Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

From Feeds to Inboxes: A Comparative Study of Polarization in Facebook and Email News Sharing

Hema Yoganarasimhan, Irina Iakovetskaia

To cite this article:
Hema Yoganarasimhan, Irina Iakovetskaia (2024) From Feeds to Inboxes: A Comparative Study of Polarization in Facebook and Email News Sharing. Management Science 70(9):6461-6472. https://doi.org/10.1287/mnsc.2023.04134

https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 70, No. 9, September 2024, pp. 6461–6472
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

From Feeds to Inboxes: A Comparative Study of Polarization in Facebook and Email News Sharing

Hema Yoganarasimhan,[a,*] **Irina Iakovetskaia**[b]

[a] Foster School of Business, University of Washington, Seattle, Washington 98195; [b] Stanford Graduate School of Business, Stanford University, Stanford, California 94305

*Corresponding author

Contact: hemay@uw.edu, https://orcid.org/0000-0003-0703-5196 (HY); irinay@stanford.edu, https://orcid.org/0009-0008-2447-8398 (II)

Received: December 16, 2023
Accepted: February 3, 2024
Published Online in Articles in Advance: July 9, 2024

https://doi.org/10.1287/mnsc.2023.04134

Copyright: © 2024 INFORMS

Abstract. This study explores the polarization of news content shared on Facebook compared with email using data from the *New York Times*' Most Emailed and Most Shared lists over 2.5 years. Employing latent Dirichlet allocation and large language models (LLMs), we find that highly polarized articles are more likely to be shared on Facebook (versus email), even after accounting for factors like topics, emotion, and article age. Additionally, distinct topic preferences emerge, with social issues dominating Facebook shares and lifestyle topics prevalent in emails. Contrary to expectations, political polarization of articles shared on Facebook did not escalate post-2020 election. We introduce a novel approach to measuring polarization of text content that leverages generative artificial intelligence models, like ChatGPT, and it is both scalable and cost effective. This research contributes to the evolving intersection of LLMs, social media, and polarization studies, shedding light on descriptive patterns of content dissemination across different digital channels.

History: Accepted by Duncan Simester, marketing.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mnsc.2023.04134.

Keywords: polarization • social media • large language models (LLMs) • Facebook • news

1. Introduction

Over the last two decades, we have seen significant shifts in how people consume and share news. Historically, users obtained news directly from established news sources by reading newspapers or news websites. However, over the years, the way that users are exposed to news and consume news has changed because of the availability of digital communication formats that allow users to share news articles with their peers (e.g., emails and messages). More recently, with the rise of social media platforms, such as Facebook and Twitter, users share news from news websites on their social media pages, which in turn, can be reshared by others. Thus, users can now obtain news without going to the original news website. Although these changes have led to easier access to news, there have been concerns that these changes have skewed news sharing and consumption toward polarizing topics.

Earlier research has provided some support for the idea that social media platforms promote polarized discourse through mechanisms such as homophilic network formation, echo chambers, and filter bubbles, and it has suggested that personalization algorithms can amplify these effects (Pariser 2011, Barberá et al. 2015, Sunstein 2018, Allcott et al. 2020, Persily and Tucker 2020, Cinelli et al. 2021, Iyer and Yoganarasimhan 2021, Levy 2021, Shin and Kadiyala 2022). On the other hand, some emerging research has questioned this narrative (Gentzkow and Shapiro 2011, Boxell et al. 2017, Bail et al. 2018, Eady et al. 2019). For instance, some early work shows that users' exposure to political news on Facebook is primarily driven by the content shared by their friends rather than the algorithm used to determine users' news feeds (Bakshy et al. 2015).

In this paper, we examine one specific aspect of this issue. Is the content seeded on social media (Facebook) systematically more polarized than that shared via more personal channels (email) after controlling for a variety of article features, such as the content and topics in the article, and how has such polarization changed over time?

To answer these questions, we use the data from the *New York Times* (*NYTimes*) Most Emailed and Most Shared in this study. Our analysis consists of four main steps. First, for our data collection efforts, we focus on the Trending section of *NYTimes*, which lists the top 20 most emailed and the top 20 most shared articles on Facebook over the last 24 hours. Both of these lists are based on the data gathered by the share buttons on each *NYTimes* article. We collect daily data on these rank-ordered lists for an ≈2.5-year period starting from January 1, 2019 and ending May 30, 2021. In addition, we collect the metadata and text for each article that appears at least once across the two lists (a total of 13,508 unique articles). In the second step, we use a latent Dirichlet allocation (LDA) model to recover the distribution of topics in each article based on the unstructured text in the entire corpus. Third, we obtain measures of polarization for all the articles in our corpus using the newly developed large language models (LLMs). We validate these LLM-based measures of polarization using a user survey. Finally, we quantify the relationship between a topic's prevalence in an article and its relative popularity on Facebook versus email using a descriptive model. In the process, we control for time-varying article-specific shocks that can affect an article's ranking on both lists.

We now discuss our main findings. We find that more polarizing articles have a higher relative likelihood of being shared on Facebook compared with email, even after controlling for a series of confounding factors, such as the topics covered in the article, news section, emotional content, and age of the article. In addition, we find that socially relevant and general interest topics are more likely to be posted on social media compared with being shared via email (after controlling for the polarization score and other control variables). Specifically, articles on topics such as *Books, Business, Animals, Food, Real Estate, Nature, and Health Research and Lifestyle Advice* are more commonly sent through email. In contrast, articles on topics such as *Election Investigations, Covid Vaccine, Russia, Women's Issues and Sexual Harassment, Coronavirus Pandemic, Black Lives Matter*, etc. are more likely to be posted on social media. Next, we examine if the articles seeded on Facebook become more politically polarized over time. This question is motivated by the discussions around the exacerbation of polarization on social media platforms after the 2020 election (Jurkowitz et al. 2020). Interestingly, we find the opposite; polarization scores of articles play a smaller role in predicting the differences across the two lists after the election.

Our paper contributes to the literature on social media and polarization in two ways. First, we show that the content seeded on social media websites is systematically different from that shared through other media formats both in terms of polarization as well as in terms of the distribution of topics covered. Although the exact causes of these descriptive findings are hard to pin down with our data (we discuss a nonexhaustive set of possibilities in Section 3.3.2), they nevertheless suggest that social media websites tend to attract more polarized content even at the content seeding stage, even before the explicit influence on recommendation algorithms on social media. Second, we show that the recently developed generative artificial intelligence models, such as ChatGPT, can be used to measure the polarization of text content in a scalable low-cost fashion (something not feasible with user surveys). Although some recent research has shown that these models produce responses consistent with true user preferences in certain settings (Brand et al. 2023), there is no consensus on the types of user preferences that these models can accurately simulate. Our findings provide some initial evidence in support of their accuracy in characterizing polarization and political bias. We hope that our findings will help spur further research on this topic.

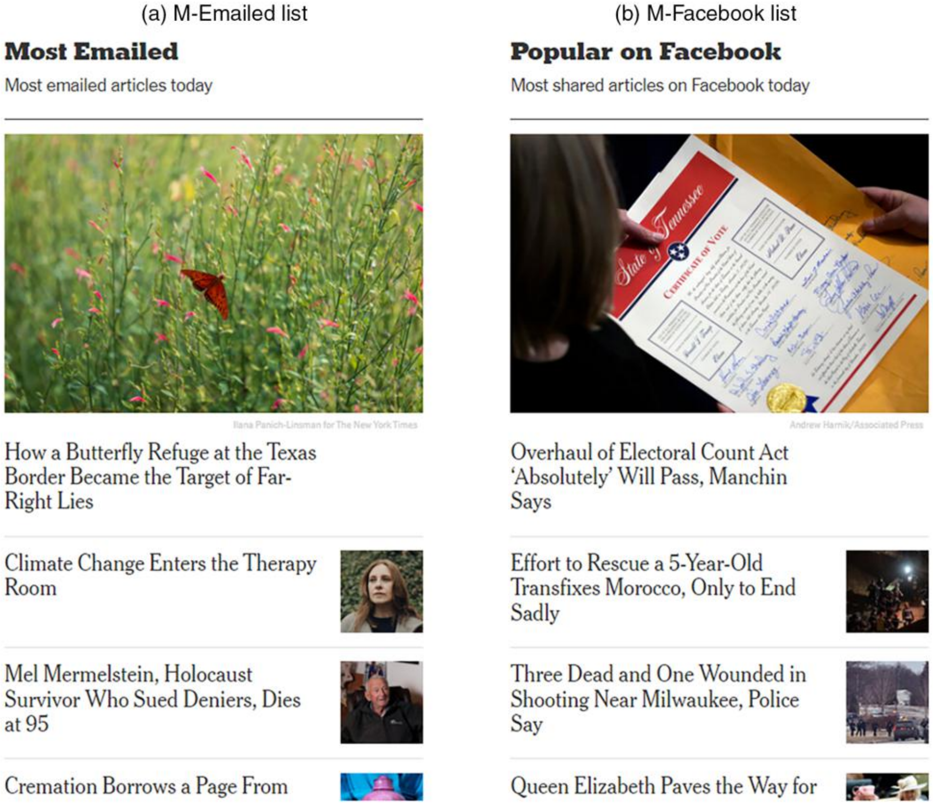
2. Data

Our data come from the Trending section on the *NYTimes* website, which has two rank-ordered lists: (1) Most Emailed and (2) Popular on Facebook, henceforth referred to as M-Emailed and M-Facebook, respectively, for convenience. Figure 1 illustrates an example of the two lists. The two lists are constructed based on the data from the share buttons on each article and then ranked accordingly. To be able to share articles (on email or Facebook) using the share buttons, the user needs to be logged in.[1]

We obtained data on these two lists using Internet Archive for an ≈2.5-year period starting from January 1, 2019 and ending May 30, 2021.[2] We parse these data for each day for the time closest to noon. Data are missing at random for some days, and we have data on a total of 697 days in our observation period. Over this period, we see 13,688 unique articles across both lists. In addition, we used Article Search Application Programming Interface to retrieve article metadata, such as headline, publication date, abstract, and section (Dev Portal 2022).[3] See Table A.1 in the Online Appendix for details. Of the 13,688 articles, 13,508 were accessible and had both metadata and full text available.[4]

We now conduct a preliminary analysis of the similarities and differences between the two lists. First, we pool all the articles over the entire observation period and examine the overlap between the two lists. There is only a small amount of overlap between the two lists; only 2,884 of the 13,688 articles (i.e., 20% of articles) appear on both lists at least once. Among the rest of the articles, 6,440 articles appear only on the M-Emailed list, whereas 4,404 articles appear only on the M-Facebook list. Next, we examine the overlap in the two lists on any given day. On average, 5.5 articles (of 20) appear on both lists on any given day (i.e., over 14 articles are different across the two lists at any given point in time) (see Figure A.1 in Online Appendix A.1). Thus, there are significant differences in the articles on the two lists (both across days and on any given day).

Figure 1. (Color online) Snapshot of the *New York Times* Trending Lists



(a) M-Emailed list

Most Emailed

Most emailed articles today

Ilana Panich-Linsman for The New York Times

How a Butterfly Refuge at the Texas Border Became the Target of Far-Right Lies

Climate Change Enters the Therapy Room

Mel Mermelstein, Holocaust Survivor Who Sued Deniers, Dies at 95

Cremation Borrows a Page From

(b) M-Facebook list

Popular on Facebook

Most shared articles on Facebook today

Andrew Harnik/Associated Press

Overhaul of Electoral Count Act 'Absolutely' Will Pass, Manchin Says

Effort to Rescue a 5-Year-Old Transfixes Morocco, Only to End Sadly

Three Dead and One Wounded in Shooting Near Milwaukee, Police Say

Queen Elizabeth Paves the Way for

Next, we examine whether the articles that appear on these two lists are systematically different on basic attributes derived from metadata. We find that articles shared on Facebook, on average, are shorter, with a mean of 1,390 words, compared with articles on the M-Emailed list, which average 1,540 words (a *t*-test confirmed that this difference is significant).[5] Further, on the M-Emailed list, over 30% come from the Opinion section followed by the United States section (at 10%) and then, Health, Well, Business Day, etc. In contrast, on the M-Facebook list, over 30% of the articles come from the United States section followed by Opinion, World, and New York (see Figure A.2 in Online Appendix A.1 for details). In summary, the length and section headings of articles that appear on the two lists are quite different.

Nevertheless, metadata, such as section names and the type of material, have limited ability to categorize news content or explain the difference in the news content across the two lists for two reasons. First, newsworthy topics change regularly, and it is hard for rigid and long-established news structures (e.g., section names) to capture constantly evolving topics covered in news cycles. Second, most news articles cover two or more topics, which makes it difficult to categorize them under one section name or news desk. See Online Appendix A.2 for a detailed example with two articles that highlight these challenges. Therefore, we need to go beyond the metadata and learn more about the content and tone of the articles to quantify the differences between the articles shared through the two different mediums.

3. Empirical Analysis and Results

Our empirical analysis consists of two steps. In the first step, we quantify the content of the articles using topic models and measures of polarization in Sections 3.1 and 3.2. Next, in the second step in Section 3.3, we use the polarization scores and topics to model the relationship between how polarized an article is versus how it is shared on Facebook versus email.

3.1. Topic Modeling Using LDA

Topic models help researchers organize and provide insights into large collections of unstructured text data. LDA is the most common topic model, and it models each document (text) as a distribution over topics and each topic as a distribution over words (Blei et al. 2003). This allows documents to be a part of different topics rather than being separated into discrete groups. LDA models have been extensively used in marketing; typically, researchers apply the LDA model (or derivatives of it) to derive topics and then use them as inputs in downstream models. The applications range from product reviews (Tirunillai and Tellis 2014) to social media content (Zhong and Schweidel 2020), restaurant menus (Puranam et al. 2017), online search queries (Liu and Toubia 2018), and entertainment products (Toubia et al. 2019).

We performed LDA on the corpus of 13,508 articles using the Gensim package for Python, which is based on the variational Bayes algorithm described by Hoffman et al. (2010). Details of the data processing and hyperparameter tuning to identify the optimal number of topics are in Online Appendix B. Table 1 shows the summary of the 40 topics recovered by our LDA model in decreasing order of prevalence in the corpus (and we use the same ordering of topics throughout the paper).[6] In our data, the most prevalent topic is *Family*, and the least prevalent topic is *Judaism*. The third column in Table 1 shows the top 10 words for each topic, listed in decreasing order of their share in that topic. The topic names were chosen manually by heuristically combining these keywords into a single phrase. For instance, topic 4 contains such words as virus, coronavirus, health, test, case, and pandemic, and therefore, it was named *Coronavirus Pandemic*.

Overall, we find that the LDA model can uncover the latent topics in the corpus quite effectively. Figure 2 shows how a topic's prevalence changes over time for the 10 most prevalent topics in the corpus. Notice that the *Pandemic* topic was almost nonexistent until the end of 2019 but became the most popular topic at the beginning of 2020. We refer interested readers to Online Appendix B.3 for additional results on the LDA analysis, including detailed word clouds of the predominant words in each topic and links to the top three articles with the highest proportion of the topic (in our corpus).

3.2. Polarization Measures

Recall that our main research question asks whether the polarization of an article predicts the relative likelihood of being shared on Facebook versus email. As such, a key concept that we need to define and measure is the political polarization of news content. Formally, a news article is considered politically polarizing if the content, text, and opinions expressed diverge away from the center and are closer to either of the extreme ends of the ideological spectrum (DiMaggio et al. 1996, Baldassarri and Gelman 2008).

Prior research has used a variety of approaches to derive or predict the polarization of text/speech by an agent. Typically, these approaches fall into two broad categories. In the first set of methods, researchers have labeled data on the political party/affiliation of the agent or the outlet that created the text. Then, taking the affiliation of the agent/outlet as the ground truth, they characterize the differences in text or speech of the two parties/groups and use these differences to quantify the extent of polarization in a given piece of text; see Gentzkow and Shapiro (2010) and Gentzkow et al. (2019). In the second set of methods, the researchers use crowdsourced methods (Amazon Mechanical Turk) and surveys to score the slant/polarization of individual news articles (Budak et al. 2016). The former approach can only work when there are clear and well-known political affiliations for each piece of text (e.g., in the case of congressional speech), whereas the latter approach is not scalable beyond a small set of articles.

In this paper, we adopt a novel approach to measure political polarization that overcomes the scalability and lack of affiliation problems; we turn to the newly developed LLMs to obtain polarization scores and also validate these scores using standard surveys from human raters, whose scores can be considered an objective and true measure of human opinions on the polarization of content but are not scalable when the number of articles/content is high. We describe both approaches here.

3.2.1. LLM Measures of Polarization.

We used GPT-3.5-turbo, a large language model, to generate polarization scores, and we obtained three types of polarization scores (on a range from one to five).

- Article-level polarization score. Here, we provide the text of each article and ask the model to provide a polarization score.
- Topic-level polarization score. Here, we provide the topic names derived from the LDA model and ask the model to provide a polarization score for each of the 40 topics.
- Topic keywords-level polarization score. One concern with the topic-level score is that the LLM may be very sensitive to the particular way we named the topics. Therefore, for robustness, we provided the model with only the top 10 keywords for each topic (and no topic name) as in Table 1, and we obtained a new set of polarization scores.

We refer readers to Online Appendix C for a detailed discussion of the prompts, the temperatures for each prompt (that drive the stochasticity of the answers), the number of iterations per question, and the procedure used to standardize the polarization scores.

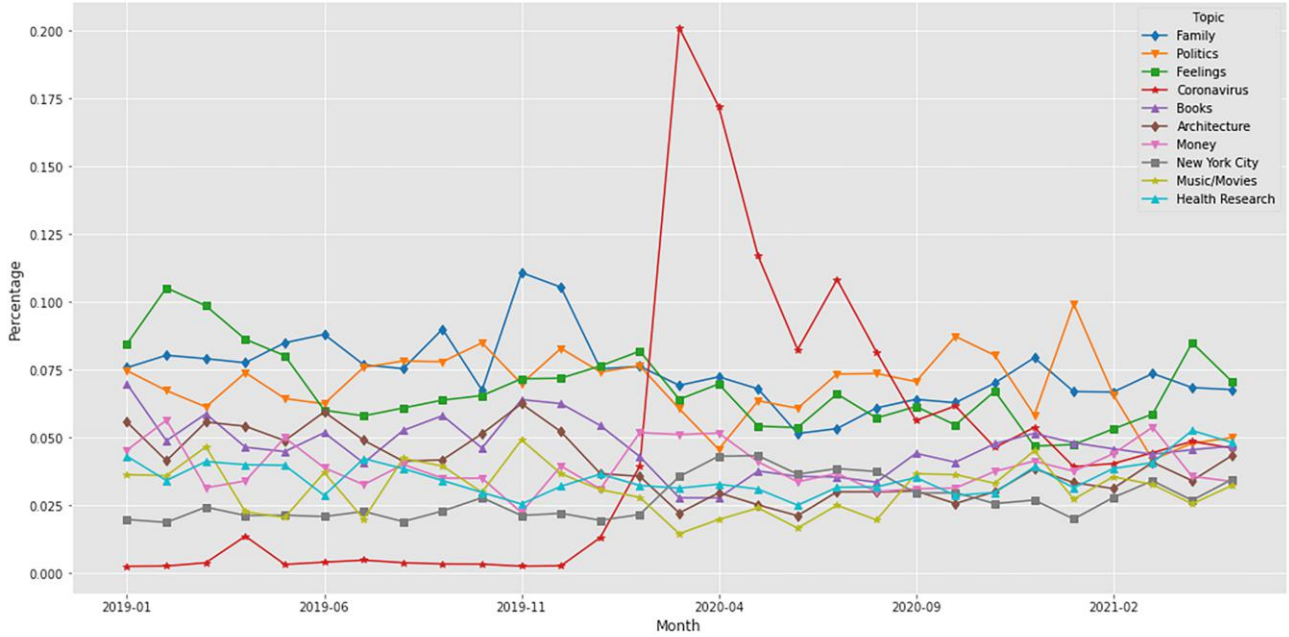
3.2.2. Survey Measures of Polarization.

Next, we conducted a survey to measure the extent to which each of the topics identified from the LDA analysis is considered to be politically polarizing. Note that we use topic-level polarization measures here (instead of article level) because the number of articles (and the length of each article) made it prohibitively expensive to obtain article-level scores. The subjects were undergraduate students at a large state university on the West Coast.

Table 1. Topics

No.	Topic	Prevalence	Top 10 words
1	Family	0.067	Family, home, friend, feel, child, old, mother, love, never, live
2	Politics	0.066	Political, America, article, editor, hear, commit, letter, email, power, diversity
3	Emotions and feelings	0.059	Really, feel, lot, mean, kind, talk, start, happen, ask, question
4	Coronavirus pandemic	0.049	Virus, coronavirus, health, test, case, pandemic, Covid, death, spread, infection
5	Books	0.043	Book, write, story, world, read, writer, man, death, author, novel
6	Architecture	0.042	Open, place, street, room, city, old, design, house, century
7	Money, personal finance	0.040	Pay, money, percent, tax, economic, job, worker, government, economy, income
8	New York City	0.031	City, York, county, home, resident, local, community, area, restaurant, MS
9	Music/movies	0.031	Music, play, film, movie, song, star, watch, character, series, theater
10	Health research, lifestyle advice	0.030	Study, Dr., researcher, research, percent, university, scientist, risk, body, health
11	Nature	0.030	Water, tree, fire, mile, island, river, area, park, foot, town
12	Black Lives Matter	0.029	Police, officer, protest, protester, kill, man, death, video, arrest, fire
13	Women's issues, sexual harassment	0.029	MS, woman, interview, family, man, girl, sexual, member, sex, write
14	Donald Trump	0.029	Trump, president, house, white, administration, news, Washington, fox, Donald, former
15	Elections	0.028	Republican, election, vote, party, senator, house, Democrat, senate, president, Trump
16	Joe Biden	0.028	Biden, campaign, democratic, candidate, voter, party, president, political, Trump, presidential
17	Political investigations	0.027	Case, lawyer, investigation, charge, prosecutor, attorney, justice, report, office, department
18	Public health and medicine	0.025	Patient, hospital, doctor, medical, health, care, drug, Dr., treatment, die
19	Racial identity and history	0.024	Black, White, racial, African, race, man, history, woman, community, America
20	Business	0.024	Company, business, executive, employee, industry, market, sell, product, chief, Amazon
21	Social media	0.024	Facebook, video, post, medium, online, app, social, Twitter, datum, digital
22	Education, school system	0.023	School, student, child, parent, college, university, class, teacher, education, family
23	Supreme Court and judicial system	0.022	Court, law, justice, rule, judge, case, federal, supreme, legal, administration
24	Food	0.021	Food, wine, restaurant, eat, cook, recipe, meat, add, dish, flavor
25	World news	0.017	European, world, Europe, Germany, Britain, British, France, German, united, French
26	American military	0.017	Military, war, Iran, force, united, official, general, troop, Iraq, Iranian
27	Russia	0.017	Official, Russia, Russian, intelligence, security, Ukraine, report, department, government, agency
28	Covid vaccine	0.015	Vaccine, dose, Johnson, vaccination, health, agency, receive, federal, administration, government
29	China, India, international travel	0.014	China, Chinese, government, travel, united, India, flight, passenger, airport, airline
30	Real estate	0.014	Home, building, estate, house, property, apartment, real, rent, buy, housing
31	Power, energy supply, and climate	0.014	Climate, change, power, energy, oil, car, environmental, plant, gas, water
32	Sport	0.012	Game, team, player, play, league, sport, season, coach, club, baseball
33	Art, planes	0.011	Art, museum, artist, bird, plant, painting, plane, paint, pilot, Boeing
34	Science	0.011	Dr., science, space, scientist, university, human, laboratory, team, paper, earth
35	Covid protection	0.011	Mask, wear, risk, face, hand, air, bike, safe, coronavirus, indoor
36	Israel	0.008	Israel, gun, Israeli, Palestinian, Jewish, Muslim, group, Jew, violence, attack
37	Christianity and vhurch	0.007	Church, abortion, religious, Christian, woman, Catholic, gay, faith, god, evangelical
38	Horse racing and farms	0.004	Farmer, farm, run, horse, Japan, race, Japanese, sport, runner, Olympic
39	Pets and animals	0.003	Animal, dog, human, cat, specie, pet, wild, wildlife, park, fish
40	Judaism	0.002	Jewish, funeral, smell, Kelly, community, Allen, wedding, Brooklyn, rabbi, Jew

Figure 2. (Color online) Topic Prevalence over Time



Details of the survey questions, the demographics of the respondents, and their news-reading habits are shown in Online Appendix D. Survey respondents were presented with 10 random topics and were asked to rate how politically polarized the news coverage on a topic is on a scale from one (not at all polarized) to five (extremely polarized); this was followed by questions about demographics and news-reading and -sharing habits. To understand people's motivation for sharing news articles, we also asked respondents to rate how important it was for them that their social circle knew of their opinions on each of the 10 random topics presented to them. This tells us the extent to which sharing opinions on a topic is relevant from an identity-signaling perspective.

3.2.3. Polarization Scores Summary. In Table 2, we present a summary of all four polarization scores. To aggregate the polarization scores from the article to the topic level (the last column in Table 2), we use the weighted average of the polarization score across all articles, where the weights are the proportion of the topic in a given article. We then correlate all the polarization scores and find that there is an extremely high correlation between survey measures of polarization and LLM-generated polarization scores; see Table 3. This is a useful finding because it suggests that future researchers can use LLMs to score news content and text on polarization and ideological issues. In our specific context, this finding allows us to use the LLM-generated article-level polarization scores in our empirical analysis. Further, Table A.4 in Online Appendix C provides examples of the rationale that the LLM provides for its polarization scores. As we can see from this table, the LLM is quite good at explaining why it scores certain articles higher and others lower.

In terms of substantive findings, we see that topics such as *Politics, Elections, Joe Biden, Political Investigations, Black Lives Matter, Women's Issues and Sexual Harassment, and Racial Identity and History* are considered to be the most polarizing. On the other hand, topics such as *Pets and Animals, Architecture, Food, Horse Racing and Farms, Sports, and Books*, are not considered to be polarizing. In Section 3.3, we examine this issue further and estimate the impact of polarization *after* controlling for the topics in the article (Table 3).

3.3. Polarization and Sharing Behavior

3.3.1. Empirical Model. We now specify a simple descriptive model to quantify the difference in the polarization of articles across the two lists. Recall that the rank is the position of an article on the M-Emailed or M-Facebook list and that it can go from 1 (most popular) to 20 (least popular). We define Y_{it} as a measure of the difference between the rank of an article i on the M-Emailed list and the rank of an article i on the M-Facebook list on day t. We consider two measures for Y_{it}: (1) a simple difference metric[7] and (2) an indicator for whether the article was ranked higher (i.e., closer to the top of the list) on the M-Facebook list compared with the M-Emailed list. Therefore, lower values of Y_{it} indicate that article i is more popular on the M-Emailed list compared with on the M-Facebook list.

Table 2. Polarization Measures (Standardized)

Topic	Survey	LLM_Topics_Avg	LLM_Keywords_Avg	LLM_Article
Family	−0.32	−1.23	−1.44	−0.97
Books	−0.82	−1.44	−1.33	−0.74
Feelings	−0.34	−0.69	−0.78	−0.61
Horse racing	−0.93	−0.61	−0.98	−0.79
Music/movies	−0.62	−1.36	−1.25	−1.32
Joe Biden	0.78	0.90	1.26	0.89
Money	−0.18	−0.69	0.53	0.44
Elections	0.79	1.28	1.26	1.67
Public health	0.33	0.06	0.46	−0.66
Pets and animals	−1.27	−1.44	−0.96	−0.98
Donald Trump	0.88	1.58	1.46	1.57
New York City	−0.33	−0.68	−0.98	−0.22
Women’s issues	0.69	0.89	0.51	0.11
Architecture	−1.26	−1.37	−1.44	−1.43
Coronavirus	0.45	1.11	1.08	−0.07
Science	−0.12	−0.83	−0.69	−1.20
American military	0.46	0.14	0.70	1.01
Food	−1.00	−1.28	−1.44	−1.94
Health research	−0.46	−0.37	−0.03	−0.76
Business	0.20	−0.45	0.17	0.13
Social media	0.08	0.07	0.05	−0.09
Black Lives Matter	0.72	1.26	0.98	1.06
Art, planes	−0.95	−1.23	−1.23	−1.19
Racial identity	0.57	1.12	1.07	0.74
World news	0.34	0.14	−0.60	0.14
Nature	−0.45	−1.44	−1.35	−1.27
Judicial system	0.57	0.67	0.44	1.40
Covid vaccine	0.44	0.97	0.79	−0.07
China, India	0.25	0.42	0.62	0.26
Real estate	−0.60	−0.69	−0.78	−0.54
Russia	0.14	0.67	0.81	1.58
Power and climate	0.38	0.83	0.98	0.36
Politics	0.89	1.58	1.08	1.57
Sports	−0.90	−0.69	−0.96	−0.48
Political investigations	0.77	0.81	0.71	1.32
Israel	0.33	1.04	1.07	1.59
Church	0.25	0.37	0.90	0.91
Covid protection	0.36	0.53	0.06	−0.61
Education	−0.04	0.14	−0.39	−0.20
Judaism	−0.18	−0.08	−0.33	−0.58

Next, we specify Y_{it} as a function of the polarization score of article i and other controls as follows:

$$Y_{it} = \alpha + \beta P_i + \sum_{j=1}^{n-1} \gamma_j \cdot p_{ij} + \delta Article_i + \zeta Age_{it} + \epsilon_{it}, \qquad (1)$$

where P_i is the polarization score of article i and p_{ij} is the proportion of topic j in article i. The proportions of all the topics in an article adds up to one, so we exclude the proportion of the topic *Social Media* to avoid collinearity. Because the total number of topics n is 40, we have $n - 1 = 39$ topics in the model. Next, $Article_i$ consists of article-specific attributes, such as its length, the length of its headline, and the section name. We also include controls for the age of the article (defined as the number of days since release) because descriptive evidence suggests that there is a difference in the stickiness of articles across the two lists (see Figure A.3 and the accompanying discussion in Online Appendix A). Finally, for data, we use observations at the day-article level, where for each day t, we include all articles that were ranked at least in one of the lists on that day. This gives us 20–40 observations for each day, which amounts to a total of 23,580 observations over a period of 697 days.

Because Y_{it} is the difference in the popularity of article i across the two lists on day t, it differences out common time-specific shocks that affect an article’s popularity. For example, if the topic *Joe Biden* was popular during elections, then articles on this topic will appear in both M-Emailed and M-Facebook lists. Thus, Y_{it} captures the incremental popularity of the article on email (compared with Facebook) after controlling for other time-varying shocks to the article’s popularity. Further, this

Table 3. Correlation Among the Different Polarization Scores

	LLM_Topics_Avg	LLM_Keywords_Avg	LLM_Article
Survey	0.91	0.88	0.81
LLM_Topics_Avg	1.00	0.93	0.84
LLM_Keywords_Avg		1.00	0.86

specification captures the impact of polarization after controlling for the topic distribution of the article. Thus, differences in topics' inherent tendency to be shared via social media versus shared privately are already captured/controlled for. Further, even if some topics are more polarizing than others (as shown in Table 1), this captures the effect of within-topic variation in polarization on Y_{it}.

3.3.2. Results and Discussion. Table 4 shows the regression results, where Model (1) uses the difference in ranks as the dependent variable and Model (2) in Table 4 uses the binary indicator as the outcome variable. In both regressions, we see that the polarization score has a positive coefficient, which means that more polarizing articles are more commonly shared on Facebook compared with email. Note that because we control for the prevalence distribution of the topics in each article, this estimate is the incremental impact of polarization after controlling for the topic distribution of the article (and the effect of the topic on the tendency to be shared on the two mediums). In Online Appendix E, we consider a model where we do not control for topics and find that the effect of polarization is overestimated in that case. This is understandable because topics that tend to have more polarized reporting also tend to be shared more on Facebook, and this also emphasizes the importance of controlling for topics. Further, in Online Appendix E, we present a series of robustness checks to show that these results are valid even when we vary the model specification and control for the emotional content of the article using Linguistic Inquiry and Word Count measures (Berger and Milkman 2012, Boyd et al. 2022).[8]

In addition, we find that certain topics are more likely to be posted on social media compared with being shared via email (after controlling for the polarization score and other control variables). Specifically, articles on topics such as *Books, Business, Animals, Food, Real Estate, Nature, and Health Research and Lifestyle*

Table 4. Results from Equation (1) Capturing the Difference in the Polarization of Articles Across the Two Lists

Dependent variables	*Difference in ranks (E − F)*		*Higher in F (binary)*	
Model	(1)		(2)	
Variables				
Polarization score	1.088***	(0.108)	0.041***	(0.004)
Family	12.783***	(1.639)	0.467***	(0.056)
Politics	−4.377**	(1.853)	−0.184***	(0.062)
Emotions and feelings	5.542***	(1.809)	0.204***	(0.059)
Coronavirus pandemic	18.352***	(1.554)	0.629***	(0.055)
Books	−2.659	(1.983)	−0.098	(0.068)
Nature	2.017	(1.843)	0.029	(0.061)
Women's issues, sexual harassment	27.063***	(1.954)	0.915***	(0.069)
Business	−4.341*	(2.399)	−0.195**	(0.082)
Education, school system	3.072	(1.895)	0.080	(0.066)
American military	15.975***	(1.780)	0.520***	(0.061)
China, India, international travel	23.807***	(2.263)	0.793***	(0.080)
Power, energy supply, and climate	6.178***	(2.141)	0.187**	(0.076)
Judaism	−26.599***	(6.431)	−1.226***	(0.250)
Architecture	−12.896***	(1.849)	−0.545***	(0.063)
Money, personal finance	7.010***	(1.598)	0.228***	(0.054)
New York City	7.669***	(2.047)	0.266***	(0.072)

Table 4. (Continued)

Dependent variables	*Difference in ranks (E − F)*		*Higher in F (binary)*	
Model	(1)		(2)	
Music/movies	10.841***	(2.089)	0.393***	(0.074)
Health research, lifestyle advice	−3.980**	(1.768)	−0.132**	(0.058)
Black Lives Matter	21.575***	(1.720)	0.682***	(0.055)
Donald Trump	14.857***	(1.876)	0.417***	(0.065)
Elections	13.087***	(1.688)	0.425***	(0.056)
Joe Biden	11.072***	(1.750)	0.359***	(0.061)
Political investigations	15.250***	(1.571)	0.459***	(0.054)
Public health and medicine	7.452***	(1.822)	0.255***	(0.063)
Racial identity and history	6.107***	(1.903)	0.173**	(0.068)
Supreme Court and judicial system	11.040***	(1.852)	0.381***	(0.065)
Food	−2.080	(1.813)	−0.080	(0.062)
Covid vaccine	17.559***	(1.799)	0.508***	(0.063)
Art, planes	−5.755**	(2.536)	−0.216**	(0.089)
Covid protection	−3.213*	(1.899)	−0.181***	(0.068)
Christianity and church	6.937**	(2.730)	0.198*	(0.103)
Horse racing and farms	6.992	(4.390)	0.267*	(0.155)
World news	2.044	(2.351)	0.026	(0.084)
Russia	9.526***	(1.967)	0.322***	(0.068)
Real estate	−7.314***	(2.318)	−0.244***	(0.081)
Sports	8.431***	(2.669)	0.188**	(0.087)
Science	4.450	(2.723)	0.082	(0.097)
Pets and animals	−21.780***	(4.689)	−0.758***	(0.168)
Israel	5.820*	(2.966)	0.198**	(0.100)
Headline length	0.087***	(0.033)	0.003***	(0.001)
Snippet length (standardized)	0.558***	(0.095)	0.019***	(0.003)
Word count (standardized)	−1.152***	(0.094)	−0.041***	(0.003)
In print (binary)	−1.863***	(0.194)	−0.071***	(0.007)
(Intercept)	−22.660***	(1.566)	−0.243***	(0.053)
Controls				
Days after release (quadratic)	Yes		Yes	
Fixed effects				
Section name	Yes		Yes	
Fit statistics				
Observations	23,580		23,580	
R^2	0.33600		0.32589	
Adjusted R^2	0.33346		0.32331	

Notes. Clustered (date_id) standard errors are in parentheses. Significance is indicated with asterisks.
*0.1; **0.05; ***0.01.

Advice are more commonly sent through email. In contrast, articles on topics such as *Election Investigations, Covid Vaccine, Russia, Women's Issues and Sexual Harassment, Coronavirus Pandemic, Black Lives Matter*, etc. are more likely to be posted on social media. Although we do not take a stance on why certain topics are shared more widely on social media, it is worthwhile to note that these patterns are consistent with some natural explanations and earlier works. Notice that the topics shared on social media tend to be of broader interest (e.g., politics, elections), and understandably, they are shared on social media with a larger set of acquaintances compared with other topics. Further, we find that the correlation between a topic's social signaling score (see Table A.5 in Online Appendix D.2) and its coefficient from the regression results of Model (1) in Table 4 is

0.35. This positive correlation is consistent with earlier work that suggests that identity signaling can be a strong motivator behind users' actions on social media (Berger 2008, Reed et al. 2012, Reed and Forehand 2019, van der Does et al. 2022).

We now provide some additional discussion and interpretation of the results. First, our results are descriptive and should be interpreted carefully. Because we do not manipulate polarization or slant within an article exogenously, our results do not state that certain polarization causes an article to be more/less likely to be posted on Facebook. Rather, it simply states that after controlling for a series of observables, such as the topics covered, length, news section, and emotional content of an article, polarizing articles are more likely to be shared through Facebook rather than email. That said, the observed patterns may stem from differences in the segments of consumers who post on social media versus those who share on email or differences in how the same users employ the two communication media (the same user may share cerebral articles with close friends through email but post political/polarizing content on Facebook). Alternatively, these patterns may also reflect the implicit effect of social media algorithms. That is, Facebook users may have learned that their posts on more polarizing topics are more likely to be popular and/or amplified by the internal algorithm and hence, favor those types of articles when posting on Facebook.

Nevertheless, we believe that documenting these descriptive results can help further discussion on this topic, and future research could further examine the sources and channels of polarized content.

3.3.3. Did Sharing Behavior Change over Time?

Recent research has shown that polarization in preferences, behavioral intentions, and actual purchase decisions for consumer brands increased after the election of Donald Trump in 2016 (Schoenmueller et al. 2023). Moreover, polls and anecdotal evidence suggest that polarization on social media platforms has exacerbated after the 2020 elections (Jurkowitz et al. 2020). Indeed, discussions about the polarization of social media platforms have gained urgency and prominence after the January 6 assault on the U.S. Capitol. Experts have argued that these incidents were fomented by the divisive discourse on social media, and these issues have been the subject of a recent senate investigation (Reuters Staff 2021).

Motivated by these arguments, we examine whether the sharing patterns are different after the 2020 elections. We consider two subsets of our data: (1) Pre, which includes data from January 1, 2020 to October 30, 2020, and (2) Post, which includes data from December 1, 2020 to May 30, 2021. We then run the same model (as shown in Equation (1), with the difference in ranks as the outcome variable) but also include a *Post* variable and the interaction between *Post* and polarization. We find that (1) the main effect of polarization continues to be positive (i.e., more polarized articles have a higher relative likelihood of being shared on Facebook in both periods) and that (2) the interaction effect is negative (i.e., the polarization score of articles plays a smaller role in predicting the differences across the two lists after the election) (see Online Appendix F for details). To further understand these patterns, we summarize the average polarization scores of articles that appeared at least once on the Most Emailed and Most Shared on Facebook lists in the pre- and postelection periods in Table 5. Interestingly, we see that the articles shared on both channels were less polarized after the elections, although this drop is higher for Facebook (which explains the results from the regression). In summary, at least in this setting, we do not find that there was any significant increase in the polarization of articles shared on social media after the elections (compared with email).

Table 5. Average Polarization Scores (Standardized) of Articles That Appeared at Least Once on the Most Emailed and Most Shared on Facebook Lists in the Pre- and Postelection Periods

	Pre	Post	Difference (*t*-statistics)
Emailed	−0.08	−0.22	6.96
Facebook	0.21	−0.00	10.82
Difference (*t*-statistics)	−20.57	−9.01	

Notes. The last column shows the difference in the difference in the average polarization scores in the post- and preperiods for each medium and the *t*-statistics. Similarly, the last row shows the difference in the average polarization scores for each period (pre-election and postelection) for a given medium and the *t*-statistics.

4. Conclusion

In this paper, we examine if and how the content of articles seeded on social media (specifically, Facebook) differs from that sent via email. We use data from the N*ew York Times* Most Emailed and Most Shared on Facebook lists for a 2.5-year period for our study. For each article, we recover the topic distribution using LDA and the polarization score using LLMs, and we connect the difference in the article's ranking across the two lists with its polarization score and topic. We show that more polarizing articles are more likely to be seeded on social media (compared with email) after controlling for a series of confounding factors, such as the topic, news section, emotion, age, etc. Our results are descriptive and should be interpreted as summarizing sharing patterns on different channels, and they should not be interpreted as the causal effects of polarization on users' sharing behavior.

Our analysis comes with a set of caveats, which can serve as avenues for further research. First, because we do not observe individual-level data, we cannot comment on whether the same user shares different content across the two media formats or whether the set of users posting on Facebook is systematically different from those users who share news through email. Second, although our analysis shows that these patterns exist before the explicit impact of Facebook's algorithms, it is not clear if there is an implicit impact. It would be useful to examine whether users post more polarizing articles on Facebook anticipating that such articles will be more popular (because Facebook's algorithm promotes such articles) or if this behavior is purely exogenous. Studies that separate the explicit and implicit role of algorithms on user behavior would be an excellent next step. Third, the findings are specific to the setting that we study (i.e., *NYTimes* readers who share articles on Facebook and/or email). Future research that expands that scope to other news websites and social media websites can help with establishing the generalizability of these patterns. Finally, although we do not delve too much into the incentives of news platforms, future research could build on our findings and the growing analytical work that examines platform and news aggregator incentives to create content and price their products (Amaldoss et al. 2021, Amaldoss and Du 2023).

Acknowledgments

The authors thank Mark Forehand, Ali Goli, Simha Mummalaneni, and Omid Rafieian for feedback that has significantly improved the paper.

Endnotes

[1] Reshares of Facebook posts containing *NYTimes* articles or direct link sharing on Facebook/email are not counted when generating these ranks because those are internal to Facebook. Thus, these lists represent the sharing behavior of logged-in *NYTimes* users. Further, these rankings are only based on the sharing behavior from web browsers, and shares from *NYTimes* applications (iOS and Android) are not included when calculating the rank orderings.

[2] Internet Archive has data from late 2015, but the *NYTimes* Trending section was in beta until 2017. Data from 2018 exist, but there are numerous missing observations; therefore, they are not reliable. Hence, we focus our data collection efforts from 2019 onward.

[3] *NYTimes* sometimes changes the headlines of news articles. However, this does not affect our analysis because we work with individual article identifications, and the same article with different headlines is still treated as one unique article.

[4] Some popular articles do not have text because they are in multimedia formats, such as videos or questionnaires.

[5] All the statistics in this section count each article as many times as it appears in the daily data (i.e., they are the summary statistics of the observations, not unique articles on the M-Emailed and M-Facebook lists). However, these summary statistics are largely the same if we instead count each article only once per list.

[6] Formally, prevalence of topic j is defined as $prevalence_j = (\sum_d p_{jd} \cdot \text{length}_d)/(\sum_j [\sum_d p_{jd} \cdot \text{length}_d])$, where d denotes a document and p_{jd} denotes the proportion of topic j in document d. Prevalence for all topics in a document sums to one.

[7] If article i is not ranked in the top 20 in a given list on day t, then we specify its rank as 25 for that list to calculate y_{it}. In Online Appendix E, we present robustness checks with other numbers.

[8] Berger and Milkman (2012) show that content that evokes high-arousal positive (awe) or negative (anger or anxiety) emotions is more likely to be shared by email, whereas content that evokes low-arousal or deactivating emotions (e.g., sadness) is less likely to be shared. Like us, they also use data from the *NYTimes*. However, they confine their analysis to the most emailed articles, whereas our goal is to contrast the differences in the sharing patterns of news articles across social media and emails.

References

Allcott H, Braghieri L, Eichmeyer S, Gentzkow M (2020) The welfare effects of social media. *Amer. Econom. Rev.* 110(3):629–676.

Amaldoss W, Du J (2023) How can publishers collaborate and compete with news aggregators? *J. Marketing Res.* 60(6):1114–1134.

Amaldoss W, Du J, Shin W (2021) Media platforms' content provision strategies and sources of profits. *Marketing Sci.* 40(3):527–547.

Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, Hunzaker MF, Lee J, Mann M, Merhout F, Volfovsky A (2018) Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. USA* 115(37):9216–9221.

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on Facebook. *Science* 348(6239):1130–1132.

Baldassarri D, Gelman A (2008) Partisans without constraint: Political polarization and trends in American public opinion. *Amer. J. Sociol.* 114(2):408–446.

Barberá P, Jost JT, Nagler J, Tucker JA, Bonneau R (2015) Tweeting from left to right: Is online political communication more than an echo chamber? *Psych. Sci.* 26(10):1531–1542.

Berger J (2008) Identity signaling, social influence, and social contagion. Prinstein MJ, Dodge KA, eds. *Understanding Peer Influence in Children and Adolescents* (The Guilford Press, New York), 181–199.

Berger J, Milkman KL (2012) What makes online content viral? *J. Marketing Res.* 49(2):192–205.

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3(January):993–1022.

Boxell L, Gentzkow M, Shapiro JM (2017) Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proc. Natl. Acad. Sci. USA* 114(40):10612–10617.

Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW (2022) *The Development and Psychometric Properties of LIWC-22* (University of Texas at Austin, Austin).

Brand J, Israeli A, Ngwe D (2023) Harvard Business School Marketing Unit Working Paper No. 23-062, Harvard Business School, Boston.

Budak C, Goel S, Rao JM (2016) Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quart.* 80(S1):250–271.

Cinelli M, De Francisci Morales G, Galeazzi A, Quattrociocchi W, Starnini M (2021) The echo chamber effect on social media. *Proc. Natl. Acad. Sci. USA* 118(9):e2023301118.

Dev Portal (2022) Article search. Accessed May 9, 2024, https://developer.nytimes.com/docs/articlesearch-product/1/overview.

DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? *Amer. J. Sociol.* 102(3):690–755.

Eady G, Nagler J, Guess A, Zilinsky J, Tucker JA (2019) How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *SAGE Open* 9(1):2158244019832705.

Gentzkow M, Shapiro JM (2010) What drives media slant? Evidence from US daily newspapers. *Econometrica* 78(1):35–71.

Gentzkow M, Shapiro JM (2011) Ideological segregation online and offline. *Quart. J. Econom.* 126(4):1799–1839.

Gentzkow M, Shapiro JM, Taddy M (2019) Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica* 87(4):1307–1340.

Hoffman M, Bach F, Blei D (2010) Online learning for latent Dirichlet allocation. *Adv. Neural Inform. Processing Systems* 23:856–864.

Iyer G, Yoganarasimhan H (2021) Strategic polarization in group interactions. *J. Marketing Res.* 58(4):782–800.

Jurkowitz M, Mitchell A, Shearer E, Walker M (2020) U.S. media polarization and the 2020 election: A nation divided. *Pew Research Center* (January 24), https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/.

Levy R (2021) Social media, news consumption, and polarization: Evidence from a field experiment. *Amer. Econom. Rev.* 111(3):831–870.

Liu J, Toubia O (2018) A semantic approach for estimating consumer content preferences from online search queries. *Marketing Sci.* 37(6):930–952.

Pariser E (2011) *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Penguin, London).

Persily N, Tucker JA (2020) *Social Media and Democracy: The State of the Field, Prospects for Reform* (Cambridge University Press, Cambridge, UK).

Puranam D, Narayan V, Kadiyali V (2017) The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Sci.* 36(5):726–746.

Reed A II, Forehand M (2019) *Handbook of Research on Identity Theory in Marketing* (Edward Elgar Publishing, Cheltenham, UK).

Reed A II, Forehand MR, Puntoni S, Warlop L (2012) Identity-based consumer behavior. *Internat. J. Res. Marketing* 29(4):310–321.

Reuters Staff (2021) Facebook does not believe it is a primary cause of polarization - Exec to CNN. *Reuters* (October 13), https://www.reuters.com/article/facebook-whistleblower-clegg-idCAKBN2GT0EO.

Schoenmueller V, Netzer O, Stahl F (2023) Frontiers: Polarized America: From political partisanship to preference partisanship. *Marketing Sci.* 42(1):48–60.

Shin D, Kadiyala B (2022) Social learning with polarized preferences on content platforms. Preprint, submitted October 11, https://dx.doi.org/10.2139/ssrn.3916284.

Sunstein CR (2018) *Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, Princeton, NJ).

Tirunillai S, Tellis GJ (2014) Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.

Toubia O, Iyengar G, Bunnell R, Lemaire A (2019) Extracting features of entertainment products: A guided latent Dirichlet allocation approach informed by the psychology of media consumption. *J. Marketing Res.* 56(1):18–36.

van der Does T, Galesic M, Dunivin ZO, Smaldino PE (2022) Strategic identity signaling in heterogeneous networks. *Proc. Natl. Acad. Sci. USA* 119(10):e2117898119.

Zhong N, Schweidel DA (2020) Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Sci.* 39(4):827–846.