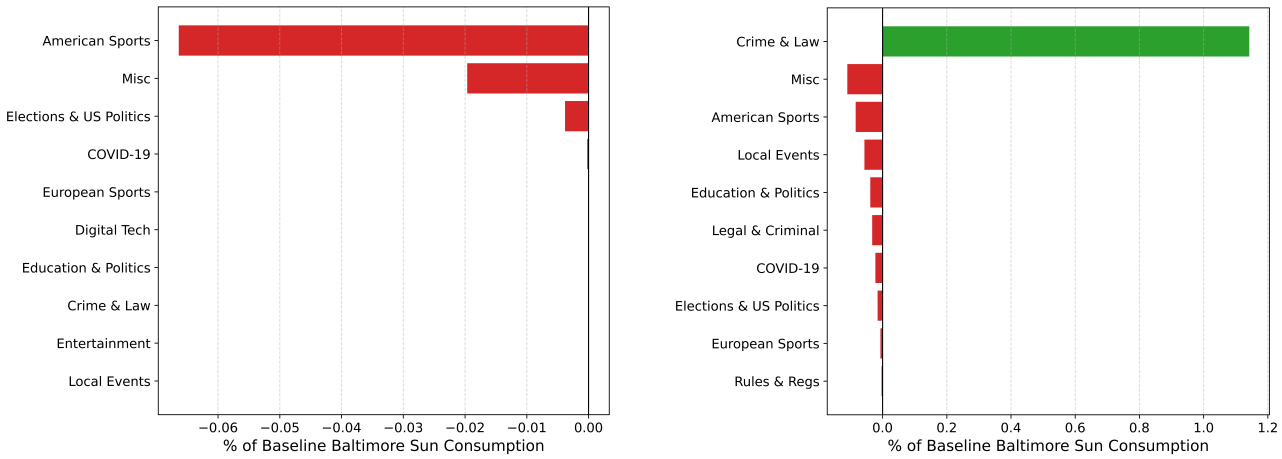


Figure 20: Substitution patterns among Baltimore Sun readers: Consumers decreasing (left) and increasing (right) their reading after increased crime coverage.

F Supplemental Figures

F.1 Additional Descriptive Figures

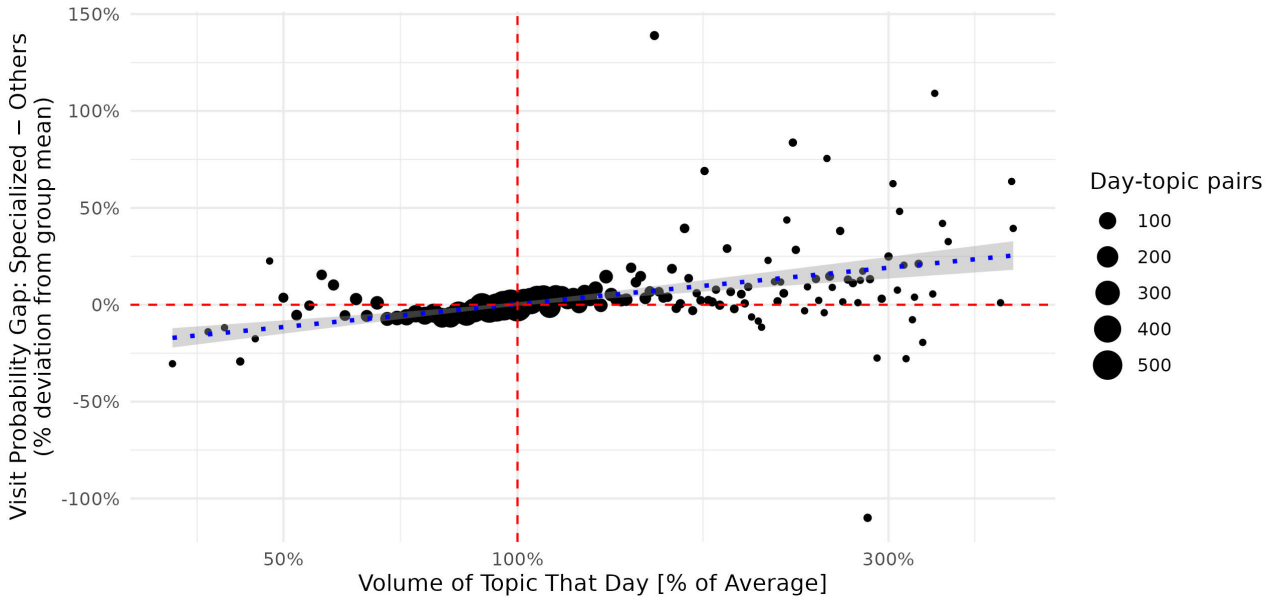
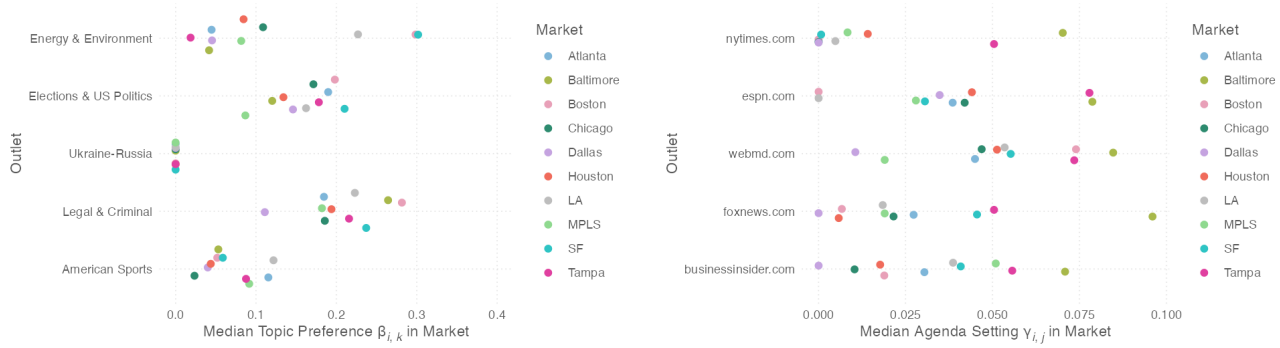


Figure 21: Specialized outlets attract new readers on high-event days. Binscatter analogous to panel (b) of Figure 5, restricted to visits by panelists who had not previously visited the outlet in our panel. The x-axis is the aggregate production share of a topic on a given day, relative to that topic's average. The y-axis is the gap in visit probability between outlets specialized in the topic and other outlets, each normalized as the percentage deviation from the group's own time-series average. Dashed red lines mark the average-day reference point. Point size reflects the number of topic-day pairs in each bin.

F.2 Additional demand descriptives

Table 12 reports the counterpart of Table 3 for frequent consumers, the top decile of panelists by total news reading time (more than 2.8 hours over 2022). Table 13 places the volume and concentration statistics for all news consumers and for frequent consumers side by side.



(a) Median estimated $\beta_{i,k}$ across markets for selected topics *(b) Median estimated $\gamma_{i,j}$ across markets for selected outlets*

Figure 22: Market heterogeneity in topic preferences and agenda setting. The left panel plots median topic-preference parameters $\beta_{i,k}$ for selected substantive topics; the right panel plots median agenda-setting parameters $\gamma_{i,j}$ for selected national outlets. Each point is a market (DMA).

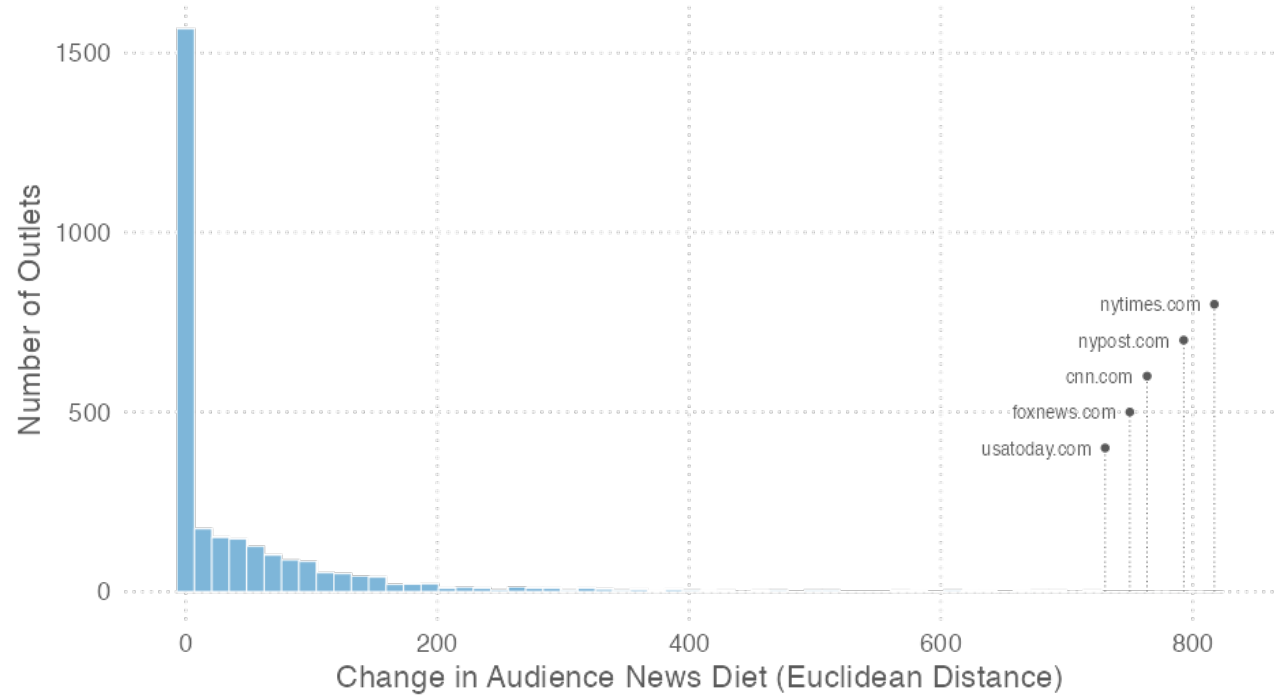


Figure 23: Change in audience news diet when an outlet's agenda setting is removed, by outlet. The horizontal axis is the Euclidean distance between each outlet's audience's topic attention vector in the baseline and in the counterfactual with $\gamma_{i,j} = 0$ for that outlet's reader clusters. Labeled points are selected large national outlets.

	Any (%)	10%	25%	Median	75%	90%	Mean	SD
Panel A: Volume and concentration of news reading								
Minutes Reading News		190.98	240.04	386.26	786.84	1,817.97	844.97	1,592.58
Minutes Online, Non-News		1,483.95	3,359.88	7,019.37	13,252.23	22,012.06	9,904.33	9,542.84
Number of Outlet-Day Visits		52.00	92.00	152.00	276.00	490.80	235.71	269.25
Number of Unique Outlets Read		5.60	12.00	27.00	52.00	82.00	37.68	36.77
Number of Unique Topics Read		17.00	22.00	25.00	27.00	29.00	24.05	4.67
% Consumption of Top Topic		16.99	21.56	29.76	44.66	62.08	34.95	17.53
% Consumption of Top 3 Topics		39.86	46.31	57.21	74.99	86.03	60.44	17.27
Panel B: Share of reading time on each topic, in percent (top 10 topics)								
American Sports	96.5	0.60	1.77	4.60	22.77	62.80	17.06	23.79
Entertainment	98.0	0.79	2.31	5.75	11.26	22.04	9.46	11.83
Elections & US Politics	87.9	0.00	0.51	2.91	9.29	18.38	6.53	8.96
Misc	98.2	1.51	3.11	5.19	8.65	14.28	7.65	9.37
Weather	85.9	0.00	0.31	1.25	3.85	14.09	5.23	11.26
Crime & Law	92.2	0.08	0.87	3.15	7.21	12.83	5.26	6.55
Financial Markets	85.7	0.00	0.26	1.19	3.55	10.45	4.52	10.20
Digital Tech	96.2	0.37	0.97	2.26	4.68	9.56	4.60	7.89
European Sports	94.9	0.27	0.96	2.11	4.55	9.47	4.11	6.94
Ukraine-Russia	85.8	0.00	0.31	1.86	5.38	9.87	3.89	5.86

Table 12: Distribution of news consumption across panelists (2022): Frequent consumers.

Notes: Panel A reports the distribution across the 7,837 frequent consumers in the Comscore panel over 2022 (the top decile by total reading time, more than 2.8 hours). "% Consumption of Top Topic" is the share of a panelist's news reading time on their single most-read topic; "Top 3 Topics" is the combined share of their three most-read topics. Panel B covers the 7,836 panelists with time on substantive topics; shares are in percent of a panelist's reading time excluding the junk/sections bucket. "Any" is the share of panelists with any time on the topic. Mean and SD are taken across all panelists, counting those with no time on the topic as zeros; the percentiles also include those zeros.

Var	10%	25%	Median	75%	90%	Mean	SD
Panel A: All news consumers (N = 83,028)							
Minutes Reading News	6.18	8.77	17.63	48.25	156.90	105.47	545.24
Minutes Online, Non-News	409.51	1,011.27	2,367.27	5,202.25	10,292.12	4,326.38	5,894.33
Number of Outlet-Day Visits	3.00	5.00	10.00	26.00	73.00	36.82	106.24
Number of Unique Outlets Read	2.00	4.00	7.00	13.00	25.00	11.61	16.47
Number of Unique Topics Read	4.00	5.00	9.00	15.00	22.00	10.79	6.92
% Consumption of Top Topic	20.76	26.32	35.36	48.87	64.05	39.29	17.29
% Consumption of Top 3 Topics	47.89	57.79	71.86	86.47	97.75	71.93	17.75
Panel B: Frequent consumers (N = 7,837)							
Minutes Reading News	190.98	240.04	386.26	786.84	1,817.97	844.97	1,592.58
Minutes Online, Non-News	1,483.95	3,359.88	7,019.37	13,252.23	22,012.06	9,904.33	9,542.84
Number of Outlet-Day Visits	52.00	92.00	152.00	276.00	490.80	235.71	269.25
Number of Unique Outlets Read	5.60	12.00	27.00	52.00	82.00	37.68	36.77
Number of Unique Topics Read	17.00	22.00	25.00	27.00	29.00	24.05	4.67
% Consumption of Top Topic	16.99	21.56	29.76	44.66	62.08	34.95	17.53
% Consumption of Top 3 Topics	39.86	46.31	57.21	74.99	86.03	60.44	17.27

Table 13: Distribution of news consumption across panelists (2022).

Notes: Each row reports the distribution across Comscore panelists over 2022. Panel A includes all panelists with at least five minutes of news reading over the year; Panel B restricts to frequent consumers, the top decile by total reading time (more than 2.8 hours). "Minutes Online, Non-News" is time spent on all non-news websites over the year. Outlet-day visits count outlet-days with positive reading time. "% Consumption of Top Topic" is the share of a panelist's news reading time devoted to their single most-read topic; "Top 3 Topics" is the combined share of their three most-read topics.